Data-Centric Engineering (2024), 5: e34
doi:10.1017/dce.2024.39

CAMBRIDGE
UNIVERSITY PRESS

RESEARCH ARTICLE

The inconvenient truth of ground truth errors in automotive datasets and DNN-based detection

Pak Hung Chan, Boda Li, Gabriele Baris, Qasim Sadiq and Valentina Donzella

WMG, University of Warwick, Coventry, UK
Corresponding author: Pak Hung Chan; Email: Pak.Chan.1@warwick.ac.uk

Received: 15 December 2023; **Revised:** 24 July 2024; **Accepted:** 15 September 2024

Keywords: machine learning; automated vehicles; automotive dataset; labeling

Abstract

Assisted and automated driving functions will rely on machine learning algorithms, given their ability to cope with real-world variations, e.g. vehicles of different shapes, positions, colors, and so forth. Supervised learning needs annotated datasets, and several automotive datasets are available. However, these datasets are tremendous in volume, and labeling accuracy and quality can vary across different datasets and within dataset frames. Accurate and appropriate ground truth is especially important for automotive, as "incomplete" or "incorrect" learning can negatively impact vehicle safety when these neural networks are deployed. This work investigates the ground truth quality of widely adopted automotive datasets, including a detailed analysis of KITTI MoSeg. According to the identified and classified errors in the annotations of different automotive datasets, this article provides three different criteria collections for producing improved annotations. These criteria are enforceable and applicable to a wide variety of datasets. The three annotations sets are created to (i) remove dubious cases; (ii) annotate to the best of human visual system; and (iii) remove clear erroneous BBs. KITTI MoSeg has been reannotated three times according to the specified criteria, and three state-of-the-art deep neural network object detectors are used to evaluate them. The results clearly show that network performance is affected by ground truth variations, and removing clear errors is beneficial for predicting real-world objects *only* for some networks. The relabeled datasets still present some cases with "arbitrary"/"-controversial" annotations, and therefore, this work concludes with some guidelines related to dataset annotation, metadata/sublabels, and specific automotive use cases.

Impact Statement

The proposed work can strongly impact the automotive community in manifold ways: (i) the development of several automotive perception algorithms rely on the data from big annotated datasets, highlighting how errors in annotations can affect neural network performance, and which neural networks are more robust can inform future algorithm design and deployment; (ii) proposing some clear and enforceable criteria for annotation (applicable to any automotive datasets), with different levels of formality and enforceability, this work might promote a more uniform way of labeling in the automotive community.

We believe this work can have strong implications for neural network research, as well as on their deployment in industry, combined with the use of well-known automotive benchmarking datasets.

This research article was awarded an Open Materials badge for transparent practices. See the Data Availability Statement for details.

Pak Hung Chan, Boda Li and Gabriele Baris contributed equally.

1. Introduction

With the recent advances in the field of artificial intelligence (AI) and machine learning (ML), deep neural networks (DNNs) are becoming commonly used in many fields, from agriculture to medical, from manufacturing to robotics (Dokic et al., 2020; Retson et al., 2019; Park et al., 2021). Compared to traditional algorithms, ML algorithms can provide better flexibility to unforeseen and uncommon circumstances. This flexibility is critical for applications like assisted and automated driving (AAD) functions due to the high variability (of the environment and road stakeholders) that can be encountered during a vehicle journey. There are countless factors that can affect an AAD *sense-perceive-plan-control* pipeline, from the degradation of perception sensor data to unpredictable actor actions, and they can compromise the overall safety of the vehicle (Chen et al., 2021).

1.1. AAD and DNNs

Assisted driving functions are currently deployed in commercial vehicles, and automated driving is a major research area in the vehicle industry and academia (Ingle & Phute, 2016; Kukkala et al., 2018; Li & Shi, 2022; Du, 2023). Functions are developed to reduce driver workload, thus improving safety and comfort. These functions can range from simple audiovisual warnings for the driver to systems taking partial or full control of the vehicle. The Society of Automotive Engineers (SAE) has published the J3016 standard, defining six levels of driving automation, from Levels 0 to 5 (SAE J3016_202104, 2021). As the level of automation increases, so does the amount of control and driving situations which the AAD function can handle. Due to the complexity and variability of the driving environment, researchers are increasingly turning to DNN-based functions for their flexibility and ability to handle previously unseen inputs (Koopman & Wagner, 2016; Li & Ibanez-Guzman, 2020; Cui et al., 2021).

