Garbage in, Garbage out: How does ambiguity in data affect state-of-the-art pedestrian detection?

Jannes Scholz

BACHELOR THESIS

for the acquisition of the academic degree Bachelor of Science (B.Sc.)

Course of Studies: Applied Artificial Intelligence

Department of Electrical Engineering, Medical Engineering and Computer Science
Offenburg University

29.02.2024

Carried out by the company Quality Match GmbH

Supervisors
Prof. Dr.-Ing. Janis Keuper, Hochschule Offenburg
Dr. Daniel Kondermann, Quality Match

Scholz, Jannes:
Garbage in, Garbage out: How does ambiguity in data affect state-of-the-art pedestrian detection? / Jannes Scholz. –
BACHELOR THESIS, Offenburg: Offenburg University, 2024. 48 pages.

Scholz, Jannes:
Garbage in, Garbage out: Wie wirkt sich Mehrdeutigkeit in Daten auf state-of-the-art Fußgängererkennung aus? / Jannes Scholz. –
BACHELORARBEIT, Offenburg: Hochschule für Technik, Wirtschaft und Medien Offenburg, 2024. 48 Seiten.

Abstract

Garbage in, Garbage out: How does ambiguity in data affect state-of-the-art pedestrian detection?

This thesis investigates the critical role of data quality in computer vision, particularly in the realm of pedestrian detection. The proliferation of deep learning methods has emphasised the importance of large datasets for model training, while the quality of these datasets is equally crucial. Ambiguity in annotations, arising from factors like mislabelling, inaccurate bounding box geometry and annotator disagreements, poses significant challenges to the reliability and robustness of the pedestrian detection models and their evaluation. This work aims to explore the effects of ambiguous data on model performance with a focus on identifying and separating ambiguous instances, employing an ambiguity measure utilizing annotator estimations of object visibility and identity. Through accurate experimentation and analysis, trade-offs between data cleanliness and representativeness, noise removal and retention of valuable data emerged, elucidating their impact on performance metrics like the log average miss-rate, recall and precision. Furthermore, a strong correlation between ambiguity and occlusion was discovered with higher ambiguity corresponding to greater occlusion prevalence. The EuroCity Persons dataset served as the primary dataset, revealing a significant proportion of ambiguous instances with approximately 8.6% ambiguity in the training dataset and 7.3% in the validation set. Results demonstrated that removing ambiguous data improves the log average miss-rate, particularly by reducing the false positive detections. Augmentation of the training data with samples from neighbouring classes enhanced the recall but diminished precision. Error correction of wrong false positives and false negatives significantly impacts model evaluation results, as evidenced by shifts in the ECP leaderboard rankings. By systematically addressing ambiguity, this thesis lays the foundation for enhancing the reliability of computer vision systems in real-world applications, motivating the prioritisation of developing robust strategies to identify, quantify and address ambiguity.

Zusammenfassung

Garbage in, Garbage out: Wie wirkt sich Mehrdeutigkeit in Daten auf state-of-the-art Fußgängererkennung aus?

In dieser Bachelorarbeit wird die kritische Rolle der Datenqualität auf dem Gebiet von Computer Vision untersucht, insbesondere im Bereich der Fußgängererkennung. Die Verwendung von Deep-Learning-Methoden hat die Bedeutung umfangreicher Datensätze für das Training von Modellen hervorgehoben, wobei die Qualität dieser Datensätze von entscheidender Bedeutung ist. Die Mehrdeutigkeit von Annotationen, die durch Faktoren wie falsches Labeling, ungenaue Bounding-Box Geometrie und Unstimmigkeiten zwischen den Annotatoren entsteht, stellt eine große Herausforderung für die Zuverlässigkeit und Robustheit der Fußgängererkennungsmodelle und deren Evaluation dar. Diese Arbeit zielt darauf ab, die Auswirkungen mehrdeutiger Daten auf die Modellleistung zu untersuchen, wobei der Schwerpunkt auf der Identifizierung und Abtrennung mehrdeutiger Instanzen liegt, wozu ein Mehrdeutigkeitsmaß verwendet wird, das die Schätzungen von Annotatoren bezüglich der Objektsichtbarkeit und -identität nutzt. Durch präzise Experimente und Analysen konnten Kompromisse zwischen der Datenreinheit und Repräsentativität der Daten, der Entfernung von mehrdeutigen Daten und der Beibehaltung wertvoller Daten gefunden werden, die sich auf Leistungskennzahlen wie die Log Average Miss-Rate, den Recall und die Precision auswirken. Darüber hinaus wurde eine starke Korrelation zwischen Mehrdeutigkeit und Okklusion festgestellt, wobei eine höhere Mehrdeutigkeit mit einer größeren Okklusionsprävalenz einhergeht. Der EuroCity Persons Datensatz diente als primärer Datensatz, der mit etwa 8,6% mehrdeutigen Instanzen im Trainingsdatensatz und 7,3% in den Validierungsdaten einen erheblichen Anteil mehrdeutiger Daten aufwies. Die Ergebnisse zeigten, dass das Entfernen mehrdeutiger Daten die Log Average Miss-Rate verbessert, insbesondere durch die Verringerung von False Positives. Eine Erweiterung der Trainingsdaten mit Beispielen aus benachbarten Klassen verbesserte den Recall, verringerte aber die Precision. Die Fehlerkorrektur von falschen False Positives und False Negatives wirkt sich erheblich auf die Ergebnisse der Modellevaluation aus, wie Veränderungen in der ECP-Rangliste zeigten. Durch den systematischen Umgang mit Mehrdeutigkeit legt diese Arbeit den Grundstein für die Verbesserung der Zuverlässigkeit von Bildverarbeitungssystemen in realen Anwendungen und motiviert dazu, robuste Strategien zur Identifizierung, Quantifizierung und Behandlung von Mehrdeutigkeit zu entwickeln.

Contents

1 Introduction

More data beats clever algorithms, but better data beats more data.[1]
Deep learning has achieved significant success due to the availability of vast amounts of data. More samples, features and dimensions increase the probability of containing useful data, which can lead to better predictive performance. However, not only data quantity is important to be able to train robust and accurate models. There is a point where adding more data will not improve the model accuracy. Any machine learning model is useless if the features are too noisy or there is not enough variation in the data, regardless of the amount of data available. This underscores the growing importance of data quality in the domain of machine learning.

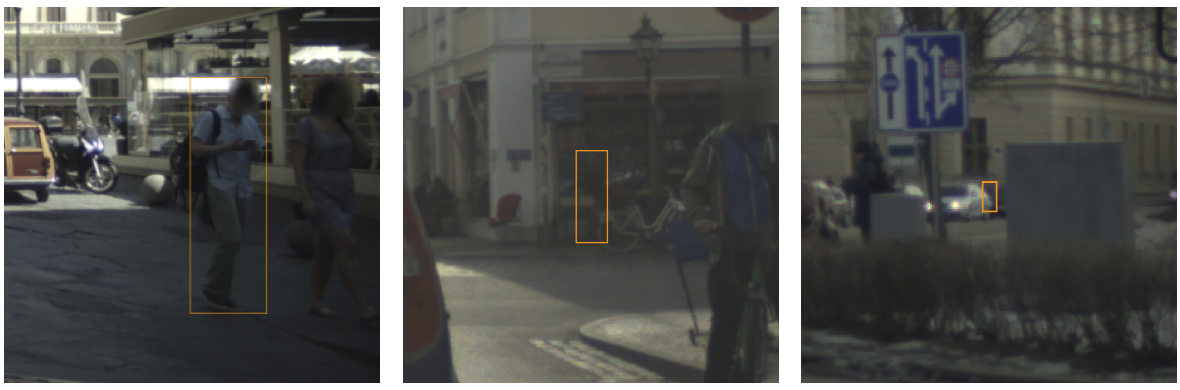
Data Quality and Errors in Computer Vision. Especially in the realm of computer vision, where visual information is interpreted, the impact of data quality becomes even greater. In the context of object detection, for instance, the identification of multiple objects within images involves the delineation of bounding boxes around them, which can then be depicted on the images. These boxes are mainly drawn by humans as most tasks only require common knowledge, e.g. drawing boxes around every car in an image. While doing annotations on images, it can happen that the humans doing this make errors, which lead to inaccurate or incorrect annotations assigned to the objects in the images. These errors can be categorized into three categories: mislabelling, wrong bounding box geometry and inconsistency. Mislabelling happens when there are multiple classes or types of objects that the annotators are asked to annotate. An object in an image might be incorrectly labelled with the wrong class, for example when an object has the label "car" but is actually a human. The second type of error has to do with the drawing of the bounding box that could be wrong in multiple ways. The box can be to big and therefore includes more than only the object that is requested. It can also be too small so that parts of the object are missing.

Error Source Ambiguity. The third potential error is inconsistency, wherein annotators may assign different labels to the same object or may not agree on the existence of an object in question. For example, when completing the task to mark every person in an image with a bounding box. There could be cases where some annotators mark a specific person as such and others don't. This often happens for objects that are ambiguous or difficult to classify,

[1] Peter Norvig - ex Director of Research at Google Inc

leading to a subjective labelling that varies among annotators. These misunderstandings between annotators are not easy to model and therefore lead to a human bias that is inherent in the annotations of the data. Labels that contain a lot of these errors are referred to as noisy labels and data that contains a lot of noisy labels is defined as low quality data. High quality data is therefore accurately annotated, free from errors, consistent (free from conflicting annotations) and representative, which means that it covers the variability and diversity present in the real-world scenarios relevant to the task.

This section delves into the **motivation** for exploring the effects of ambiguous data on pedestrian detection. Algorithms in self-driving cars heavily rely on the data they are trained on. In crucial applications such as pedestrian detection, the failure to detect obstacles can result in severe consequences. Unlike traditional machine learning tasks where some degree of error might be acceptable, in autonomous driving, there is no room for mistakes. Pedestrian detection ensures safety for both pedestrians and passengers in autonomous vehicles. It demands a big amount of data but also great data quality. Diverse scenarios with different lighting conditions and perspectives have to be covered. A clean dataset with precise annotations is needed to provide the model with accurate labels and information so it can perform well in real-world situations. As pedestrian detection is a hard task with zero room for errors, the performance of the models has to be properly evaluated as the quality of the evaluation directly reflects the reliability of pedestrian detection models. Due to occlusion, truncation, high distances and different lighting conditions there is a lot of ambiguity in images of street scenes used to train and evaluate object detectors for this task. Often there are huge annotator disagreements, even when labelled by experts. Ambiguous data has no underlying true label that all of the annotators would agree on. If a lot of ambiguous data is contained in the training and test datasets, this could have a great effect on the training performance but also on the evaluation of the models.



(a) (b) (c)

Figure 1.1: Image examples with ascendingly more difficult identification of a person. 1.1a shows a really clear pedestrian while 1.1b is more difficult to identify and 1.1c is a very ambiguous sample. The 1.1c type of samples are the central objective of this work since it is questionable if they should be in the training or test data.

Ambiguity in Pedestrian Detection. Some examples of the problem are depicted in 1.1. While in 1.1a, it is very clear that there is a walking man looking at his phone in the orange bounding box, for 1.1b it already becomes more difficult even though most humans would probably still agree that there is a human inside the box. However, in 1.1c, the task becomes considerably more challenging as there are no discernible human features visible, only some dark pixels indicating a potential presence of a human. These are exactly the cases, which are questionable to include in the training or test data.

To summarize, this introduction has outlined the importance of high quality data in computer vision, particularly in the context of object detection and bounding box annotations. Especially in the realm of pedestrian detection it is crucial to be aware of the data that is used for the training and the evaluation of the models that could one day decide over a human life. Having concluded the motivation for investigating the effects of ambiguous data on pedestrian detection, the relevant literature will be reviewed in the next section and existing approaches to address these challenges will be explored.

1.1 Related Work

In this section, the related work will be presented which essentially consists of three topics: pruning or improving data prior to the training, using ambiguous data during the training and estimating the uncertainty of annotators. Many different approaches exist because gathering and annotating clean datasets of high quality is a hard task as label errors are found even in public, widely used datasets like CIFAR, MNIST, ImageNet and IMDB [1]. Therefore, label errors and human biases are a common problem in machine learning, which are addressed by the following approaches.

Pruning Data. The first topic consists of methods that clean the data by pruning the ambiguous or incorrect samples and methods that improve the quality of the data, both prior to the training. One way to detect and prune unwanted data this is to train classifiers that serve as noise filters for the training data. The evaluation of a single algorithm, majority voting and consensus filters on five datasets showed that classification accuracy can be improved by up to 30%. For less data, consensus filters seem to perform better and if more data is available, majority voting is the option to choose [2]. This also works well for automated land cover mapping [3]. Biological data, especially in the case of gene expression data, is often redundant and very noisy due to errors during the data collection or inaccurate equipment and tools. To counter this, distance-based pre-processing techniques for noise detection are used in [4] to define different types of data: mislabeled cases, redundant data, outliers, borderlines (close to decision border) and safe cases. This lead to a simplification of machine learning classifiers and a reduction of their error rates through excluding unclear or simply wrong data from the training.

The second area of study of the first topic is to deal with ambiguous data by leveraging techniques to mitigate label ambiguities and enhance the quality of training data. In visual saliency estimation, where accurate identification of visually important image and video contents is crucial, label ambiguities may arise due to inaccurate user data obtained through manual labelling like in other fields [5]. They proposed a multi-instance learning approach that incorporates correlations between image patches into an ordinal regression framework. By training a ranking model and relabelling the samples according to their mutual correlations to other image patches, label ambiguities could be effectively removed from the training data.

Using Ambiguous Data during Training. Another way of dealing with ambiguous and erroneous data besides pruning or correcting it, is to make use of it during the training. The approaches of the second big topic suspect that even noisy data can be useful in the process of learning. This becomes an even better solution when there is limited training data available. Recent research explores leveraging incorrect and ambiguous training data, recognizing its potential usefulness despite the containing errors and potential annotator disagreements. In a study focusing in-home health monitoring via IoT data, a new framework called LeMAL is proposed to learn from mislabeled data through ambiguous learning [6]. This is done by converting the original data to ambiguous data and then applying a distance-based ambiguous learning algorithm to it. [7] examined the effect of different types of label noise on the performance of an object detector. Noisy labels are handled with a method called co-teaching which uses the memorisation characteristics of neural networks in a way that noisy labels don't fit in the simple patterns that are learned before memorising examples. In this work, simulated noise on the KITTI dataset is used, no real noise from annotators for example. Also the benefits of co-teaching might diminish with larger datasets which would not be helpful for pedestrian detection where large amounts of data are needed.

Modelling and Evaluating Annotators. The last topic concentrates on the uncertainty and behaviour of the annotators since most of the labelling today is still done by human annotators even though there is a lot of research in progress to replace humans by foundational models, which would save a lot of money and time. Several approaches try to estimate the behaviour of human annotators. The Inter Annotator Agreement is used to determine the agreement between the annotators. It is defined as a measure of objectivity as it makes it possible to determine the extent to which the annotation results are independent of the annotators. [8] asked annotators to indicate the certainty of their annotation. The outcome was filtered for high and low agreement to understand and analyse causes of disagreement. Another approach is to learn an annotator model and the true label distribution at the same time [9]. In this work, a regularization term is added to the loss function that makes the model converge to a true confusion matrix for each annotator. This way, the skills of annotators can be estimated even if there is only one label available per image. The problem of this

approach is that the assumption is made that there is only one ground truth label for each input which no longer holds true when input is truly ambiguous. This is often the case in pedestrian detection data.

The reviewed literature provided valuable insights into various approaches for handling and utilizing incorrect or ambiguous data. While pruning ambiguous data seems to be an efficient and simple approach, none of the presented approaches used repeated annotator estimations in an object detection topic to filter the data prior to the training and evaluation, which is the purpose of this work. In the next section, the research objectives will be presented.

1.2 Research Objectives

In the preceding discussion, the significance of high quality data in the content of pedestrian detection was highlighted. The challenges posed by errors in annotations, such as mislabelling, incorrect bounding box geometry and annotator disagreements directly inform the objectives of this thesis. The ambiguity present in pedestrian detection images serves as the focal point of the investigation and the research aims to draw attention to the effect that ambiguity has by identifying and separating ambiguous data, leveraging annotator answers and quantifying the impact of such data on model performance. Part of this is to use the estimations of annotators to develop a measure which ranks the samples by their ambiguity so that the data can be split and investigated at different ambiguity thresholds. This information is then used during the training and evaluation of a state-of-the-art object detector for pedestrians to measure the impact of ambiguous data in different ways. Not only the result but also the severity of the error is important. Given the existence of a leaderboard showcasing the best performing models for the dataset that will be used, the primary objective is to assess and quantify the impact of ambiguous data on this ranking.
In this chapter, the motivation behind the thesis topic was explained including the challenges within the problem domain. Related literature was reviewed and the objectives of this work were defined. The subsequent chapter will delve into the experimental design, the selected datasets, model specifications, as well as the definition and detection methods of human ambiguity.

2 Experimental Design

This chapter delineates the research design, presenting the overall strategy that outlines how the research will be conducted, along with a comprehensive definition of human ambiguity. Subsequently, the used dataset will be introduced including the evaluation methodology. This is followed by an explanation of the detector and backbone model trained in conjunction. Lastly, the computation of the ambiguity measure using a general binary model is presented, together with its application to the dataset.

2.1 Research Design

A subset of the Eurocity Persons dataset was re-labelled with noisy human-generated labels by a labelling company to reflect real-world noise like it exists in human annotated datasets in general. To quantify ambiguity of instances, human annotators were asked precise questions concerning the visibility and identification of the objects inside the bounding boxes. This way, information about the variability of annotator answers, label credibility and ambiguity could be derived. The retrieval of this information happened semi-automatically during the pipeline treatment for quality assurance at Quality Match. Furthermore, a correction of false negative and false positive instances within the training subset was conducted, resulting in potentially higher quality data used to train an object detector that achieved state-of-the-art performance on the entire original ECP dataset. After the derivation of an ambiguity measure that sorts the samples, the data was segmented at different thresholds to perform further experiments on the whole ECP dataset, evaluating the impact of ambiguous data on both training and test datasets.

Assumptions. Hypothesizing that models would benefit from training on enhanced data with higher quality and fewer ambiguous samples compared to the original dataset. Also, models trained on ambiguous data are expected to produce more false positive errors since more edge cases are contained in their training data so they reproduce the "hallucination error" of their training data. In contrast to that, adding samples from neighbouring classes to the training data should result in fewer false negatives. Moreover, given that the test data also contained ambiguous samples, the pruning of highly ambiguous data from the test dataset

was predicted to increase the recall and the differences between models trained with varying amounts of ambiguous instances. Furthermore, the presence of ambiguous data in the test dataset was expected to influence the ranking of the leaderboard.

2.2 Human Ambiguity

The quest for precise and clear data is a great subject in the realm of machine learning, deep learning and especially computer vision. However, inherent to the human behaviour and experience is a phenomenon that challenges this pursuit: ambiguity. Human ambiguity introduces subjectivity and uncertainty into the data which can create significant problems for tasks like pedestrian detection. Ground truth labels work as the foundation of detection models but human annotators may exhibit uncertainty when labelling images. This can have various reasons: occlusion of objects, bad image quality, lighting conditions or different interpretations of pedestrian boundaries. The introduced ambiguity leads to discrepancies among ground truth labels. Furthermore, in some complex domains, finding a single correct answer is often impossible due to the inherent ambiguity in real-world data [10]. While machine learning and deep learning systems try to be as certain and precise as possible, human ambiguity challenges this by the subjective nature of perception and interpretation. Different humans or individuals may perceive and interpret images differently due to different experiences, expertise or cognitive biases. Inter-observer variability is another reason for ambiguity in data. Additional challenges exist in image segmentation where pixels at object boundaries are often very ambiguous which leads to even more inconsistencies. In multi-class problems, ambiguity in ground truth labels often results in label correlation, where some features are associated with multiple labels. This complicates the task immensely by creating dependencies within the data which have to be explored to produce robust detection algorithms.

Using Ambiguity. Ambiguity can also improve recognition performance if effectively exploited. By leveraging the inherent uncertainty in the data, models can become more robust to variations in real-world scenarios. However, it is challenging to find the right balance between exploiting ambiguity and maintaining accuracy. By reducing the number of training images with the knowledge of ambiguity, it may be possible to achieve comparable performance with a smaller training dataset [11]. This would have a significant impact for resource-constraint environments or real-time applications.

Figure 2.1: Two example images with drawn bounding boxes from the Eurocity Persons dataset. The left image contains less pedestrians which are more far away than the crowd of persons on the right image. Overlapping, occluded and truncated bounding boxes exist in the data.

2.3 Eurocity Persons Dataset

The Eurocity Persons dataset provides a large number of highly diverse annotations of pedestrians, cyclists and other riders in urban traffic scenes [12]. Its images were collected by a camera mounted on a moving vehicle in 31 cities of 12 European countries during different seasons and weather conditions and were manually annotated by human annotators. Figure 2.1 shows one less crowded image in Barcelona and one bigger crowd on a square in Roma. The amount of wrong annotations (missed and hallucinated objects) is claimed to lie within the 1% range. In table 2.1, the properties of different person detection benchmarks are compared, showing that Eurocity Persons was the most diverse dataset at this point. The test data labels were not publicly available because they are used to compare state-of-the-art models on a detection leaderboard. As part of scientific work, anyone can submit the predictions of their model on the test data and thus appear on the leaderboard. For this work, the validation dataset of ECP was used as the test dataset because the ground truth labels were fully available.

	Caltech	KITTI	CityPersons	TDC	**EuroCity Persons**
# countries	1	1	3	1	**12**
# cities	1	1	27	1	**31**
# seasons	1	1	3	1	**4**
# images (day / night)	**249884** / -	14999 / -	5000 / -	14674 / -	40217 / **7118**
# pedestrians (day / night)	**289395** / -	∼9400 / -	31514 / -	8919 / -	183004 / **35309**
# riders (day / night)	- / -	∼3300 / -	3502 / -	**23442** / -	18216 / **1564**
# ignore regions (day / night)	57226 / -	∼22600 / -	13172 / -	- / -	**75673** / **20032**
# orientations (day / night)	- / -	∼12700 / -	- / -	- / -	**176879** / **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

Table 2.1: Comparison of person detection benchmarks in vehicle context [12]. The columns represent the different datasets and the rows compare various aspects. EuroCity Persons offers the most variety when it comes to countries, cities, seasons and weather conditions which makes it one of the most useful datasets for real-world applications where models need to perform robustly across diverse environmental conditions and scenarios.

subsets	height	occlusion	truncation
reasonable	>40px	<40%	<40%
small	30-60px	<40%	<40%
occluded	>40px	40-80%	<80%
all	>20px	<80%	<80%

Table 2.2: Data subsets of EuroCity Persons for the evaluation. The rows depict the four different subsets while the columns compare three filter options. "Reasonable" represents minimal occlusion and truncation with box sizes >40px which should include most relatively clear cases. "Occluded" is expected to be the hardest subset to predict due to exclusively high occluded instances and "all" provides an average representation across all subsets.

Tags. Pedestrians were additionally annotated with tags for occlusion, truncation at the image border, non-upright poses (sitting, lying) and being behind glass, a reflection (e.g. in store windows) or a depiction (e.g. large posters). If persons were occluded, the bounding box was estimated to its full extent. There are four possibilities for occlusion and truncation: no tag, >10%, >40%, >80%. The evaluation was done on four subsets of the whole validation/test data which are shown in table 2.2. On the leaderboard, the evaluation metric was calculated for each subset.

2.4 Evaluation Details and Metric

The matching of the detected bounding boxes to the ground truth was done using the intersection over union (IOU) that expresses the degree of overlap of two boxes. The detected boxes were sorted by their score which was output by the model, indicating the certainty of the decision, and then matched with the ground truth from highest to lowest score. To evaluate the detection performance of the models, the miss-rate (mr) was plotted against the number of false positives per image (fppi):

$$mr(c) = \frac{fn(c)}{tp(c) + fn(c)}, \tag{2.1}$$

$$fppi(c) = \frac{fp(c)}{\#img}, \tag{2.2}$$

where *fn(c)* is the number of false negatives, *tp(c)* is the number of true positives and *fp(c)* is the number of false positives. c is the confidence threshold of the samples so that only detections with a confidence value greater or equal c are used. Decreasing c means taking more possibly unsure detections into account which would result in more true or false pos-

itives and less false negatives. The log average miss-rate (LAMR) combines the miss-rate and false positives per image as follows:

$$LAMR = exp\left(\frac{1}{9}\sum_{f} log\left(mr(\underset{fppi(c)\leq f}{\operatorname{argmax}} fppi(c))\right)\right), \tag{2.3}$$

where the 9 points f are equally spaced in the log space. This means that they are ranging from 10^{-2} to 10^{0} with increasing the exponent by 0.25 each step. The expression $\operatorname{argmax}_{fppi(c)\leq f} fppi(c)$ finds the maximum false positive per image value that is less than or equal to *f*. Then the miss-rate corresponding to the *fppi* value is taken in the logarithm base 10.

Ignore Regions. Similar to other object detection datasets, EuroCity Persons includes ignore regions in cases where it is uncertain whether an object belongs to the correct class, instances are grouped and cannot be separated properly or the object is exceptionally small. In the ECP evaluation, the following regions are designated as ignore regions:

- Persons smaller than 20 pixels
- Groups of people that cannot be clearly separated ("person-group-far-away" label)
- Neighbouring classes depending on the application
- Persons with the tags "behind glass" or "sitting-lying"

For the pedestrian detection task, "rider" is considered a neighbouring class and thus an ignore region. Depending on the actual application of the model, the settings for neighbouring classes could be more or less strict. For certain applications, it may be critical to avoid misclassifying riders as pedestrians, making it necessary for "rider" not to be a tolerated class. Detections that could not be matched with the regular ground truth boxes in the test or validation data, can be matched with ignore regions if they have an intersection of at least 0.5. Those detections are not counted as false positives and therefore neither rewarded nor penalized.

2.5 Model and Training Setup

The Pedestron [13] repository, whose top-performing model ranks third on the ECP leaderboard, was used for the modelling. Since the model on the first place uses neural architecture search (NAS), the resulting model architecture from NAS is very specific to the dataset and the aim to draw broader conclusions regarding the impact of ambiguous data on model performance, this model was excluded from consideration. The second-ranked model is an evolution of Pedestron. Another reason for Pedestron [13] was the availability of the entire repository which could be used very flexibly with freely downloadable weights.

Pedestron Model Specifications. Their best model uses an HRNet [14] as the backbone network architecture into a Cascade R-CNN [15] general object detector. When evaluating on the Eurocity Persons dataset, their findings indicated that training on the ECP dataset also yields optimal results (refer to tables 6 and 7 in [13]). This is the case because Eurocity Persons is the most diverse dataset among the tested ones in the autonomous driving context. Due to limited processing capabilities and time on one RTX 4090, another backbone network, the MobileNetV2 [16] was chosen for this work. Notably, prior research has demonstrated that even lighter-weight backbone architectures can perform competitively to state-of-the-art methods [13]. Depending on the experiment, the training was conducted either on a subset or the entire training data of ECP and the testing was consistently performed on the validation set of ECP unless stated otherwise. Models from the Pedestron repository accept annotations in the .json format similar to COCO [17] annotation style. All of the models are trained from scratch without any pre-training with the following parameters:

- Batch size: 2
- Optimizer: Stochastic Gradient Descent
  - Momentum: 0.9
- Learning rate: 0.02
  - Weight decay: 1e-4
  - Policy: Cosine
  - Warm-up strategy: Linear
  - Warm-up iterations: 500
  - Warm-up ratio: 1/3

Training Time. Given the time required for one training cycle lasting approximately 4 days for 50 epochs, the decision was mode to limit each training to this duration. However, in 2.2, the training was continued until epoch 150 and lead to further improvements. Despite this, the relative performance within the constraint of only training for 50 epochs is still comparable and the trade-off of investing an additional 8 days of training for minimal improvements had to be considered.

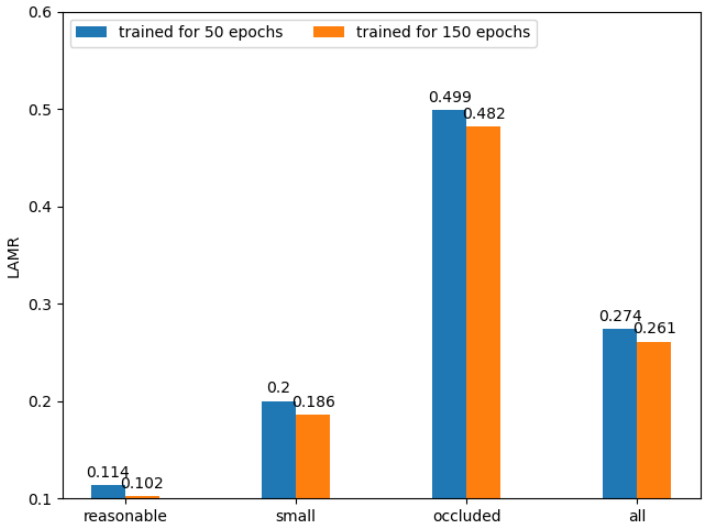


Figure 2.2: LAMR comparison on the test data of the same model trained for 50 and continued to 150 epochs. On the x-axis are the four subsets defined by ECP and the y-axis shows the log average miss-rate of the two models for each subset. The model trained longer was able to generalize better in each evaluation subset while the substantial improvement of 10.5% in the "reasonable" subset underscores the significance of long training for enhanced model performance.

2.5.1 MobilenetV2

MobileNetV2 [16] is a convolutional neural network (CNN) primarily developed for mobile devices. It uses depth-wise separable convolutions which split the computation into two steps: depthwise convolution and pointwise convolution. Depthwise Convolution applies one single filter for each input channel unlike regular 2D convolution where the filter is as deep as the input. Pointwise convolution is then applied on the output of the depthwise convolution to create a linear combination of it. This way is much more efficient than regular convolution, reducing computational complexity and model size. One key feature of MobileNetV2 is the usage of inverted residuals. A shortcut connection with linear bottleneck layers is introduced to improve the information flow through the network while keeping the computational cost low. The linear bottleneck structure minimizes non-linearity within the residual block which allows an efficient information flow and maintains feature expressiveness. The architecture contains repeated building blocks that have depthwise separable convolutions, inverted residuals and linear bottlenecks. This flexible architecture allows customization in terms of depth and width for different use cases. Within the Pedestron repository, the architecture comprises 17 inverted residual blocks (see 2.3). Each of these blocks includes a sequence of ConvBNReLU layers, involving convolution, batch normalization, and ReLU activation. Additionally, the architecture features a ConvBNReLU layer at the beginning and at the end.

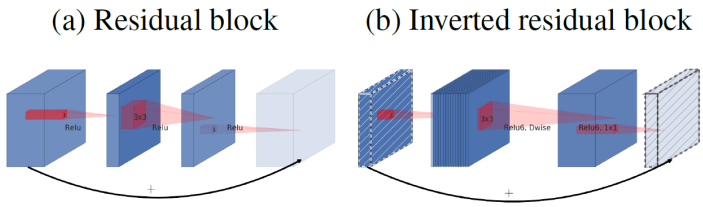


Figure 2.3: The difference between a residual block and an inverted residual. Classical residuals connect layers with high number of channels while inverted residuals connect the bottleneck layers [16]. Inverted residual blocks have a better information flow while keeping the computational complexity low.

2.5.2 Cascade R-CNN

The Cascade R-CNN [15] [18] addresses the problem of degrading model performance with increased intersection over union (IOU) thresholds. This has two main causes: overfitting during training to a specific IOU threshold and mismatch between training IOU and inference IOU. The widely used IOU threshold of 0.5 often leads to more ambiguous and noisy detections while with higher IOU thresholds, the detection performance degrades. The multi-stage architecture of the Cascade R-CNN combines multiple detector stages trained with increasing IOU thresholds in a sequence. In Figure 2.4b, the output of each stage is the input of the next stage which is always more selective against close false positive cases. Therefore, every stage tries to find a set of close false positive cases for the training of the next stage. That is how the sequence of detectors with increasing IOU thresholds vanishes the problem of overfitting. During the inference, the same cascade architecture is applied and therefore eliminates the mismatch problem.

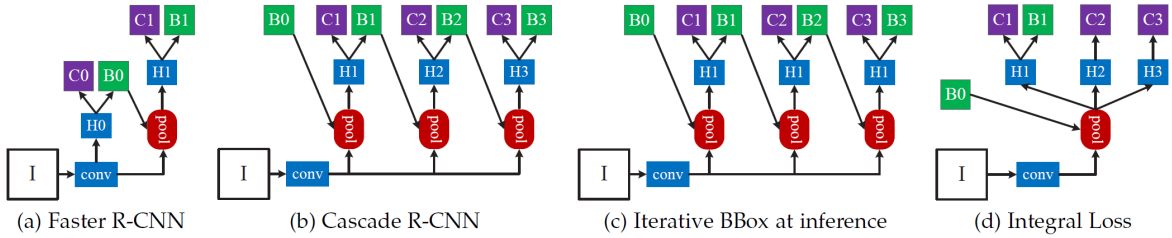


Figure 2.4: The architectures of different frameworks for detection. "I" is input image, "conv" is backbone, "pool" is feature extraction, "H" is one network head, "B" is the predicted bounding box, "C" is the classification of the box and "B0" is region proposals [15]. The Cascade R-CNN (b) has multiple detector stages with increasing IOU thresholds, challenging the problem of overfitting to one threshold like in (a).

2.6 Data Subset and Correction

To facilitate the identification and pruning of ambiguous data from the training dataset, a subset comprising 6000 randomly sampled images from the original training dataset was generated. During the sampling process, the distributions of images per city were taken into account. This means that approximately the same proportion of images was chosen from each city. The subset was re-labelled by a labelling company and therefore expected to contain human real-world label noise. The aim was to correct the subset so it contains less noisy data which can be filtered by different criteria. To achieve this, annotators were asked the following questions about the ground truth samples of the re-labelled subset:

- Is the object in the box a human being? - yes/no
- Can you identify the object inside the box? - yes/no
- Is the person in the box on a motorbike/bike? - yes/no
- Is the object in the box a reflection of a person? - yes/no
- Is the object in the box a statue/mannequin? - yes/no
- Is the person in the box on a poster/picture/billboard? - yes/no

Annotator Estimations. For every ground truth box, each question was asked between 5 and 10 annotators, depending on their answers. For example, if the first 5 annotators agreed on a question, it was not put to any other annotators. In the opposite case, they could also strongly disagree and it would end with 10 asked annotators. It is possible that the votes are even at the end with 5 votes each for "yes" and "no". These cases, where annotators did not agree on one answer are considered as very ambiguous cases. Furthermore, the variability and distribution of annotator answers could be used to calculate 3 measures: solvability, convergence and credibility.
In addition to the answers "yes" and "no" to the binary questions here, annotators could also vote for "can't solve" if they were not able to decide for an answer. **Solvability** describes how many annotators found the task unsolvable. A value close to 1 suggests that few annotators voted for "can't solve".
The **convergence** is a measure which indicates how quickly a probabilistic process reaches a stable state. In this case, it is high if the opinion of a group of annotators tends to settle around one common answer ("yes" or "no") over time. The implementation uses Bayesian updating of probabilities along different branches of a binary tree. It expresses the level of stability in the answers provided by those branches.