There are many datasets collected for developing, testing, and benchmarking some of the perception and prediction functions used to support AAD tasks. They present annotated and labeled ground truth data appropriate to the different perception/prediction tasks (i.e. bounding boxes (BBs) for object detection, pixel masks for segmentation) (Guo et al., 2020). Commonly used and established benchmarking datasets are the KITTI (Geiger et al., 2012), Berkeley DeepDrive (BDD) (Xu et al., 2017), nuScene, (Caesar et al., 2020) and CityScapes (Cordts et al., 2016). These datasets publish and regularly update leaderboard tables to compare the performance of novel neural networks (for different tasks), and all developers can submit their results.

However, for automotive datasets, improvements in detection performance are becoming smaller and smaller (Valada, 2022). One difficulty is the presence of inaccurate annotations, see Sec. 3.2. In particular, incorrect labeling of the testing/evaluation data split will lead to wrong classification of the predicted boxes. In this work, an investigation into different criteria for dataset annotation is performed, and the effect of different labeling criteria is analyzed.

1.2. Contributions

This article discusses, analyses, and classifies the annotation errors in widely used automotive datasets, namely KITTI and nuScene (Siam et al., 2017; Geiger et al., 2012; Caesar et al., 2020). These identified errors are used as a guide to propose some improved criteria for dataset annotation. As the criteria present some arbitrary aspects (further discussed in the paper), three different sets of criteria are used to generate an equal amount of annotation versions, and these annotation sets are compared using three DNNs, covering the state-of-the-art architectures for object detection. In this context, the main contributions are as follows:

1. the authors demonstrate that incorrect annotations have a detrimental effect on the DNNs' learning and performance, and even small improvements in the training labels can improve performance;
2. the performance is dependent on the DNN architecture and the annotation criteria, but overall training with more accurate labels seems beneficial;

https://doi.org/10.1017/dce.2024.39 Published online by Cambridge University Press

3. removal of BBs not belonging to visible objects in training is beneficial for the used one and two-stage detectors but not for the used transformer architecture;
4. proposing different sets of labeling criteria with different levels of formality can be key to better understand the learning of the DNNs and can support future more accurate annotation processes;
5. proposing annotation criteria which can be applied to any automotive datasets and automatically enforced.

The results show that by properly improving the quality of the annotations, an increase of up to 9% can be achieved in the DNN performance when evaluating mAP_{50}. Moreover, removing obviously incorrect BBs in the test sets improves the evaluated performance metrics. This step is critical as a better-measured accuracy improves the public perception related to the use of DNNs for AAD tasks. Errors in BBs labels may also hinder the maximum potential of the DNN, affecting the training loss and the adjustments of weights. It is worth noting that this article focuses on the quality of the labeling of the data (real or synthetic), considering that datasets will be far from ideal and will contain sources of noise and imperfections, which will make the application of labeling criteria a complex procedure.

2. Background

The training data can highly affect the performance of the trained neural network. In object detection, BBs are used to define the ground truth. However, the quality of the BBs can differ between datasets, based on tools/annotators creating the labeling, and also within a dataset due to some ambiguities in the data or annotation process. This section presents some annotation processes of datasets and works on understanding errors in datasets.

2.1. Datasets and annotation criteria

Datasets have moved from simple classification tasks of the MNIST dataset, where each sample contain one class, to current big curated datasets, which include frames with multiple objects and different annotations and classes (Deng, 2012; Xiao et al., 2017). Existing works on dataset annotation cover different issues that are strongly related to the specific task of the neural network. Some datasets used for object detection have been manually annotated using proprietary software or annotation tools, e.g. WoodScape (Uřičář et al., 2019), RADIATE (Sheeny et al., 2021), or nuScenes (Caesar et al., 2020). nuScenes has also published instructions for their labeling; however, these labels are open to interpretation and produce errors similar to the ones identified in other datasets, e.g. KITTI MoSeg dataset, as shown in Figure 1. Other datasets have been developed and annotated using deep learning methods such as active learning and neural networks (Angus et al., 2018; Janosovits, 2022; Meyer & Kuschk, 2019). These methods of annotation are otherwise known as semiautomatic. Further techniques include the development of new annotation tools to adapt current datasets to specific use cases in AAD (Arief et al., 2020; Wang et al., 2019). However, in the above-mentioned datasets, the specific criteria used to define how annotations should be implemented are missing, not explicit or ambiguous.