The third measure provides information about the reliability or trustworthiness. The more answers are consistent, the lower the **credibility** but if answers are conflicting, credibility should be higher. It is calculated based on a high-density interval obtained from a binary analytical model. The interval indicates a range where the true value will most likely be.

subsets	is object human being?	can you identify?	person on a bike?
QM 1	"no", solvability<=0.45	"no"	"yes"
QM 2	count("yes")=count("no")	count("yes")=count("no")	count("yes")=count("no")
QM 3	"yes", credibility>=0.6	"yes", convergence<=0.2	"no", credibility>0.5

Table 2.3: Overview of the pruning conditions for the subsets filtered by annotator answers. "no" and "yes" mean the aggregation of annotator answers. In ascending order, the conditions of the predecessors apply backwards, e.g. QM 3 is also filtered by the conditions of QM 2 and QM 1. QM 3 experienced the most filtering by ambiguity which results in containing the least ambiguity among all subsets. Models trained on QM 3 are likely to exhibit improved performance and generalization capabilities.

The implementation is a little irritating since the value directly represents the width of the interval. The narrower the interval, the lower the credibility because it suggests a more precise estimation.

These measures were used to create 3 additional training data subsets to the originally and externally annotated ones with 6000 images. The first 3 questions of the listing at the beginning of this chapter were found to be the most valuable ones and were therefore used for this task.

Additional Training Datasets. In table 2.3, the conditions of the data that was pruned from the re-labelled 6000 images training subset are displayed. The thresholds were mainly chosen by manually inspecting the data at different filtering options. In addition to the false positive analysis of the boxes provided by the labelling company, 4942 false negative cases were found by asking the annotators to draw a tight box around any person which was not marked by a box already. Only boxes with a very high solvability were added to each of the 3 new datasets.

2.7 Ambiguity Measure

While until now the efforts were focusing on pruning ambiguous data from the training data, it is equally crucial to understand the extent of ambiguity present in the test data. Because these efforts included a lot of manual data inspection, parameter setting and demanded a high data understanding, an ambiguity measure was developed to assess the uncertainty of a crowd from the distribution of responses per binary task [19]. This score aims to sort the data according to its ambiguity where 1 stands for very ambiguous. That way, datasets can easily be split into clear and ambiguous cases by setting thresholds for the ambiguity measure. This enabled the investigation of different degrees of ambiguity.

Assuming that each annotator chooses from a discrete set of answers $\{yes, no, cs\}$, n^{yes}, n^{no} and n^{cs} represent the frequencies of responses (cs stands for "can't solve"). n is the total number of answers of a task. The following equations define the disambiguity as a baseline:

$$\tilde{a} = \begin{cases} 2 \cdot |\tilde{p}^{\text{yes}} - 0.5| & \text{if } n^{\text{yes}} + n^{\text{no}} > 0 \\ 1 & \text{otherwise.} \end{cases} \tag{2.4}$$

$$\tilde{p}^{\text{yes}} = n^{\text{yes}}/(n - n^{\text{cs}}) \tag{2.5}$$

$\tilde{p}^{\text{yes}}$ is the rate of observed "yes" responses within the responses that consider the task solvable [19]. A disambiguity of 1 indicates a non-uniform distribution of "yes" and "no" answers. If all answers equal "cs", the value is set to 1. To account the number of "cs" answers, another term is added to scale the disambiguity:

$$a = \eta \cdot \tilde{a} \qquad \text{where} \qquad \eta = 1 - \frac{n^{\text{cs}}}{n} \tag{2.6}$$

Now, if "can't solve" receives all votes, $\eta = 0$ which results in $a = 0$ and the disambiguity is minimal. This also applies for $n^{\text{yes}} = n^{\text{no}}$. The ambiguity measure b can be obtained by subtracting the disambiguity from 1: $b = 1 - a$. If $n^{\text{yes}} = n$ or $n^{\text{no}} = n$, $a = 1$ and the minimum ambiguity is achieved.

Application. To apply the ambiguity measure to the data, annotator answers for at least one question for every single sample are needed. Since the annotations for larger datasets are expensive and time-consuming, a model was trained on the annotator answers of the 6000 images subset. It is able to simulate the votes for the test dataset and the rest of the training data for the question "Is the object in the box a human being?". The predictions can be used to calculate the ambiguity measure for non-human-annotated data so that it can be split and inspected at different thresholds. This way, ambiguous samples can be pruned from the whole training and test data. The next section explains the functionality of this model.

2.7.1 General Binary Model

The main goal of the used work is to reduce costs and enhance efficiency in generating labelled data for supervised machine learning with a model-based approach [20]. The focus is set on visual datasets annotated by humans or machines. The responses of crowd workers to simple categorical questions, e.g. if the object in the box is a human being, were modeled using a multinomial-Dirichlet approach. This allowed for an analytical assessment of posterior distributions with the aim of making them a learning objective. To automate the annotation process, a convolutional neural network (CNN) was trained on visual representations, predicting continuous distributions for each object instead of discrete labels. These continuous distributions worked as prior distributions for Bayesian inference.

Modelling Annotator Responses. In detail, each annotator $r \in R = \{1, ...R\}$ is choosing an answer from the set $C = C' \cup \{cs\}$ where C' are the response options to a specific task, here $C' = \{yes, no\}$ as answers to the human being question, and "cs" again for "can't solve" when the task is considered unsolvable. The assumption was made that every annotator answers each task $t \in T$ maximal one time. For the observation number $i \in N$, the labeller $r_i \in R$ and the observed response $a_i \in C$, a dataset is defined as $D = \{(i, t_i, r_i, a_i)\}_{i=1}^{N}$.
Now, the responses of the annotators to the question "Is the object in the box a human being?" were used to update a chosen prior distribution with Bayesian updating. A simple first approach ignored the can't solve cases to have a dichotomous task where $C = C'$ holds. Each annotation task and the responses provided by different labellers for a task were considered independent. As it was the aim to make predictions about the tasks based on the responses obtained, the inference problem was defined as $\{q_t\}_{t=1}^{T}$ with T independent parameters. A beta distribution served as a prior belief about the success probability of each task. The parameters (0.5, 0.5) represent an uninformative prior, suggesting an equal chance of success and failure. The likelihood function modelled the probability of observing the responses a_i given the success probability q_{t_i}.

$$\text{Prior on success probability: } q_t \sim \text{Beta}(0.5, 0.5) \quad \forall t \in T \tag{2.7}$$

$$\text{Likelihood: } a_i \sim \text{Bernoulli}(q_{t_i}) \quad \forall i \in N \tag{2.8}$$

Due to the simplicity of neglecting annotator identity and leveraging the conjugacy of the beta prior and Bernoulli likelihood, the posterior distribution of the parameters q_t could be derived:

$$p(q_t|D_t) = \text{Beta}(q_t; 0.5 + n_t^+, 0.5 + n_t^-) \tag{2.9}$$

The observations of a task t are denoted by D_t and n_t^+ is the number of positive answers given by the annotators. Accordingly, n_t^- is the number of negative answers for a task t.

Considering Unsolvable Tasks. An extension of this model was needed to consider the tasks that might be unsolvable. For each task t, two parameters are introduced: π_t as the probability that the task is solvable and p_t, the probability of success if a task is solvable. The generative process is similar to the simple binary model before:

$$\text{Prior probabilities: } q_t \sim \text{Dirichlet}(\alpha_0) \quad \forall t \in T \tag{2.10}$$

$$\text{Solvability probability: } \pi_t = q_t^{no} + q_t^{yes} \quad \forall t \in T \tag{2.11}$$

$$\text{Success probability: } p_t = q_t^{yes}/(q_t^{no} + q_t^{yes}) \quad \forall t \in T \tag{2.12}$$

$$\text{Likelihood: } a_i \sim \text{Cat}(\pi_{t_i}(1-p_{t_i}), \pi_{t_i}p_{t_i}, 1-\pi_{t_i}) \quad \forall i \in N \tag{2.13}$$

The likelihood function a_i describes the distribution of observed responses based on the probabilities π_t and p_t. It is represented as a categorical distribution where the observations follow a probability distribution imposed by a Dirichlet distribution. Again, an uninformative prior $a_0 = (1,1,1)^T$ was used which allows the data to have a significant impact on the posterior distribution. Leveraging the conjugacy of prior and likelihood one more time to express the posterior distribution of q_t as a Dirichlet distribution:

$$q_t|D \sim \text{Dirichlet}(\alpha_0 + \sum_{k \in C} n_t^k e_k) \tag{2.14}$$

The prior distributions were updated by adding the number of responses n_t^k for each class $k \in C$. The variables e_k simply represent the unit vectors.

Training Strategies. The feature representation was done by a ResNet-50 as a backbone that was pre-trained on the ImageNet dataset. It uses binary masks in addition to image crops to focus on specific objects. Also, a better prior distribution was learned by using visual representations associated with the tasks. After the feature extraction, a lightweight stack of fully connected layers transforms the learned features into the log-transformed parameters of the Dirichlet distribution [20]. To account for varying response scales, the total number of observed answers for a task is considered as an additional input. The Chernoff distance, specifically the Bhattacharyya distance, was used as a loss function to minimize the difference between the predicted and actual distributions. To counteract imbalanced classification problems, class weighting was applied based on the relative frequencies of responses. The optimizer Adam with a learning rate of 1e-4 and a weight decay of 1e-5 achieved satisfactory results that were validated. For the training of this model, the subset of 6000 images was split into training and test data because the annotator answers are only available for this part of the data. In 2.5, the success probability and the solvability probability were predicted for one bounding box.

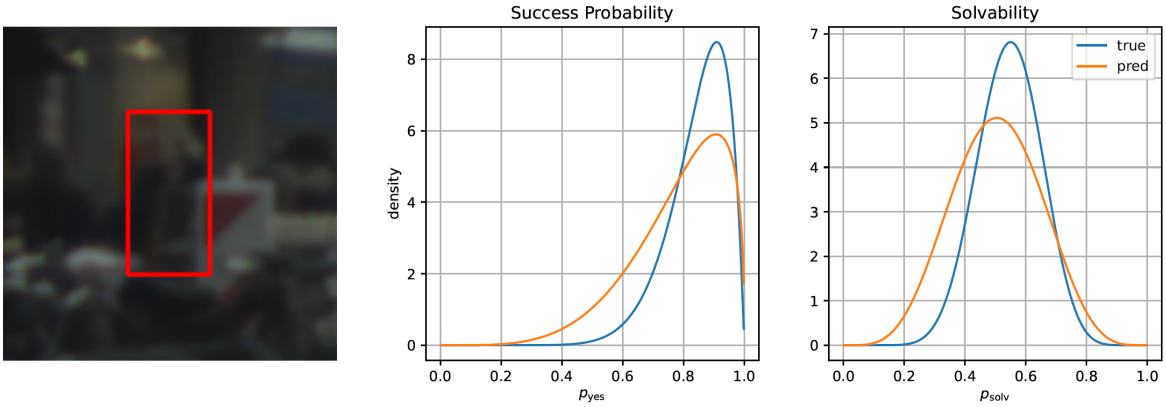


Figure 2.5: The distributions of success probability p_t and solvability probability π_t with the image crop. On the x-axis the success probability and the solvability are shown while the density is depicted on the y-axis. The blue distribution is the ground truth and the orange one the prediction [20]. Although the predicted distributions closely resemble the ground truth, they do not perfectly match the peak density.

2.7.2 Application

The trained general binary model was then applied to the whole training data and test data of the Eurocity Persons dataset. The input of the model was always the image and the coordinates of the bounding box around the object for which the Dirichlet distribution of the annotators' answers to the human being question is to be predicted. This means that the output is a vector of the three parameters of the corresponding posterior Dirichlet distribution for the responses "no", "yes" and "can't solve". Two measures were calculated using the results: the entropy and the ambiguity measure.

Entropy. The Dirichlet distribution of each bounding box or object is employed to model the uncertainty of the annotator answers. Entropy, a measure of uncertainty or disorder in the distribution, was computed using the following formula:

$$H(\alpha) = \log(\beta) + (\alpha_0 - C) \cdot \psi(\alpha_0) - \sum_{i=1}^{C} (\alpha_i - 1) \cdot \psi(\alpha_i), \tag{2.15}$$

where α is the set of parameters representing the Dirichlet distribution of one object, C is the total number of categories (here: 3), α_0 is the sum of parameters across categories, ψ is the digamma function applied to each parameter $\psi(\alpha_i)$ and the sum of parameters $\psi(\alpha_0)$ and $\log(\beta)$, where β is the beta function, which is computed as the difference between the sum of logarithms of gamma functions over parameters and the logarithm of the gamma function over the sum of parameters (2.16).

$$\log(\beta) = \sum_{i=1}^{C} \log(\Gamma(\alpha_i)) - \log(\Gamma(\alpha_0)) \tag{2.16}$$

Ambiguity Measure. The ambiguity measure was retrieved by repeating the α parameters of one bounding box 2^{14} times and sampling from the Dirichlet distributions represented by these parameters. The samples were generated from the standard gamma distribution and normalized afterwards. The number of occurrences for each category were obtained by sampling from the resulting multinomial distributions. For each realization, the ambiguity score (equations 2.4 to 2.6) could be computed which resulted in one ambiguity measure distribution per object. For simplification reasons, the median of this distribution was used as "the" ambiguity measure to evaluate and compare it to the entropy. A third very simple approach would be to just take the largest α parameter of each box and use this as a filtering measure. The downside of this idea is that there would only be the differentiation between "no", "yes" and "cs" and one would not be able to set different thresholds for finer division.

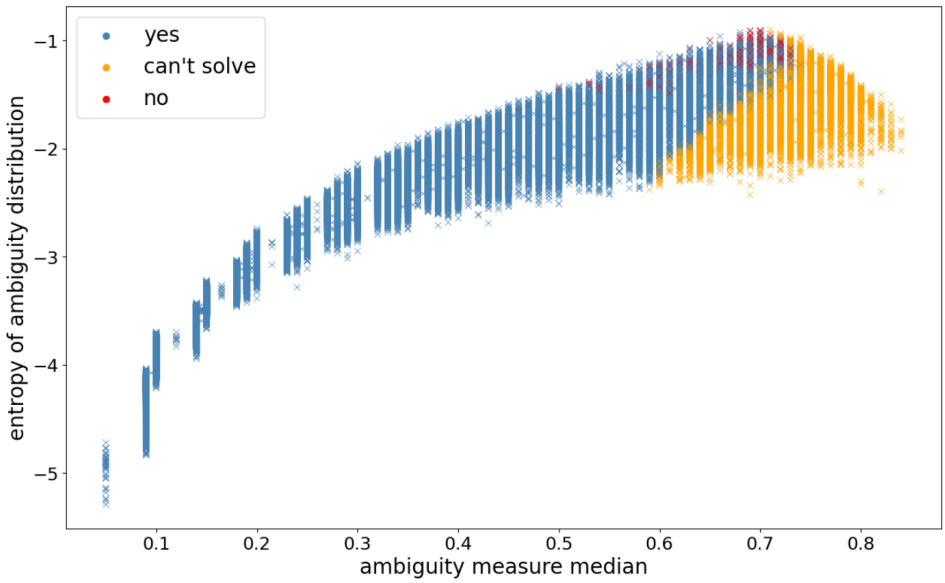


Figure 2.6: Correlation of entropy and ambiguity measure median of the whole ECP training data. On the x-axis the ambiguity measure median in comparison to the entropy of the ambiguity distribution on the y-axis. The largest of the three alpha parameters of the predicted Dirichlet distributions is highlighted with colour. Each marker represents a sample with corresponding values for both measures, displaying a correlation pattern that appears logarithmic. Even though the ambiguity measure can be low for the answer "no" (if all agree), most "no" instances have a high ambiguity score which means that there are not much samples in the data that annotators easily agree on "no human being". They rather agree on "can't solve".

Comparison of Ambiguity Measure and Entropy. In 2.6, the maximum parameter is still used as a validation for the ambiguity measure. One can see that there is a non-linear correlation between the entropy and the ambiguity median. Also, it exists a separation line between the samples with the largest alpha parameter "yes" and "can't solve", even though the markers overlap each other so it is not a very clear line. Neither the entropy nor the ambiguity median separate these samples clearly with a possible threshold but this was not required since the largest alpha parameter does not necessarily indicate the perfect split between ambiguous and unambiguous data. Both measures rank the objects with the largest alpha parameter "no" as very uncertain or ambiguous. Surprisingly, there are some samples that have a high ambiguity measure median but a medium entropy. When looking at the

interquartile range of the ambiguity measure distributions in 2.7, it is noticeable that most samples have a low ambiguity median and a small interquartile range. This has to do with the fact that the Dirichlet distribution and the ambiguity measure share certain properties. If a Dirichlet distribution has low entropy, then the discrete probability vectors with low ambiguity are drawn from a relatively narrow range. "Narrow" means that there is little variance, i.e. the interquartile range of the realised ambiguity is small. However, there is another case where this intuition can be reversed. A Dirichlet distribution can be very strongly concentrated around the "can't solve" responses. The majority of the drawn response distributions would have a large ambiguity measure median by definition. But the interquartile range would be very small again in this case because the Dirichlet distribution is still quite concentrated which explains the "arc" in 2.7. Besides that, one can see that most samples have a low ambiguity measure median which could look a little misleading in Figure 2.6. Also, some kind of quantisation of the ambiguity measure is visible which comes from the fact that the measure only takes on a finite number of values for a fixed number of observed responses.

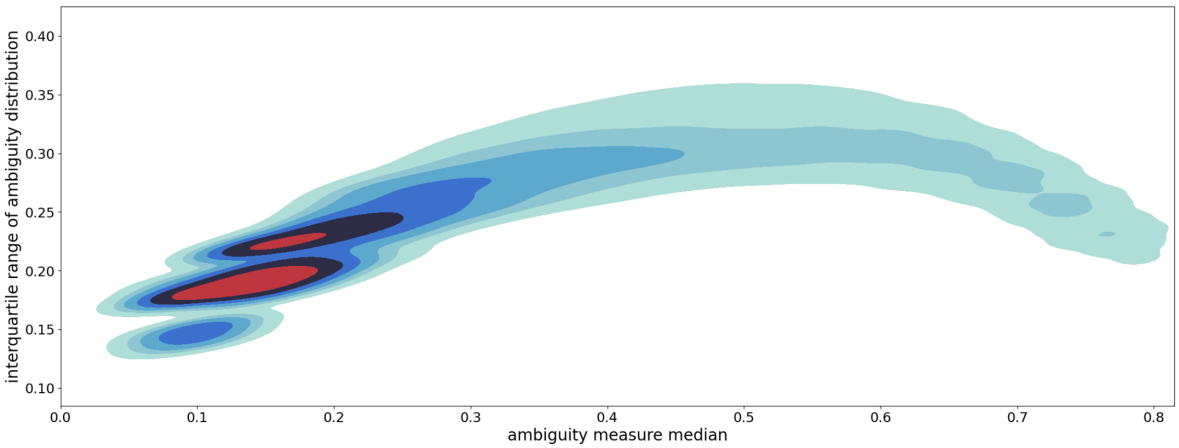


Figure 2.7: Correlation of the interquartile range of ambiguity measure distribution to the median. The density plot is coloured red for a high concentration of samples and light blue for a low density. Most samples have a low ambiguity measure median and IQR while there are samples that have high ambiguity score and low IQR. These are samples with a Dirichlet distribution concentrated around "can't solve" which is considered ambiguous by the ambiguity measure.

Impact of "can't solve". In summary, the entropy that is calculated from the model predictions expresses how large the variation within the crowd is across all permissible answers. The ambiguity measure median does something very similar, but with the crucial difference that the ambiguity indicates the "entropy" only with regard to the solvable answers (in this case "yes" and "no"). It treats the unsolvable cases separately. For example, the entropy can be low if the models predicts "can't solve" reliably. On the other hand, the ambiguity measure would deliver very high values in this case. In short, ambiguity score and entropy lead to a similar sorting of instances in cases where the model gives little weight to "can't solve". In cases where the model has a strong tendency to "can't solve", ambiguity measure and entropy lead to an almost reversed sorting of the data.

(a) $med_{amb} = 0.05$ **(b)** $med_{amb} = 0.15$ **(c)** $med_{amb} = 0.25$

(d) $med_{amb} = 0.35$ **(e)** $med_{amb} = 0.45$ **(f)** $med_{amb} = 0.55$

(g) $med_{amb} = 0.65$ **(h)** $med_{amb} = 0.75$ **(i)** $med_{amb} = 0.82$

Figure 2.8: Examples of ranking by ambiguity measure as the median of each ambiguity distribution from low to high. The persons marked by the orange bounding boxes are relatively clear from 2.8a to 2.8e, become much more ambiguous from 2.8g and are impossible to identify at 2.8i due to truncation, occlusion and lighting conditions.

"Can't solve" cases are considered ambiguous because the annotators either could not solve the task with respect to the current bounding box or were not sure whether to choose "yes" or "no". These cases in which not even a majority of people can make a decision might not bring any value to the learning process of an object detection model.

Investigation of the Ambiguity Measure. In 2.8, image crops of marked humans in the test dataset are sorted by the median of their ambiguity measure distributions in ascending order. The visual ambiguity of the samples rises with increasingly ambiguity measure. The higher the ambiguity measure, the more occlusion and truncation seems to be present. Also, the samples with a high ambiguity median are often in dark or poorly illuminated areas of

the images. When manually investigating the objects, another noticeable effect was that the height of the bounding boxes decreases with a higher ambiguity measure. However, 2.9 only partially confirms this assumption. The median height of the binned samples increases to around 75px at a higher ambiguity measure. Further investigation of the data showed that the there is a great amount of samples that are rated with a high degree of ambiguity but are also not small. These are mainly persons behind parking vehicles where only the head is visible or highly truncated instances. The orange and red line display the percentage of samples in each bin that are tagged either with an occlusion or truncation tag. For a low ambiguity measure, the proportion of truncation tags is almost zero while it increases to 2 to 3% for a higher ambiguity value. The occlusion begins with a low percentage and rises steadily up to almost 90% in the last bin, confirming the statement that these are mainly concealed people behind vehicles. To summarise, Figure 2.8, 2.9 and manual data investigation proved that the ambiguity measure is successfully sorting the data according to its ambiguity and will therefore be used to perform splits at different thresholds of the training and test data.

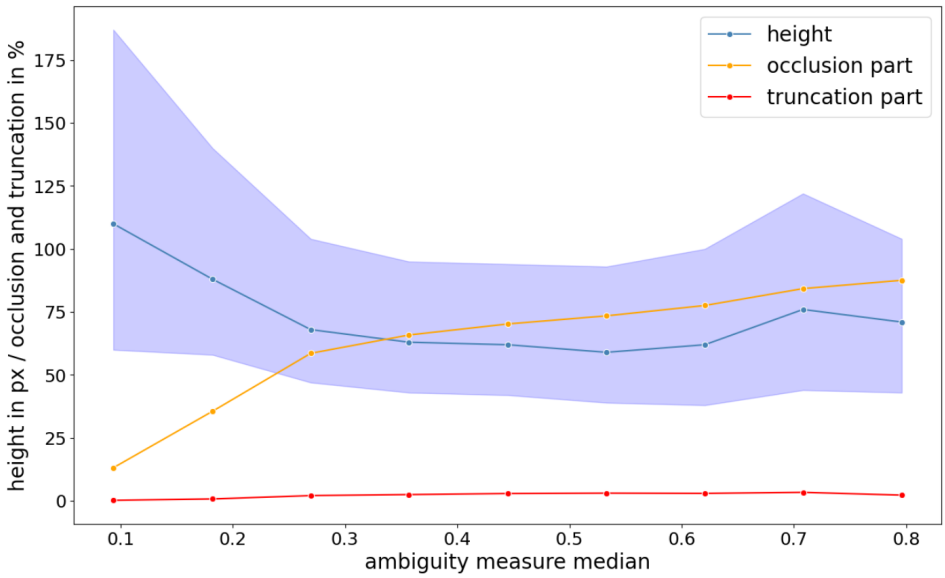


Figure 2.9: Correlation of ambiguity measure, bounding box height and proportion of occlusion and truncation tags in the ECP training data. The ambiguity measure median is depicted on the x-axis and the height and occlusion/truncation tag proportion on the y-axis. The height decreases to an ambiguity value of 0.6 and has a small peak at around 0.7 which could indicate highly truncated instances that are ranked with a high ambiguity measure but are not necessarily small or persons occluded by vehicles. Also, there is a noticeable, almost linear correlation between the ambiguity score and the part of occlusion tags which shows that samples with an occlusion tag are more likely to be ambiguous.

3 Results

The investigation into the impact of ambiguous data on state-of-the-art pedestrian detection models has yielded multiple results in different areas. This chapter presents an analysis of these results obtained from various trainings and evaluations on different datasets which will be interpreted in the next chapter.

3.1 Improving Model Performance through Ambiguity Analysis

First experiments were performed on the subset of 6000 images taken from the original training data whose annotations were improved and corrected by the false positive and false negative pipelines at Quality Match. Three corrected datasets were derived as previously shown in table 2.3.

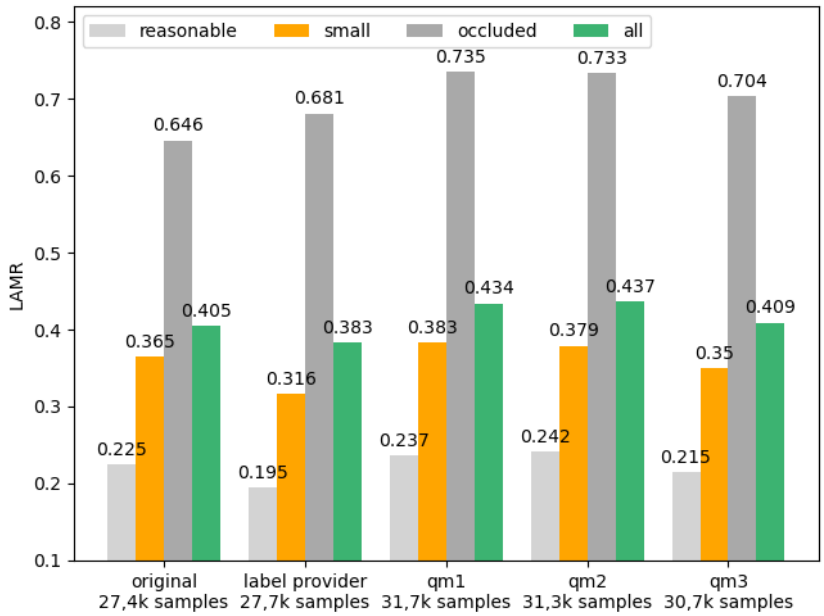


Figure 3.1: Comparison of the LAMR on the test data between models trained on different annotations of the 6000 images subset. The models with the additional information of training data sample size are on the x-axis. Each model has the LAMR calculated for every ECP evaluation subset. The qm3 model outperforms the original annotations in every category except "occluded", most likely due to the lack of occluded data in the training set which was filtered out. For humans, there also seems to be a correlation of occlusion and ambiguity.

In total, five models were trained using these three datasets, the original labels and the labels of the label provider which formed the basis for the data quality improvement with the usage of additional annotators. The log average miss-rate was calculated for all four ECP evaluation subsets of the test data: reasonable, small, occluded and all.

Filtering by Annotator Estimations. In 3.1, one can see how the increasingly stricter filtering of samples in the datasets qm1 to qm3 affects the number of samples existing in each dataset. The qm1 dataset included around 1000 more samples than qm3. Furthermore, the LAMR of all evaluation subsets significantly decreases from qm1 with more ambiguous samples in the training data to qm3 containing less, even though the values of "reasonable" and "all" experience an increase of 0.005 and 0.003 when comparing qm1 to qm2. The difference between the datasets qm2 and qm3 is particularly noteworthy. Apparently, it also useful to prune samples with the opposite aggregated answer for which the decision was very uncertain. This can be expressed by a high credibility or low convergence. Despite the improvements of the log average miss-rate along filtering the datasets with theoretically improved quality by Quality Match, the raw, not processed dataset delivered by the external label provider has the best LAMR of all datasets for all categories except the category "occluded" which is predicted best by the model trained on originally annotated data from ECP. It is also important to note that the qm3 model performs better on the evaluation subsets "reasonable" and "small" than the original dataset, according to the LAMR.

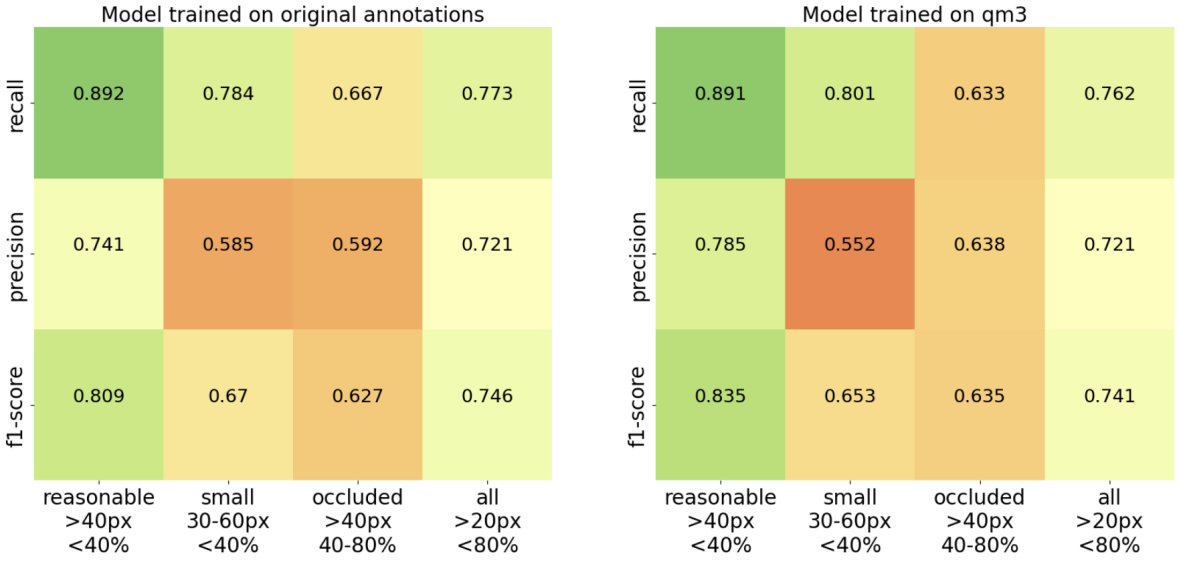


Figure 3.2: Comparison of recall, precision and f1-score between the model that was trained on the original annotations and the one trained on the qm3 dataset. The measure values are presented in each cell for the associated subset and coloured by the performance. Green stands for a high percentage which is good and vice versa for orange. There is a trade-off between recall and precision, when one model has a better recall, the other one has the higher precision. The "reasonable" subset contains little ambiguous data due to the recall difference being small.

A More Detailed Evaluation. Since the log average miss-rate is a quite generous measure and the log-log plots with the miss-rate and the false positives per image are not well readable, the recall, precision and f1-score were plotted in Figure 3.2. While the recall is a measure that focuses on finding all ground truth samples, the precision answers the question of how many of the predicted boxes were true and could be matched with ground truth boxes. The f1-score combines the recall and precision to make a statement about how effectively the model makes the trade-off between the both:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3.1}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3.2}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.3}$$

The objective of these heatmaps was to delve deeper into the types of errors made by the models and to identify potential discrepancies between them. The precision is calculated for all false positive boxes that fall into the height range of the corresponding evaluation subset, e.g. in "small", only false positive boxes with a height between 30 and 60 pixels are counted. Boxes that are matched with any ground truth box in the whole test dataset are not considered as errors, even if they could not be matched within the actual evaluation subset. This could be the case if there is a predicted box with a height of 40px in the "small" subset but it could only be matched with a ground truth box that has an occlusion tag for more than 40% which is not in "small". Therefore, all external ground truth boxes that are not within the actual subset are treated as ignore regions like neighbouring classes or persons with the tags "behind glass" or "sitting-lying". Moreover, the additional information on the x-axis in percent always refers to the ground-truth tags for occlusion and truncation. The only category where the percent range only refers to the occlusion is "occluded" because truncation tags are not allowed in this subset.
Comparing the different measures of the models in 3.2, one can see that the model trained on the original training data has an overall better f1-score for the subsets "small" and "all". Interestingly, the qm3 model achieves a better recall but a worse precision for "small". In the evaluation subset "all", they have the same precision but the original model was able to detect more pedestrians. For the "reasonable" subset where both models are similarly good in identifying pedestrians, the qm3 model produces less false-positive errors and in "occluded" the same holds but the difference in recall is greater.

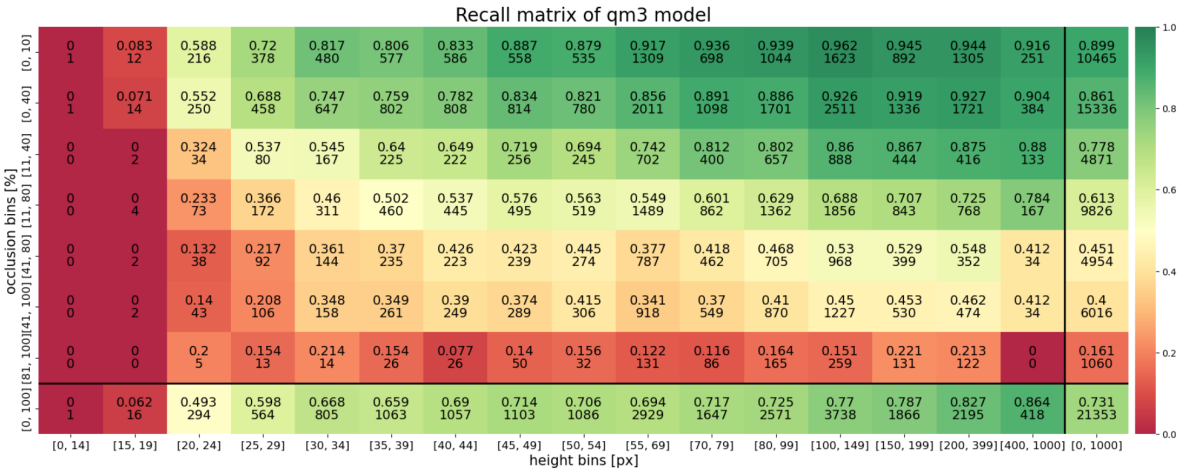


Figure 3.3: The recall of the model trained on the qm3 dataset with boxes binned by height on the x-axis and boxes binned by occlusion level on the y-axis. The upper value is the recall and the lower value in each cell represents the number of samples it contains. Separated by the 2 black lines, the recall of the entire data regardless the occlusion or height can be viewed. The model is better at detecting taller objects and those with minimal occlusion. However, it struggles notably with occlusion levels exceeding 80%.

The most unmatched boxes and therefore the lowest precision has the "small" evaluation subset which suggests that both models especially produce small false positive error boxes. The overall difference in f1-score is highest for the subset "reasonable" and then descends to the right for "small", "occluded" and "all". Because of this, the correlation between the error that happens, the height of the boxes and the amount of occlusion or truncation had to be further explored.