Interestingly, the VOC dataset has published the 2011 annotation guidelines used for labeling (Everingham et al., 2009). The guidelines provide guidance on which images to label and how they should be categorized. Examples include categorizing an object as occluded if more than 5% of the object within a BB is occluded, and there are images considered too difficult to segment and left unlabeled, e.g. a nest of bicycles.

2.2. Quality of annotations

Due to the effort and complexity required to annotate datasets, there are often errors associated with some of the labels. Recent works are trying to understand how these errors affect the DNN performance and how to improve the quality of the annotations. Notably, Ma et al. have presented a work in which they reannotated the MS COCO and Google Open Images datasets by providing a description of the object as

https://doi.org/10.1017/dce.2024.39 Published online by Cambridge University Press

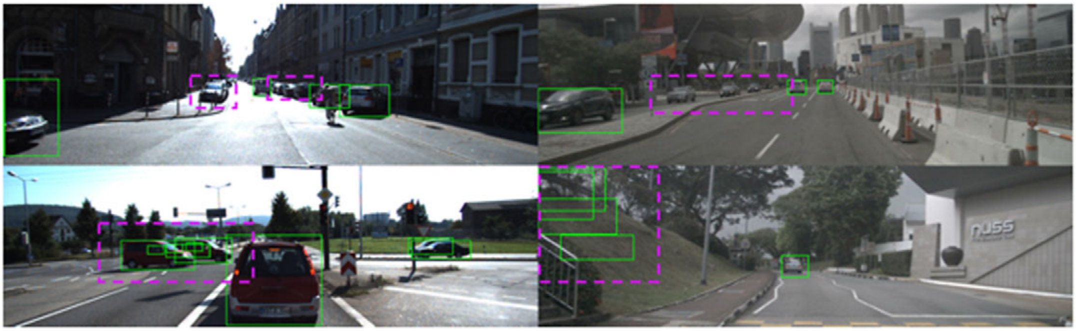


Figure 1. *Examples of errors (highlighted by dotted rectangles) in the BBs (green rectangles) of KITTI MoSeg (left) and nuScenes (right): upper frames show missing BBs, and lower frames show BBs not belonging to any objects*

well as providing some common examples of positive or negative cases which are open to the interpretation of the annotators (Ma et al., 2022). They have trained five neural network models based on the original labels and the new labels. For the COCO dataset, only some of the considered performance metrics improved due to reannotation, whereas the results on Google Open Image are improved after training and/or testing with the new annotations. Northcutt et al. provided a study into commonly used datasets for ML to understand and categorize the errors in labeling (Northcutt et al., 2021). In addition, they provided results on benchmarking datasets, comparing the original labels with corrected labels. Their results showed that lower capacity/simpler networks are more robust against the effects of erroneous labels. Tsipras et al. have focused on ImageNet dataset (Tsipras et al., 2020). The original ImageNet dataset provided only one class per image; however, each image may contain several secondary objects or clutter. Tsipras relabeled a subset of the ImageNet database (10 images per class, 10,000 images in total) and provided multiclass annotations where appropriate in their subset. The neural network performed worse on images that had multiple object classes in the scene with respect to images with only one object. Additionally, Tsipras et al. investigated the performance of the trained DNN when evaluating real-world data by comparing the neural network predictions with the classes identified by human annotators. More accurate models are in better agreement with the human annotators, and even when some network predictions are technically wrong, there is often an agreement with the classification by the human annotators (*that is* identification of dog breeds by nonspecialists).

2.3. DNN-based object detection

As previously mentioned, DNNs are expected to play a key role in AAD functions. One of the basic and most important tasks is the detection of the road stakeholders, and particularly vehicles, as they cause most of the accidents (Feng et al., 2021; Rashed et al., 2021). Object detectors can be broadly divided into two classes. Moreover, recently, there have been several implementations of object detectors using transformers (He et al., 2021; Carion et al., 2020); a comprehensive review of object detectors is given in (Zaidi et al., 2022). Vision transformers are becoming very popular due to their promising performance; however, they usually require more epochs to converge and typically do not perform well on small objects (Yao et al., 2021). In terms of the two “traditional” categories of object detectors, the *one-stage* and *two-stage* detectors, several architectures have been proposed through the years. Usually, one-stage detectors are faster but have lower accuracy than two-stage detectors. However, the two-stage Faster R-CNN have a good balance in terms of speed and accuracy; hence, they are frequently used for AAD functions.