Detailed Recall Comparison. In 3.3, the distribution follows an expected pattern. Note that bins containing zero samples are also coloured red and boxes can occur in multiple bins by occlusion since they overlap partly. The bins with a really low height and high occlusion tags, especially greater than 80%, reach the lowest recall values, therefore only a very small part of the boxes in these bins were detected by the model. In contrast to this, larger pedestrians with a low level of occlusion or even no occlusion at all are predicted much better and some bins at the top right of the heatmap even reach the 90% mark. This Figure of the recall looks similar for the other 4 trained models despite slightly lower or higher values. When comparing this to the recalls of the model trained on original ECP annotations, there exists a pattern which confirms the observations of 3.2. The differences between the recall of each bin of the qm3 model to the original model are plotted in 3.4. There is a strong tendency that the original model detects larger and lightly to strongly occluded boxes much better. On the other hand, it gets outperformed on the smaller boxes by the qm3 model which can also be seen in the recall difference for the subset "small" in 3.2. As the ECP evaluation subsets differ in height of the boxes taken into account, this also explains the higher recall of the original model for the subsets "reasonable" and "occluded".

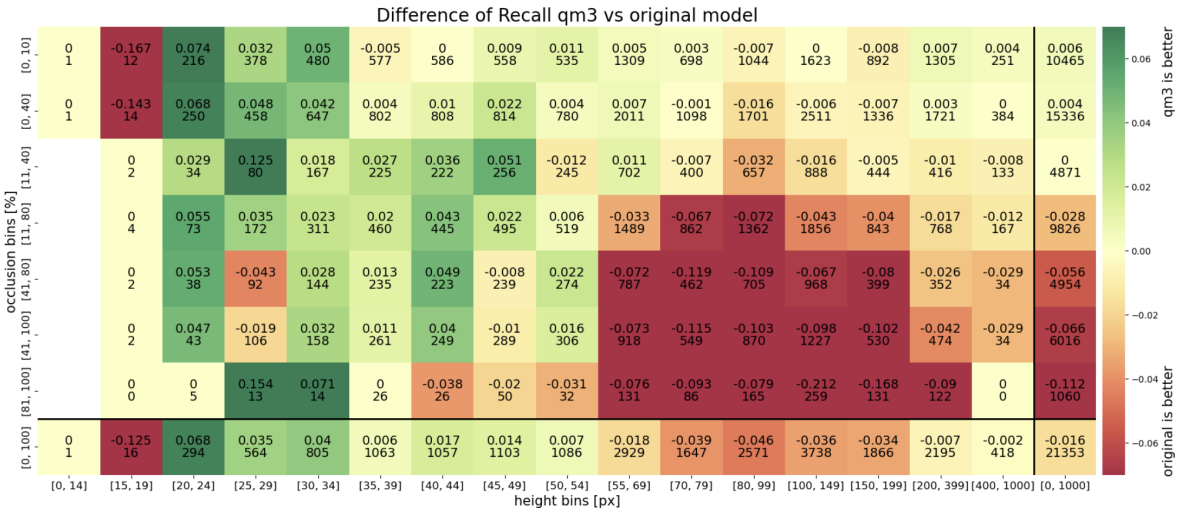


Figure 3.4: The difference of recall between the qm3 model and the model trained on the original annotations with boxes binned by height on the x-axis and binned by occlusion level on the y-axis. The lower value in each cell represents the number of samples it contains. Separated by the 2 black lines, the recall difference of the whole data regardless the occlusion or height can be viewed. The original model shows higher recall for boxes with significant occlusion due to the training data containing more ambiguous data.

Detailed Precision Comparison. While the recall plot is mostly dominated by the original model, it is the other way around for the precision. Figure 3.5 shows that the qm3 model produces much less large false-positive boxes compared to the original model. In contrast, the qm3 model demonstrates inferior precision for smaller boxes compared to its counterpart. Since small boxes predicted by object detection models are often not even considered in the evaluation and the ground truth boxes of the test dataset are mostly starting at around 35 pixels height, these errors done by the qm3 model could be considered less worse than falsely predicting large boxes. Again, the distribution of error differences supports the values in Figure 3.2 where qm3 has a higher precision for the subsets "reasonable" and "occluded" in which the boxes are larger (>40px). The opposite applies for "small" with smaller boxes (30-60px).

Conclusion and Next Steps. To summarise, one can see a trade-off between the two models when it comes to recall and precision at different box sizes and occlusion/truncation levels. When one model has a higher recall, it performs worse for the precision and vice versa. The general recall, except for small boxes, still seems to be better when training with the original ECP annotations. This could have multiple reasons. At first, the training only happened on one fourth of the total training data available because the data correction was only done for this part of the data. It could be that the differences of using higher quality data only become visible when training with more data. The subset of 6000 images is quite small for an object detector of this size which was also not pre-trained on any other dataset like it is usually done [13]. It is also possible that the training was simply to short to reveal major recall differences. The most likely reason for the results is the quality and structure of the test data. Remembering that the quality of the training data was improved by additional filtering of ground truth boxes (FP analysis) and adding of missing boxes (FN analysis) but the test

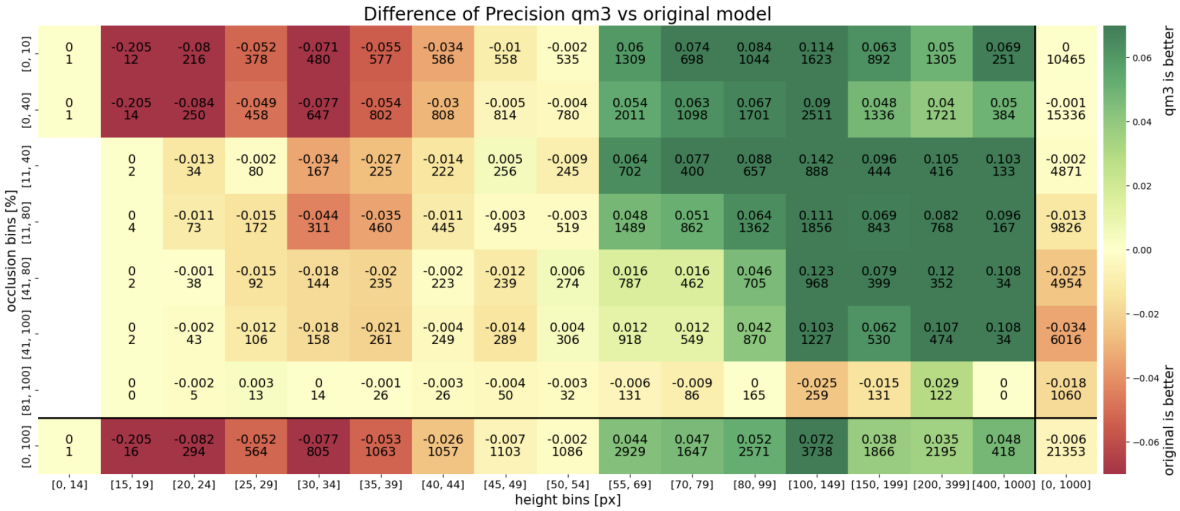


Figure 3.5: The difference of precision between the qm3 model and the model trained on the original data with boxes binned by height on the x-axis and binned by occlusion level on the y-axis. The lower value in each cell represents the number of samples it contains. Separated by the 2 black lines, the precision difference of the whole data regardless the occlusion or height can be viewed. The qm3 model has a much better precision for boxes above 54 px height which matters more than small boxes due to the subset filtering of ECP. This indicates ambiguous training data to cause more small FP boxes in particular.

data still contains the original ECP ground truth annotations. These are, just like the training data, possibly ambiguous and can therefore have a great effect on the evaluation results. To evaluate and possibly prune the test data, the general binary model was used to estimate the ambiguity of the samples and express it with the ambiguity measure. Because the general binary model is trained on the answers of humans to questions like "Is the object in the box a human being?", the human ambiguity that exists in the ground truth data because it was labelled by humans [12], is then again measured by humans. This time, it is a model that learned to answer these questions like humans would most likely do it. To counteract the assumption of training with not enough data, the ambiguity measure was also applied on the entire training data which consists of approximately 24000 images. The training length of 50 epochs was kept due to time-limiting reasons.

Ambiguity Measure Threshold. Finding the right ambiguity measure threshold for the training and test data can be a difficult task because there is no universally correct answer that applies to every dataset and use case. Figure 2.6 showed that most samples where the highest alpha parameter is "no" for the human being question lie above an ambiguity measure median of 0.6. The orange samples with "can't solve" as their largest parameter can not be clearly split by the ambiguity score. These samples exist in a range from 0.6 to the maximum of 0.83. Even Figure 2.9 could not suggest a good split for the ambiguity measure. While the median height of boxes shows a slight increase around 0.6, relying solely on height as a decision criterion is problematic due to truncated samples at the image border which can still be large. Also, it is often the case that only the head of a person standing behind a parking vehicle at the edge of the road is visible. These cases lead to a higher ambiguity measure value but are not necessarily highly ambiguous.

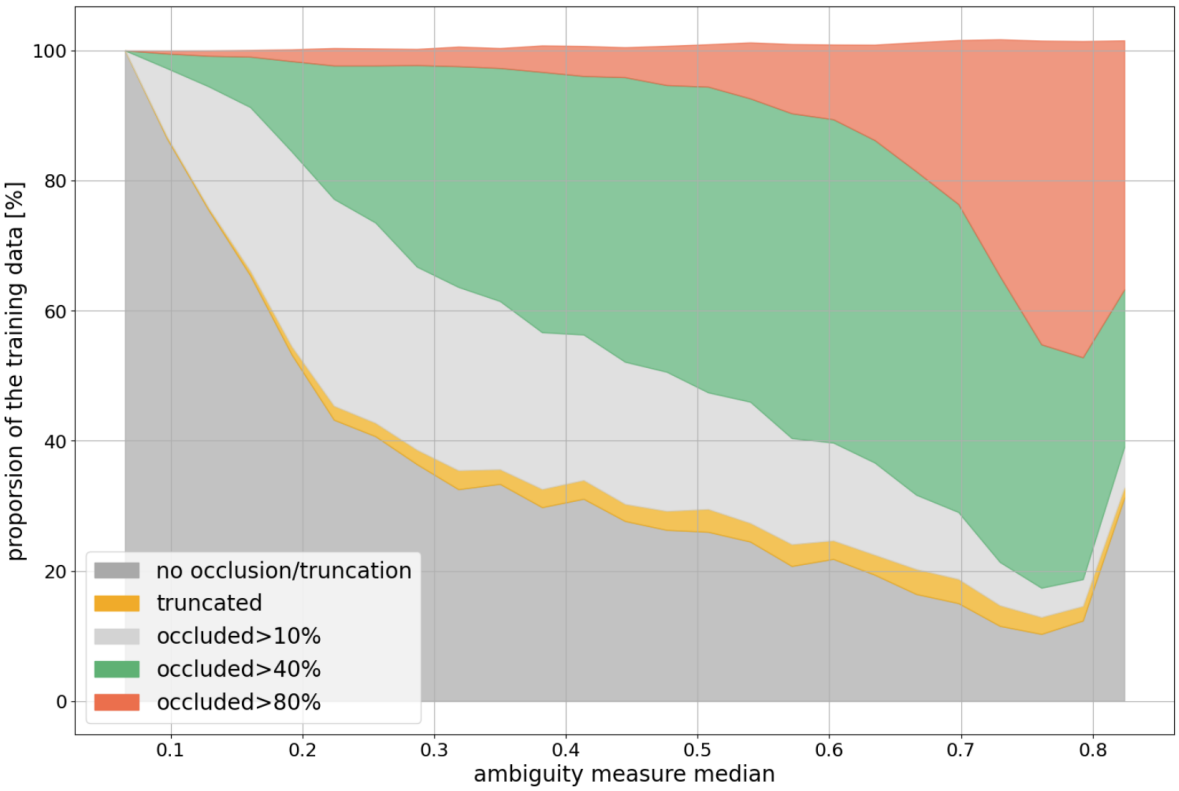


Figure 3.6: Distribution of occlusion and truncation tags of samples binned at different values of the ambiguity measure. For simplicity reasons, the truncation is only shown as one area and not split by percentage like the occlusion. The proportion of high occlusion in the data raises with the ambiguity score of the samples. A correlation between ambiguity measure and occlusion is visible with a decrease at high ambiguity values. The reason for this could be samples with missing occlusion/truncation tags.

This is a complexity of the ambiguity measure calculated from the predictions of the general binary model even though these cases seem to be challenging for annotators as well. The overall partition of boxes with an occlusion tag rises constantly in 2.9 which is also not a good indicator of where to do the ambiguity measure split for the training data.

Occlusion Tags and Ambiguity. That is why a more detailed distribution of the occlusion tags in different bins of the ambiguity measure median was plotted in 3.6. One can see that the proportion of samples without any occlusion or truncation tags decreases with an increasing ambiguity measure value and has a peak at a high ambiguity. The tag "occluded>10" forms a small part at a low ambiguity median which rises to around 25% at 0.3 and then decreases again for higher ambiguity values. Samples with the tag "occluded>40" make a much greater part of the area plot with approximately 50% at an ambiguity score of 0.6. Their proportion also decreases to the right but not as much as the 10% tag. The highest occlusion tag of greater than 80% has a very clear distribution which raises very slowly with increasing ambiguity measure median until 0.6, then rapidly increases to almost 50% with a peak at around 0.79. The overall sum of percentages can be more than 100% because there are 578 bounding boxes that have both, a tag for occlusion and truncation, which is counted double in this plot. These samples mostly occur for a high ambiguity score which can be seen at the line crossing the 100% mark.

Finding an Ambiguity Measure Threshold. Showing that there is a correlation between the occlusion and the ambiguity measure, one can infer that as the ambiguity measure increases, indicating higher uncertainty of annotators and more visual difficulties, the likelihood of occlusion also tends to increase. The partitions of occlusion tags for higher percentages of occlusion take over the other ones. This suggests that more ambiguous samples are often associated with higher levels of occlusion which aligns with the intuitive understanding of ambiguity in automotive image scenes. Extensive manual inspection of the data and the usage of information derived from the previous plots, like the rapidly increasing portion of "occluded>80%" tags at a certain point, lead to splitting the training data at an ambiguity measure threshold of 0.65. This excludes all data from the training that has a higher or equal ambiguity score of 0.65 which corresponds to a reduction of samples of around 8.6% in the whole ECP training dataset. Not only the training data but also the test data contains ambiguous data that can now be pruned with the ambiguity measure by choosing a threshold. Intentionally, the same threshold can be taken like for the training data.

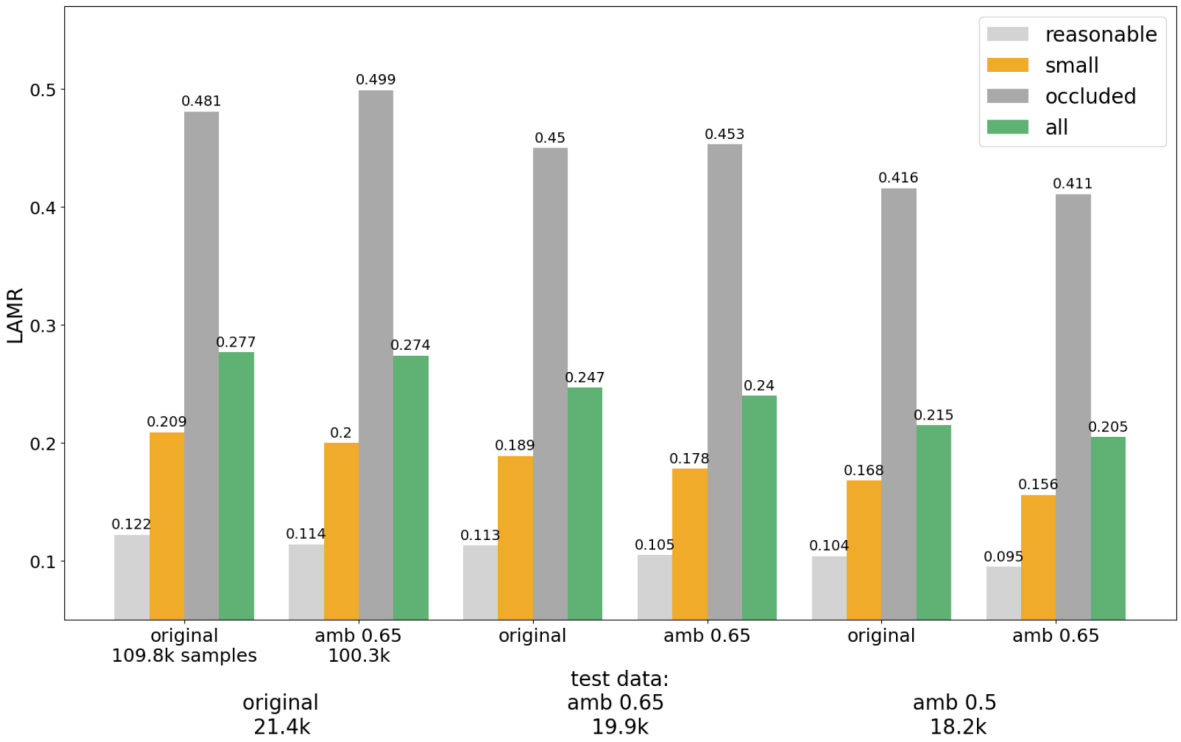
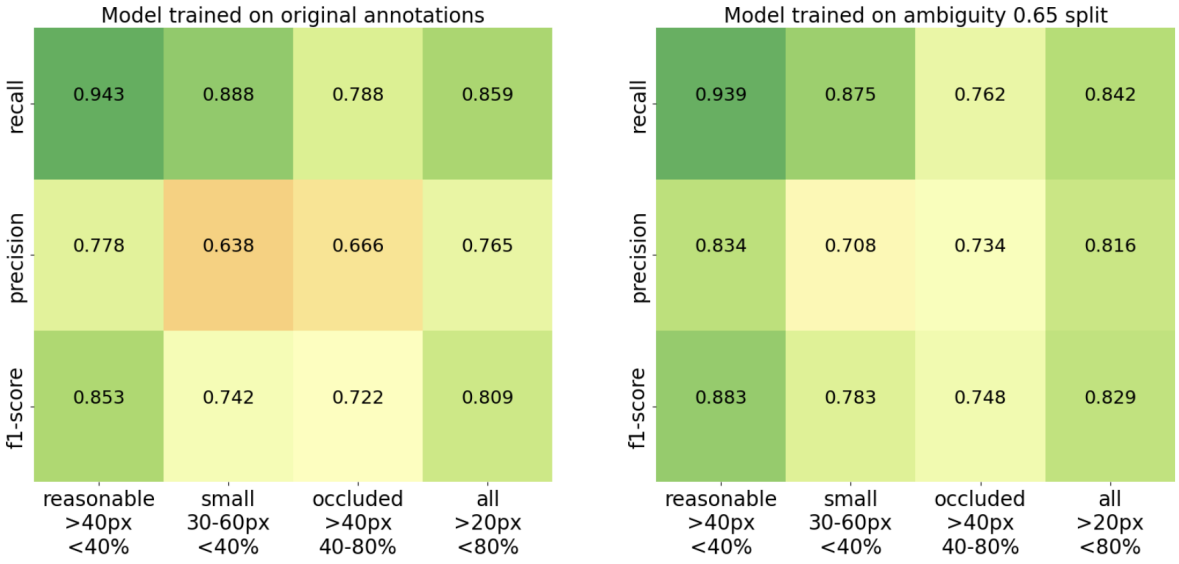
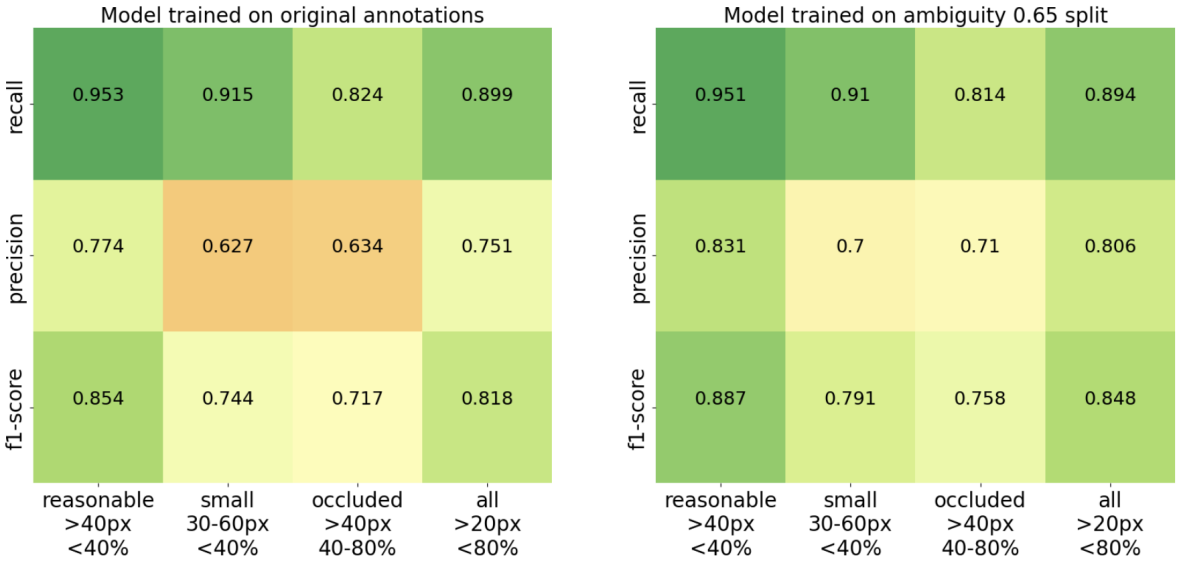


Figure 3.7: Comparison of two models, one trained on the original whole training data and the other on pruned data by ambiguity threshold. The LAMR is calculated for three test datasets on the x-axis: the original one and two pruned by ambiguity measure threshold like the training data. The contained number of samples is shown below the dataset names. The model trained on thresholded data outperforms the original one, even for the subset "occluded" when excluding all ambiguous data. Most very ambiguous data is excluded from the "occluded" subset at an ambiguity threshold of 0.5 which leads to this alignment.

Training and Evaluating with Thresholded Data. In 3.7, a second harder threshold of 0.5 was done for the test data to ensure that all of the ambiguous data is excluded, even when this also possibly includes some more clear pedestrians. The filtering of the test data results in two datasets with 7% (amb 0.65) and 15% (amb 0.5) reduced samples. A first general trend can be seen in the bar chart: Excluding ambiguous samples from the test data, in fact making them ignore regions like neighbouring classes, lowers the log average miss-rate of both models in each category. It would not make sense to simply delete them from the ground-truth because the models would be punished for detecting ambiguous data for which it should not matter what they do when even humans cannot decide if it is a human or not. The generally lower LAMR makes sense since these are the harder instances to predict and also shows that both models make more mistakes with ambiguous data than predicting it correctly. Interestingly, the model whose training data was pruned with the ambiguity measure performs better in each evaluation subset except "occluded". By excluding ambiguous samples, this model was trained on a cleaner dataset leading to improved performance overall. However, in the "occluded" subset, the exclusion of highly ambiguous data from the training data may inadvertently remove crucial examples that help the model learn to deal with occluded objects. This changes in the last test dataset with the stricter ambiguity threshold where the original model begins to underperform. There seems to be a shifting impact of the ambiguity measure on the "occluded" subset that arises from the correlation of occlusion and the ambiguity score. This suggests that while filtering out ambiguous data is beneficial for generalization of the other three evaluation subsets, there is a trade-off when it comes to handling occlusion. Another remarkable point is that the differences between the LAMR's for each of the other three categories are getting larger when pruning more ambiguous data. For example, the difference for "all" in the original test data is only 0.003 while it increases to 0.01 in the ambiguity 0.5 test dataset. The reason for that is not the logarithmic nature of the log average miss-rate but the changing dataset composition resulting from data pruning which is also proofed by comparing Subfigure 3.8a to 3.8b, specifically by examining the difference in f1-scores between the models across each subset. "All" has a difference of 2% f1-score in 3.8a for the original test data while it is 3% in 3.8b which makes a huge difference. The original model still has a better recall in all subsets but the difference of recall between the two models decreases when pruning ambiguous data from the test data. This can be explained by the fact that the original model was trained on more ambiguous data which it therefore also learned to detect. When making this data ignore regions by pruning with the ambiguity measure, the model gets no longer rewarded for detecting these ambiguous samples. Again, the precision is the main factor of the f1-score difference between the two models because it is significantly higher for the ambiguity threshold model. Each category has a difference of at least 5% between the two models in 3.8a which also increases in 3.8b.



(a) Original test dataset



(b) Test dataset pruned by ambiguity measure at 0.5

Figure 3.8: Comparison of recall, precision and f1-score between the model that was trained on the whole original annotations and the one trained on the pruned ambiguity 0.65 dataset on two test datasets. On the x-axis, the ECP evaluation subsets and the measures on the y-axis associate as one value in each cell, coloured green if close to 1 and red when reflecting a low percentage. As the proportion of ambiguous data in the test dataset decreases, the difference of recall between the two model diminishes due to less difficult ambiguous samples contained in the test data. Both test datasets indicate a negative influence of ambiguous training data on the precision.

Summary. In conclusion, there are two main effects when evaluating on less ambiguous data: the recall differences decrease and the precision differences increase. Both of these changes favour the model with pruned training data by ambiguity measure, even though it can not overcome the recall of the original model.

Summarizing this chapter, the first experiments, where ambiguous data was pruned and missed boxes were added to the training data, showed an overall better LAMR of the model with a higher training data quality. When inspecting more details of the recall and the precision, there was a trade-off between those two measures visible. Depending on the evaluation

subset, one measure was better for one model while the other one was worse. It was shown that detecting larger boxes with less occlusion is easier than smaller boxes with high levels of occlusion or truncation and that the ambiguity measure has a strong correlation with the proportion of occlusion that is inherent in the data. Apparently, only dealing with ambiguous samples in the training data is not enough as results changed when also applying the ambiguity measure on the test data. It changed the evaluation drastically as the recalls almost lined up and precision was much better for the model trained on data split by the ambiguity measure. Therefore, it was shown that ambiguous samples contribute to the generation of more false positive boxes by the model trained with them.

3.2 Exploring the Impact of Neighbouring Classes

In section 3.1, results have shown that removing ambiguous data from the training data leads to a significant improvement of the precision. On the other hand, the recall decreases slightly even though the difference becomes smaller when also pruning ambiguous data from the test dataset. This led to the hypothesis of receiving a higher recall when adding additional ground truth data to the training data, a notion supported by the results from the 6000 image subset. In this subset, not only ambiguous data was pruned but also 4942 false negative bounding boxes were added to the training data which resulted in the qm3 model being able to detect more samples in the "small" evaluation subset and nearly as many as the original model in the "reasonable" subset (see Figure 3.2). Apparently, the false negative ground truth analysis was only done for a part of the training data and an alternative approach had to be established to identify missing ground truth pedestrians.

Adding Ground Truth Annotations. In the ECP evaluation, groups of people that cannot be clearly separated are labelled as "person-group-far-away" and the class "rider" is defined as neighbouring class of "pedestrian" because there often is only a small difference between these two, e.g. standing with one foot on the ground or not. Bounding boxes labelled with "person-group-far-away" or "rider" are definitely including humans, if they are not very ambiguous. Potentially, the model can also learn the visual representation of pedestrians from these objects. Excluding boxes with the tag "depiction" (e.g. large poster) and "reflection" (e.g. in store windows) [12], the original dataset size grows to 147.6k samples. The ambiguity measure was applied to the extended dataset to filter out very ambiguous samples with a threshold of 0.75. The threshold was chosen higher than in the last section to include as much additional information as possible that could help the model generalize better.

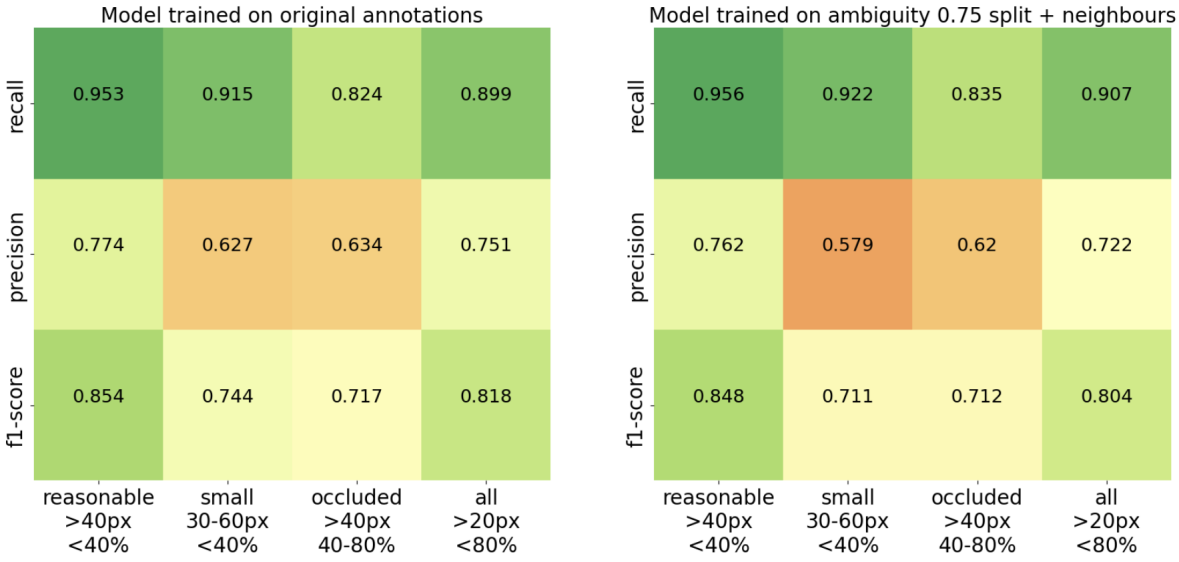


Figure 3.9: Comparison of recall, precision and f1-score between the model that was trained on the original annotations and the one trained on extended data by neighbouring classes split by the ambiguity measure at 0.75. Ambiguous samples have also been pruned from the test data with a threshold of 0.5. The measure values are presented in each cell for the associated subset and coloured by the performance. Green stands for a high percentage which is good and vice versa for orange. The model trained with more samples from neighbouring classes is able to detect more pedestrians as such at the cost of a lower precision due to additional knowledge about how humans appear during training, even if this information comes from neighbouring classes.

After filtering by the threshold and pruning samples with the largest alpha parameter equal to "no" (human being question), 136.9k samples are left in the training dataset. The ambiguity score reduced the number of objects by 7.2%.

Measure Impact. Figure 3.9 compares the model trained on extended data to the one trained on original annotations with 27.1k less objects in its training data. One can see an improvement of the recall in each category, which is specifically high for "occluded". The recall did only improve by 0.3% for the "reasonable" subset suggesting that the model primarily improved in predicting occluded or ambiguous data. Details of this distribution are visible in 3.10, where most green bins with a positive difference for the model split by ambiguity score concentrate at the bottom left and therefore have a low height and high occlusion level. Adding boxes from neighbouring classes to the training data also had a negative effect. The precision is much lower compared to the original model. Each category experienced more than 1% additional false positives up to almost 5% for "small".

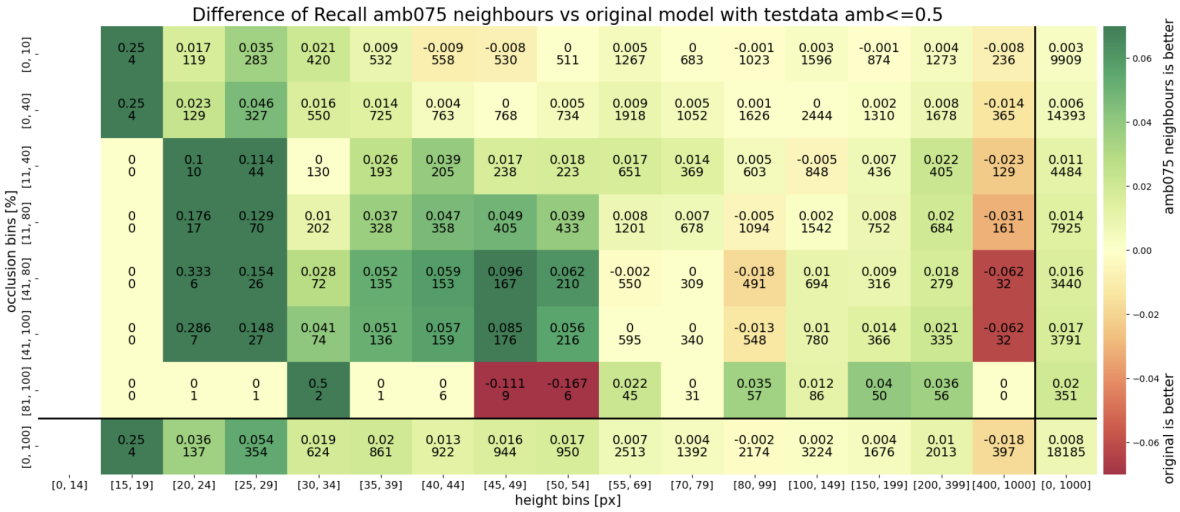


Figure 3.10: The difference of recall between the original model and the one trained on extended data by neighbouring classes split by the ambiguity measure at 0.75 with boxes binned by height and occlusion. The lower value in each cell represents the number of samples it contains. Separated by the 2 black lines, the recall difference of the whole data regardless the occlusion or height can be viewed. Smaller and more occluded boxes are detected better by the model with additional training samples. Interestingly, it struggles for tall objects because large, highly truncated boxes were pruned from the training data with the ambiguity measure.

Conclusion. In this experiment, the impact of samples from neighbouring classes on the performance of the pedestrian detection model was investigated. Interestingly, when additional ground truth annotations were added to the training data, there was a noticeable improvement in recall, especially for occluded or ambiguous instances. However, this improvement came at the cost of reduced precision. False positives increased across all categories. Further interpretations will be made in the discussion chapter.

3.3 Analysing Model and Annotation Errors

In the last section, the effect of adding ground truth objects from neighbouring classes to the training data was analysed. Except filtering by ambiguity measure, all experiments concentrated on the training data and how its quality can be improved. Manual investigations while pruning the test data with the ambiguity score and comparing the errors of the models showed that - (i) there are pedestrians on the images that are missing bounding boxes; (ii) some objects have a real high level of ambiguity so that they should not be part of any test dataset; and (iii) a lot of occlusion tags are missing or wrong. Figure 3.11 shows image crops of objects that are occluded but do not have a tag in the ground truth data. The original ECP paper claims an annotation error of $\leq 1\%$ for missed and hallucinated objects within the annotated number but does not make a statement about the accuracy of the tags. Besides the already existent ambiguity, these discoveries could also have a great effect on the results which is why they are examined in this section.

(a) (b) (c)

Figure 3.11: Image crops of objects without any occlusion or truncation tag in the ground truth data. All of the three pedestrians obviously are occluded and would need an occlusion tag. This falsifies the evaluation, e.g. due to more difficult objects being in the "reasonable" subset that should be in "occluded".

Error Distribution. Comparing the model trained on all original ECP annotations and the model with the data pruned by an ambiguity threshold of 0.65, they made different mistakes predicting the test data. Referring to the original test data, the original model had an overall better recall which results in less false negative errors and a worse precision what means that it produced more false positive boxes than the other one (see Figure 3.8a). The absolute numbers that also reflect this trade off can be seen in 3.1.

model	FP boxes	FN boxes	TP boxes
original	2929	3001	18352
amb 0.65	1982	3652	17701

Table 3.1: Absolute number of errors and correct detections done by the original and ambiguity 0.65 model. The columns describe false positives, false negatives and true positives. The trade off of recall and precision between the two models is also visible. While these differences appear substantial across the entire test dataset, subsets undergo filtering for the evaluation, impacting the total numbers.

Annotation Tasks. To gain insight into the correctness of these errors, annotators were asked to evaluate different properties of the ground truth boxes causing the false negatives and the detected boxes causing the false positive errors:

- Is the object in the box a: - human being, street sign, vehicle or other
- Can you identify the object inside the box? - yes/no
- Is the box drawn correctly around the person (including any occluded parts)? - yes/no
- only FN: To what extent is the object in the box occluded? - different occlusion levels

The initial three questions were posed for each box, whereas the final question was exclusive to the false negative errors, facilitating comparison with the occlusion tags in the ground truth. Each task-object pair received responses from a minimum of 5 to a maximum of 10 annotators with the option to select "can’t solve" provided, similar to the questions posed for the 6000 image subset.

Occlusion Tags. The question concerning the level of occlusion of the objects inside the bounding boxes provided four answers: "not occluded", ">10% occluded", ">40% occluded", ">80% occluded". The answers refer to the occlusion and truncation tag options that are in the ground truth of the original annotations in the ECP dataset and can therefore be directly compared. Since the question posed did not differentiate between occlusion and truncation, annotators likely marked objects that have a high truncation in the ground truth data as highly occluded. In 3.12, this was taken into account by - (i) extending the filtering of "not occluded" to also expect that no truncation tag is existent; and (ii) allowing the ground truth to either have an occlusion or truncation tag, for example "occluded>10%" can also be "truncated>10%". Despite the relatively low overall proportion of truncation tags (see Figure 3.6), they were more prevalent among the errors of the models and had a great impact on the category "not occluded". Without considering the truncation tags, the value for annotator aggregations "occluded>80%" and ground truth "not occluded" was 5.4% due to highly truncated pedestrians without occlusion tags. The percentages in 3.12 do not necessarily add up to 100% in any row because not all samples reached a clear decision for one of the occlusion tags, e.g. some had the same number of answers for "occluded>10%" and "occluded>40%". These undecided objects were left out of the calculation and make up the last part to 100%.

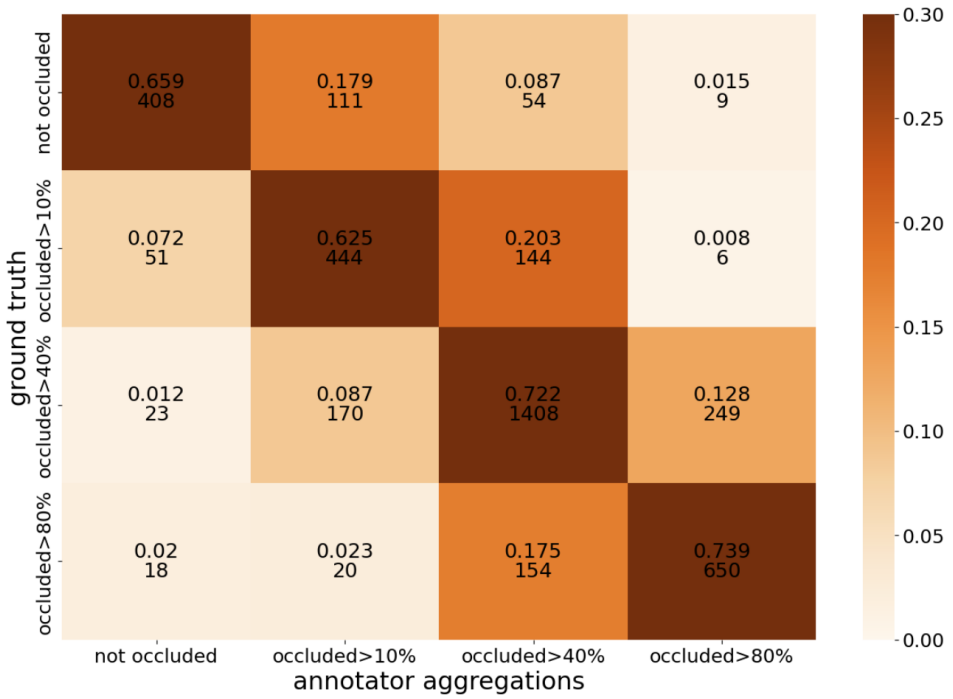


Figure 3.12: Comparison of occlusion tags between the ground truth on the y-axis and the aggregated annotator answers on the x-axis. Each row adds up to the whole ground truth samples tagged with the corresponding y-label. The upper value is the proportion and the lower the absolute sample count. For instance, 8.7% (54) of the samples without any occlusion/truncation tag are tagged as "occluded>40%" by the annotators (first row, third column). The percentages in each row may not sum up to 100% because some samples did not receive a clear consensus on their occlusion level among the annotators. There is a significant deviation toward neighbouring tags which diminishes for more divergent tags.

The false negative errors of both models, the original one and the one trained with an ambiguity threshold of 0.65, were matched and utilized in generating this heatmap. One can see that there is a deviation from the vertical line where ground truth and annotators are the same opinion. Most samples for which annotators have chosen a different category are close to the original choice. The deviation of the answers occurs in both directions, both in the direction of higher occlusion and vice versa. On average, 25.2% of samples deviate from the ground truth according to the annotators while the deviation slightly decreases with a higher level of occlusion. The tags of the error samples were rectified according to 3.12 but only resulted in small changes of the log average miss-rate. Nevertheless, these influences should not be ignored. They add additional ambiguity to the data since the tags decide to which evaluation subset one sample belongs and filtering by ambiguity measure can produce different quantities depending on the error rate of the tags.

Analysis of FP Errors. The posed questions enabled further filtering of the errors done by the models. False positives were filtered using the identity, geometry and categorical human being questions from the beginning of this section according to table 3.2. The answers for the identity and geometry question should be "yes" and the object in the box should be a "human being". Additionally, predicted boxes smaller than 33px (ECP evaluation settings for "reasonable") were not taken into account and only samples identified as humans by the general binary model were considered. This resulted into approximately 100 bounding boxes predicted by each of both models that were also manually inspected and in fact verified as humans. These are missing in the ground truth and are therefore wrong false positive errors that should not be penalized in the evaluation.

question	convergence	credibility	solvability
Can you identify the object?	≥ 0.9	≤ 0.4	≥ 0.9
Is the box drawn correctly?	≥ 0.9	≤ 0.4	≥ 0.9
Is the object in the box a human being?	≥ 0.9	≤ 0.4	≥ 0.9

Table 3.2: Filtering criteria of questions posed to annotators about the false positive errors. The aim to have certain answers with little inter annotator disagreement can be achieved by taking samples with a high convergence and solvability with a narrow high-density interval implied by the credibility.

Analysis of FN Errors. Similar filtering could be done for the false negative boxes which were filtered by the ambiguity measure before. Now, the identity and categorical human being question were used to find samples that are extremely ambiguous. Simply filtering by "no" as the answer to the question if the annotators could identify the object inside the bounding box and not "human being" to the categorical question, 220 boxes are left in total that depict very ambiguous or even hallucinated pedestrians. Calculating the ambiguity measure for these boxes in Figure 3.13, the distribution is concentrated around high ambiguity score values above 0.6. This confirms again the validity of the ambiguity measure due to the agreement with the results from the annotators.

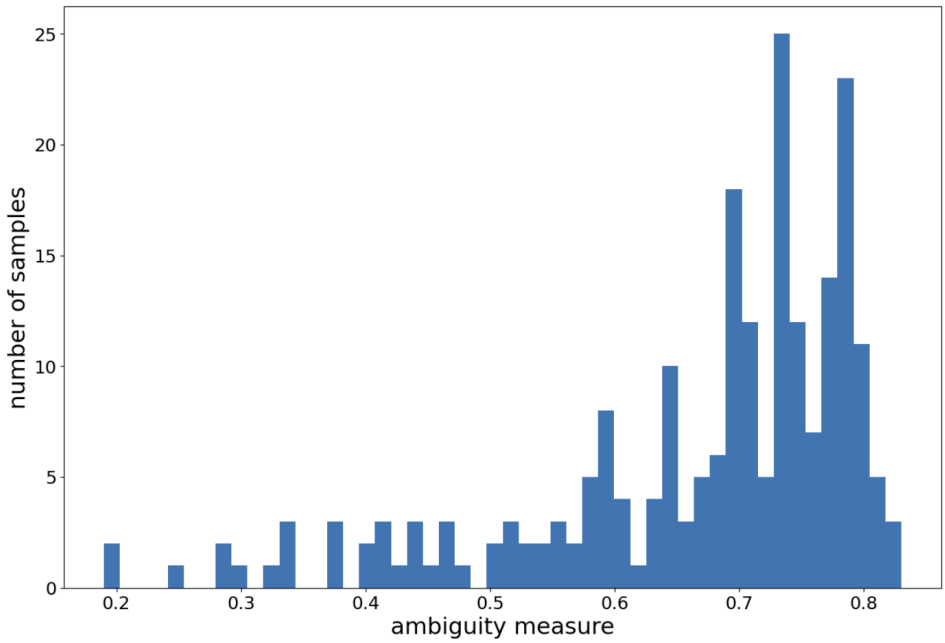


Figure 3.13: Distribution of the ambiguity measure for very ambiguous FN boxes found by annotators. The histogram is binned by the ambiguity measure on the x-axis. The distribution is concentrated around high ambiguity values, confirming that human perception is consistent with the results of the general binary model and the resulting ambiguity measure.

For added confidence in the results, further refinement was conducted by filtering out boxes with a low ambiguity score (≤ 0.5), resulting in 193 remaining bounding boxes. These samples depict highly ambiguous ground truth instances and are unsuitable for inclusion in the ground truth test dataset. Moreover, they should not be penalized if undetected.

Correction of Wrong Errors and Tags. The findings retrieved from the annotator evaluation of the errors and tags were used to correct the original test dataset. The model trained on the whole original training data and the model trained on this data pruned by an ambiguity threshold of 0.65 were evaluated on the uncorrected and corrected test dataset with 100 less false positive boxes and 193 less false negative boxes. Figure 3.14 shows the differences of the LAMR for both models which are greater for the amb 0.65 model. The occlusion tags were corrected for every evaluation in this plot. Also, the disparities within the "reasonable" subset are smaller than within "occluded" and "all". This effect can be explained by the correlation between ambiguity measure and occlusion. The evaluation subsets "occluded" and "all" allow a wider range of occlusion tags than "reasonable". As most of the corrected false negative boxes have a high ambiguity score which also means a rather high level of occlusion, most of these boxes did not even occur in the subset "reasonable". Therefore not penalizing them does not have a great effect on the LAMR of this subset. The largest difference of the correction happened for the subsets "occluded" (-0.012) and "all" (-0.01) of the ambiguity threshold model.

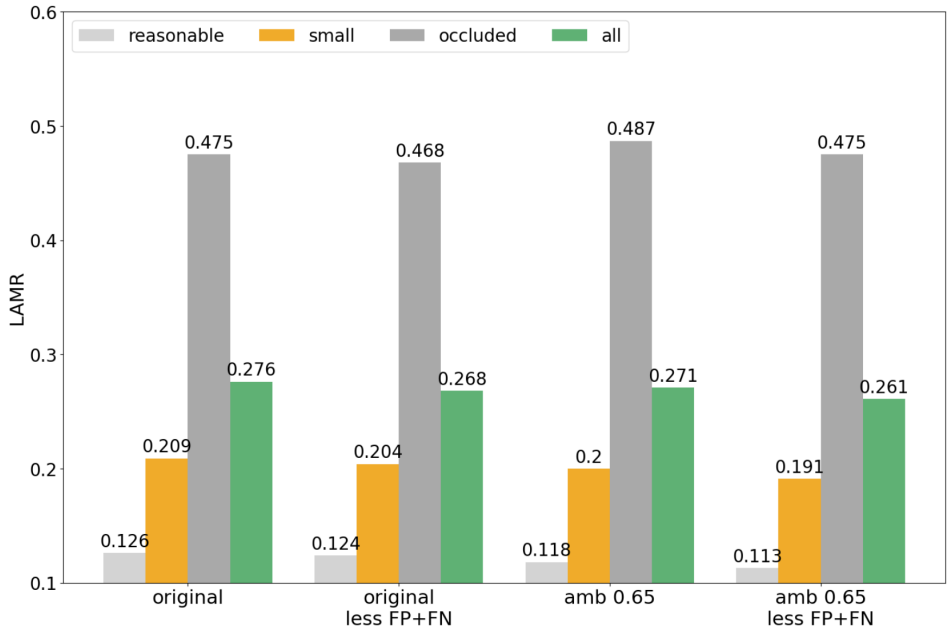


Figure 3.14: Changes of the LAMR when correcting the discovered wrong false positives and false negatives in the original test data for the original and ambiguity threshold model. On the x-axis, both models without and with the correction are displayed, each for all four evaluation subsets. The occlusion tags have been corrected beforehand. From no correction to not penalizing the identified 100 FP boxes and 193 FN boxes, the log average miss-rate decreases, especially for the "occluded" and "all" subsets. These subsets that allow high occlusion tags, include most of the ambiguous data which was partly pruned.

Conclusion. The analysis of the occlusion tags in the ground truth and the errors that the models made provided valuable insights into their performance and the quality of the test data. Discrepancies between the ground truth occlusion tags and annotator responses have been revealed which indicate inaccuracies in the dataset annotations. Annotator evaluations on the FP and FN boxes of the models provided clarity on the correctness of the errors and made identifying of very ambiguous and missed annotations possible. Approximately 5% of the false positive errors were identified as humans and 6.5% of the false negative error samples were evaluated as very ambiguous. Error filtering and correction based on the annotation tasks was done to enhance the accuracy of the evaluation and show the difference these samples make in the LAMR calculation.

In this chapter, a comprehensive exploration of the impact of ambiguous data on state-of-the-art pedestrian detection models was carried out including an investigation into the influence of neighbouring classes on the model performance. Starting by pruning ambiguous data from both training and test data, significant improvements in precision with minor decreases in recall were observed. To address this, augmentation of the training dataset with additional ground truth samples from neighbouring classes was done which resulted in an improvement in recall, particularly for occluded and ambiguous instances. However, this enhancement was tempered by a reduction in precision due to increased false positive detections. Furthermore, the exploration of model and annotation errors showed several critical insights into the dataset quality. Through systematic evaluation and analysis of annotated

samples by human annotators, missing annotations, highly ambiguous objects and inaccurate occlusion tags were identified. Leveraging annotator evaluations, the errors in the test dataset were corrected, resulting in measurable changes during the evaluation. The impact of the retrieved results on the log average miss-rate and the ECP leaderboard will be investigated in the next chapter and a broader perspective of their implications on pedestrian detection will be provided in the discussion chapter.

4 LAMR Evaluation and Impact of Ambiguity

Despite the detailed experiments conducted in the results chapter, the question of the overarching impact of ambiguity on the log average miss-rate remains. This chapter seeks to unravel the complexities surrounding the influence of ambiguity and the established results on the LAMR, underscoring its significance in model evaluation.
The log average miss-rate serves as a critical metric for evaluating pedestrian detection models. Figure 4.1 shows that it penalizes false negative boxes more than false positives, the same amount of added error boxes results in a higher LAMR value. This makes sense due to the requirements of not missing any pedestrian in the ground truth. Experiments, manipulating the annotation files and calculating the log average miss-rate with the ECP repository, have shown that the LAMR behaves the same for adding and deleting errors from the data. This means that if adding 10 FN boxes increases the LAMR by 0.0011, it would decrease by the same value if 10 FN boxes were deleted. Another characteristic of the LAMR is a higher sensitivity for lower values, e.g. the difference between 0 and 50 added FN boxes is higher than between 50 and 100 added false negatives.

The Effect of Wrong Errors and Ambiguity. With an ambiguity threshold of 0.65, 7.3% of the ground truth test dataset was identified as ambiguous and therefore pruned in earlier experiments. As this part of the data is highly ambiguous and not even humans are in agreement for these samples, one can not be sure if a model detects these objects or not. Additionally, 100 false positive errors were identified as humans and if the model detects those, which would be correct, it would still be penalized since they are not in the ground truth data. There is a best and a worst case scenario for the LAMR influenced by these samples. Best case, the model detects all ambiguous objects in the ground truth and does not predict any of the 100 FP humans. The LAMR would decrease. The worst case scenario involves the model failing to detect any of the ambiguous samples and generating 100 additional false positive errors for the humans absent in the ground truth. The LAMR would increase and punish this behaviour. Concentrating on the "reasonable" evaluation subset because it is expected to include most of the clear samples due to the filtering criteria, Figure 4.2 presents the ECP leaderboard as of February 2024 with upper and lower bounds of the LAMR for each model considering 138 wrong FN boxes identified by the ambiguity measure and 100 wrong FP boxes found by annotators. These interval estimations were made by manipulating

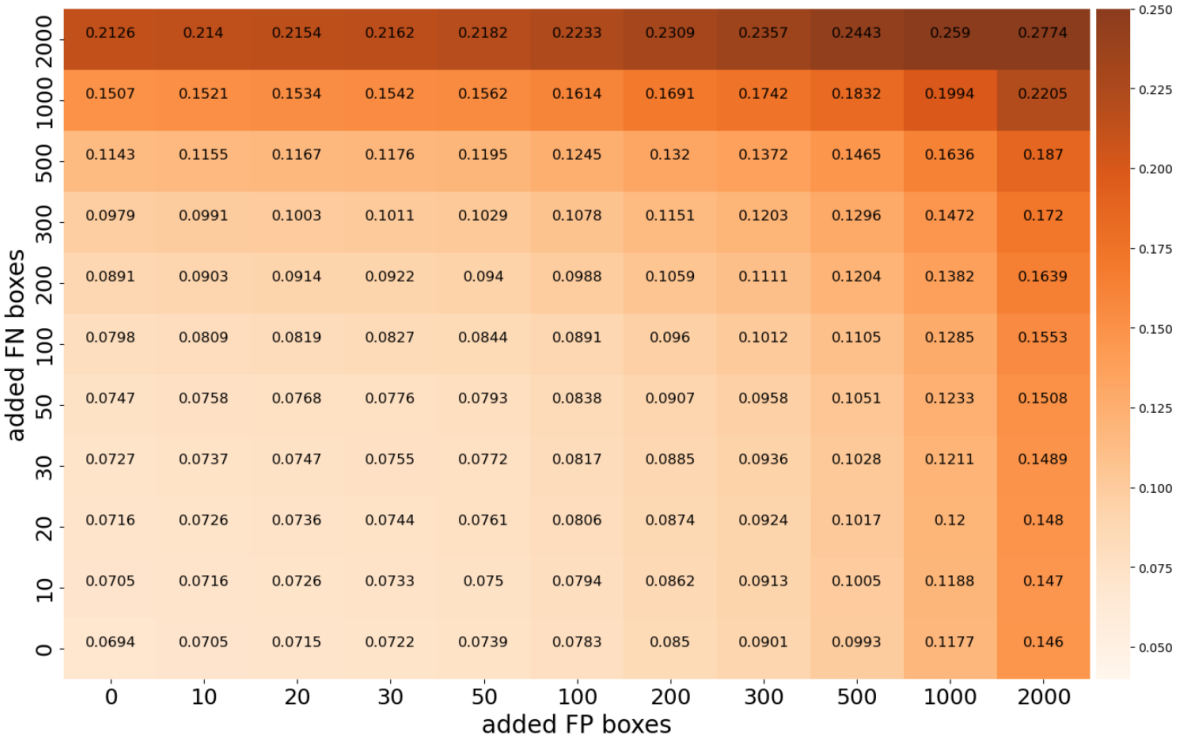


Figure 4.1: LAMR variations with added false positive or false negative errors. The x-axis represents the quantity of FP boxes while the y-axis indicates the quantity of FN boxes added to the base LAMR of 0.0694 (bottom left), calculated through inference from the available weights of the best model in the Pedestron repository on the "reasonable" test subset. Notably, false negatives are punished harder than false positives with higher sensitivity observed at lower LAMR values.

the test data so that the corresponding LAMR of the model is the result. Subsequently, false positive and false negative errors were added or deleted at this log average miss-rate to simulate the impact of wrong errors on the model performance. This method does not account for the possibility that models may partially predict the wrong errors correctly which would result in less pronounced changes in LAMR. The calculated bounds represent the extreme scenarios. These adjustments to the test data, based on the ambiguity measure, annotator estimations and manual inspections, could potentially lead to a difference of up to 8 places in the ranking.

A Better Ranking of the ECP Leaderboard. The last section motivated the significance of addressing ambiguity in model evaluation. To improve the ranking of leaderboards like the one from EuroCity Persons and make it more informative, one approach is to incorporate intervals of the metric that account for ambiguity like it was done in 4.2. This could also contain providing minimum and maximum values for different ambiguity measure thresholds in the test data. For a lower ambiguity threshold, the intervals would be larger and vice versa. Furthermore, visualizing the min and max metric intervals over various ambiguity thresholds could offer valuable insights into how model performance fluctuates with different levels of ambiguity. Additionally, creating an extra category for non-ambiguous data could help in scenarios where ambiguity is minimal or easily filterable. These approaches could enhance the interpretability and utility of leaderboard rankings in general which is necessary because

Method	User	LAMR (reasonable)▲		LAMR (small)	LAMR (occluded)	LAMR (all)	External data used	Publication URL	Publication code	Submitted on
SPNet w cascade	Huawei Noah AI Th...	0.042	▲0.018 ▼0.066	0.095	0.216	0.139	ImageNet	yes	yes	2020-03-18 23:33:33
LSFM	Abdul Hannan Khan	0.044	▲0.020 ▼0.068	0.099	0.230	0.150	ImageNet, TJU-DHD...	yes	yes	2022-10-16 16:15:54
Pedestron	IIAI, UAE	0.051	▲0.027 ▼0.075	0.112	0.254	0.162	ImageNet	yes	yes	2020-03-09 11:56:49
APD	Anonymous	0.053	▲0.029 ▼0.077	0.124	0.268	0.173	ImageNet	yes	no	2020-05-08 05:49:20
SPNet w FPN	Huawei Noah AI Th...	0.055	▲0.031 ▼0.079	0.121	0.246	0.165	ImageNet	yes	yes	2019-10-15 09:44:33
Pedestrian2	Hongsong Wang	0.056	▲0.032 ▼0.080	0.126	0.266	0.171	ImageNet	no	yes	2019-11-06 07:07:40
DAGN	DSLab	0.059	▲0.035 ▼0.083	0.142	0.263	0.175	ImageNet	yes	no	2021-07-01 06:28:00
Real-time Pedestr...	Irtiza and LiJinp...	0.066	▲0.042 ▼0.090	0.136	0.313	0.193	ImageNet	yes	yes	2020-01-13 10:18:58
Irtiza and LiJinp...	Irtiza Hasan	0.086	▲0.063 ▼0.109	0.168	0.379	0.230	ImageNet	yes	yes	2019-12-04 12:29:36
YOLOv3	ECP Team	0.097	▲0.075 ▼0.119	0.186	0.401	0.242	ImageNet	yes	no	2019-04-01 17:08:05
Faster R-CNN	ECP Team	0.101	▲0.080 ▼0.122	0.196	0.381	0.251	ImageNet	yes	no	2019-04-01 17:06:33
F2DNet	Abdul Hannan Khan	0.107	▲0.087 ▼0.127	0.175	0.387	0.261	ImageNet	yes	yes	2021-12-29 18:23:11

Figure 4.2: The ECP detection leaderboard with upper and lower bounds for the LAMR of "reasonable" due to ambiguity. The rows represent different models and methods submitted and the y-axis indicates different properties including the LAMR for the evaluation subsets "reasonable", "small", "occluded" and "all". The estimation is not able to consider the amount of samples that are predicted "correctly" by the different models.

especially in cases where the metric difference is small and ambiguity is present, they do not provide valuable information that distinguishes the models from each other.

In the upcoming discussion chapter, a nuanced interpretation of the research findings from the results and LAMR evaluation, particularly focusing on the implications of ambiguity in pedestrian detection models, will be carried out. Additionally, avenues for future research to address emerging challenges will be promoted and this work will conclude with some practical advice for handling datasets containing ambiguous data.

5 Discussion

In this chapter, the results and novelty of the research will be summarized and interpreted. The impact of the findings will be discussed and future work will be motivated. At the end, practical advice will be given for dealing with ambiguous data.

Main Results. The results obtained from the conducted experiments provide valuable insights into the impact of ambiguous data on pedestrian detection models. Five key findings can be derived - (i) removing ambiguous data from the training dataset improves the log average miss-rate, except for test data that includes heavily occluded instances (Figures 3.1, 3.7, 3.14). For example, in 3.7, the ambiguity 0.65 model reached lower LAMR values than the original model in each evaluation subset except "occluded". The reason for this improvement is the increased precision, not the recall (Figures 3.2, 3.4, 3.5). Therefore, ambiguous data contributes to the generation of false positive detections which can also be seen in 3.8a, where the model with less ambiguous training data has an overall better precision of 5.1%.
(ii) Augmentation of the training dataset with annotations from neighbouring classes enhances the recall at the expense of reduced precision across all evaluation categories (Figures 3.9, 3.10). Figure 3.9 shows a general decrease of precision by 2.9% and an increase of recall by 0.8%.
(iii) Pruning ambiguous samples from the test dataset narrows the gap in recall between models trained with and without ambiguity removal, e.g. in 3.8 by comparing the recall differences of the two models for both test datasets (1.7% to 0.5%). Consequently, this leads to greater differences in LAMR between the models because the precision differences remain, highlighting the importance of test data quality during the evaluation (Figure 3.7).
(iv) A strong correlation between ambiguity and occlusion was revealed when comparing the ambiguity measure, that reflects the distribution of human annotator estimations, with the occlusion tags in the ground truth (Figures 2.9, 3.6). Higher values of the ambiguity measure correspond to a greater prevalence of occlusion tags within the dataset. Notably, as the ambiguity measure increased, the proportion of tags indicating higher levels of occlusion also elevated which is evident in 3.6, where the peak of the "occluded>80" tag proportion is at an ambiguity measure value of 0.79.

(v) Errors in dataset annotations extend beyond bounding boxes and labels to additional tags for occlusion that can significantly impact model evaluation results (Figures 3.11, 3.12). Correcting the 25.2% wrong tags on average in the model errors lead to log average miss-rate deviations of up to 0.012 (Figure 3.7 compared to 3.14).

Trade-Offs and Implications. The main findings of this work highlight several trade-offs that must be considered in pedestrian detection models. The first trade-off between the recall and the precision often occurred when comparing different subsets and underscores the challenges of balancing model performance across different evaluation metrics. This can also be affected by the use case and the requirements to the model since in applications like vision-supported emergency braking, the recall has to be as low as possible to not miss any pedestrian. On the other hand, large false positives that are hallucinated close in front of the car would not be beneficial as well. But there might be cases where specific errors are irrelevant because the boxes are very small and far away for example.
While filtering out ambiguous data is beneficial for generalization, there is a second trade-off when it comes to handling occlusion. In this case, it is essential to find a balance between data cleanliness and representativeness, noise removal and retention of valuable training examples. Due to the correlation of ambiguity and occlusion, excessively pruning ambiguous training data may result in the model struggling to predict occluded samples effectively.
Looking back from this perspective, most of the hypotheses stated at the end of section 2.1 were confirmed, with the exception that less ambiguous training data does not generally improve model performance because that depends on several factors: the evaluation measure that the performance is quantified with, the quality and amount of ambiguity in the test data and the real-world application data and the specific requirements of the model's use case. This study showed that ambiguity is inherent in the training and the test data and therefore has a great effect on the training and the evaluation of the model. The title of this thesis challenged the simplistic notion that poor quality training data will inevitably lead to poor quality model output. In pedestrian detection, this is definitely not true as it is equally important to recognize the multifaceted nature of data ambiguity and its effects on the training and evaluation. It is important to be aware of the ambiguity inherent in the data, both in the training data and in the test data, to be able to create high-quality datasets and develop robust models.

Future Work. Moving forward, there are multiple directions for future research in the domain of handling ambiguity and occlusion in computer vision tasks. One promising way is the development of more sophisticated ambiguity measures that go beyond simple statistical summaries. For instance, multiple information could be derived from the ambiguity distribution of each object to get an even more precise estimation of the ambiguity. Additionally, while all models in this work were not pre-trained, investigating the impact of ambiguity on pre-trained models could be valuable. Pre-training often relies on large-scale

datasets which may exhibit different characteristics of ambiguity compared to smaller and task-specific datasets like ECP. This could provide important insights that could be used in transfer learning and domain adaption strategies. Furthermore, ambiguity not only affects object detection but also other areas such as semantic segmentation or instance segmentation where tailored approaches for handling the discovered issues may be necessary. In various domains, leaderboards or rankings are commonly utilized as a visual tool for comparing different approaches and methodologies based on specific metrics. As shown in chapter 4, the same applies to the EuroCity Persons dataset where the LAMR is heavily influenced by ambiguous data and ambiguity can be the reason for misleading conclusions regarding model performance. Therefore, future work should also focus on ambiguity-aware evaluation metrics and methodologies to provide more accurate and reliable assessments of model performance. Ultimately, addressing ambiguity in datasets and evaluation procedures is crucial for advancing the field of computer vision.

Concluding Summary and Advice. The EuroCity Persons dataset which was thoroughly analysed in this study, contains a notable amount of ambiguity within both training and test datasets. Specifically, the validation data that was employed as the test set in this research due to the availability of ground truth labels, includes approximately 7.3% ambiguous instances while the training dataset exhibits a slightly higher ambiguity rate of 8.6%. Given the substantial ambiguity present in the ECP dataset, it is evident that datasets of this nature pose significant challenges for model training but also for the evaluation. Ambiguity manifests in different forms ranging from incorrect labels to inaccurate bounding boxes and erroneous tags. Therefore, it is important for users encountering datasets with ambiguous data to take a proactive approach in overcoming these challenges. Advising the implementation of strategies for identifying and quantifying the degree of ambiguity in the data. This involves the development of measures or metrics to rank the level of ambiguity for each sample. Thus, one can gain insights into the impact of ambiguous data on the training and model performance to make informed decisions regarding the data preprocessing and training. It is crucial to recognize that ambiguity can be present in the training and test data. Therefore, users should also assess if the prediction of ambiguous data is necessary and if the model needs to be capable of handling ambiguous data appropriately or if it is acceptable to disregard ambiguous instances.

In summary, when confronted with datasets containing ambiguous data, it is essential to prioritize the development of robust strategies to identify, quantify and address ambiguity. By adopting systematic approaches to manage ambiguity, the quality and reliability of datasets can be enhanced, leading to more robust and accurate model development and evaluation.

List of Tables

List of Figures

Bibliography

[1] C. G. Northcutt, A. Athalye, and J. Mueller, *Pervasive label errors in test sets destabilize machine learning benchmarks*, 2021. arXiv: 2103.14749 [stat.ML].

[2] C. E. Brodley and M. A. Friedl, "Identifying mislabeled training data", *Journal of Artificial Intelligence Research*, vol. 11, pp. 131–167, Aug. 1999, ISSN: 1076-9757. DOI: 10.1613/jair.606. [Online]. Available: http://dx.doi.org/10.1613/jair.606.

[3] C. Brodley and M. Friedl, "Improving automated land cover mapping by identifying and eliminating mislabeled observations from training data", in *IGARSS '96. 1996 International Geoscience and Remote Sensing Symposium*, vol. 2, 1996, 1379–1381 vol.2. DOI: 10.1109/IGARSS.1996.516669.

[4] G. Libralon, A. de Carvalho, and A. Lorena, "Pre-processing for noise detection in gene expression classification data", *J. Braz. Comp. Soc.*, vol. 15, pp. 3–11, Mar. 2009. DOI: 10.1590/S0104-65002009000100002.

[5] J. Li, D. Xu, and W. Gao, "Removing label ambiguity in learning-based visual saliency estimation", *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1513–1525, 2012. DOI: 10.1109/TIP.2011.2179665.

[6] W. Yuan, G. Han, and D. Guan, "Learning from mislabeled training data through ambiguous learning for in-home health monitoring", *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 2, pp. 549–561, 2021. DOI: 10.1109/JSAC.2020.3021572.

[7] S. Chadwick and P. Newman, *Training object detectors with noisy data*, 2019. arXiv: 1905.07202 [cs.RO].

[8] M. Andresen, M. Vauth, and H. Zinsmeister, "Modeling ambiguity with many annotators and self-assessments of annotator certainty", in *Proceedings of the 14th Linguistic Annotation Workshop*, S. Dipper and A. Zeldes, Eds., Barcelona, Spain: Association for Computational Linguistics, Dec. 2020, pp. 48–59. [Online]. Available: https://aclanthology.org/2020.law-1.5.

[9] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander, and N. Silberman, *Learning from noisy labels by regularized estimation of annotator confusion*, 2019. arXiv: 1902.03680 [cs.LG].

[10] Schaekermann, Mike, "Human-ai interaction in the presence of ambiguity: From deliberation-based labeling to ambiguity-aware ai", Ph.D. dissertation, 2020. [Online]. Available: http://hdl.handle.net/10012/16284.

[11] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity", *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017. DOI: 10.1109/TIP.2017.2689998.

[12] M. Braun, S. Krebs, F. B. Flohr, and D. M. Gavrila, "Eurocity persons: A novel benchmark for person detection in traffic scenes", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2019.2897684.

[13] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, *Generalizable pedestrian detection: The elephant in the room*, 2020. arXiv: 2003.08799 [cs.CV].

[14] J. Wang, K. Sun, T. Cheng, *et al.*, *Deep high-resolution representation learning for visual recognition*, 2020. arXiv: 1908.07919 [cs.CV].

[15] Z. Cai and N. Vasconcelos, *Cascade r-cnn: High quality object detection and instance segmentation*, 2019. arXiv: 1906.09756 [cs.CV].

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, *Mobilenetv2: Inverted residuals and linear bottlenecks*, 2019. arXiv: 1801.04381 [cs.CV].

[17] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, *Microsoft coco: Common objects in context*, 2015. arXiv: 1405.0312 [cs.CV].

[18] Z. Cai and N. Vasconcelos, *Cascade r-cnn: Delving into high quality object detection*, 2017. arXiv: 1712.00726 [cs.CV].

[19] C. Klugmann and L. Schwirten, "Quantifying ambiguity in crowd-sourced categorical tasks", unpublished, 2023.

[20] C. Klugmann, "Human-informed automation: A dirichlet-based approach for integrating uncertainty in machine learning on visual data for efficient annotation and dataset quality assurance", unpublished, 2023.