As a part of the hereby proposed work, the authors have selected one implementation for each type of object detector (one-stage, two-stage, and transformer) to understand if the observed trends in the results, when reannotating the datasets, are common across different architectures for the same perception task, that is object detection.

3. Methodology

In this work, the annotations of different automotive datasets were reviewed, and several errors were identified. Moreover, KITTI MoSeg dataset was reannotated using three different sets of proposed criteria by the authors, that is C1, C2, and C3, see Section 3.2. Three DNNs were fine-tuned on the dataset four separate times: with the original ground truth labels as well as with the three new sets of labels generated according to the defined criteria, resulting in 12 trained networks (4 per DNN type), see Section 3.3. These 12 networks were then used to evaluate the testing datasets on the 4 sets of labels (*that is* MoSeg original and C1-C2-C3).

3.1. KITTI MoSeg

The KITTI MoSeg dataset is a subset of the highly cited KITTI Benchmarking Vision Suite dataset (Siam et al., 2017; Geiger et al., 2012). Compared to the KITTI dataset, KITTI MoSeg provides chronological sequences of frames, totaling 1449 frames (Siam et al., 2017). The KITTI MoSeg dataset provides 2D BB labels for *car* and *van*; the proposed work merged them into a single *vehicle* class. This choice was made due to the poor ratio between the two classes in the dataset.

In the KITTI MoSeg dataset, the original annotations were expanded by using the information available in the original dataset (3D BB, odometry information), where 3D ground truth BBs were converted into 2D BBs, associating BBs across chronological frames to obtain estimated velocity per BB. Finally, a filtering process was applied to keep objects consistently identified across frames (Siam et al., 2017). As the experiments proposed in this paper entail manually reannotating the same dataset three times, a dataset with moderate size was the best choice. However, the BB errors tackled in this paper are common across different automotive datasets.

3.2. Re-annotation criteria

From preliminary findings, the original labels in some automotive datasets were visually inspected to identify potential BB problems which can affect DNN evaluation metrics. The majority of the identified errors can be categorized using the definitions below:

- Missing: There are clear and obvious vehicles in the frames which are not labeled.
- Incorrect: There is no object in the labeled class in the BB.
- Bad Fit: The BB size and/or position are not appropriate for the identified object.
- Occlusion: The BB predicts the full size of the object, not only what is visible.

Based on these identified problems, three sets of BB criteria were defined to ensure that the annotation process is more consistent, see Table 1. These sets of criteria were implemented one at a time to completely reannotate the KITTI MoSeg manually, using an ad hoc Matlab app developed by the authors. Although human labeling error can still exist, some of the criteria can be coded to ensure the drawn BB meets the criteria. The defined sets of criteria are not specific to this dataset and can be applied to any other datasets.

For Criteria 1 (C1), the guidelines are defined as much as possible in an objective way and based on what features are expected to be learnt by the neural network (that is wheels, shapes, etc.). C1.1 specifies the minimum dimension of a BB based on an analysis of box sizes in the original KITTI MoSeg labels. In the specific automotive use case, vehicles in the far distance cover less pixels, and they are not of immediate safety concern. However, the minimum BB dimension can be tailored based on the detection requirements. For C1.2, a 3-pixel error was defined to consider annotation inaccuracies to identify the exact edge of an object and issues due to the lack of contrast at the edges, but avoiding significant errors at the smallest size based on C1.1. C1.3 ignores occluded regions from the BB, avoiding incorrect features to be learnt by the network. A minimum threshold of 20% is applied in C1.4 to determine the minimum visibility of the object. The 20% visibility is inline with the 20% visibility by nuScenes nuImages Annotator Instructions (Caesar et al., 2020). However, the determination of the 20% is subjective to the annotator or even to an “automated” implementation as it is based on estimation.

https://doi.org/10.1017/dce.2024.39 Published online by Cambridge University Press

Table 1. *The three sets of criteria for the reannotation of the KITTI MoSeg dataset for object detection*

Criteria 1 (C1)	Criteria 2 (C2)	Criteria 3 (C3)
1.1 BB shall be no less than 15 pixels in width or height	2.1 Object is annotated if human annotator can identify it	3.1 Incorrect* BBs are removed from original KITTI MoSeg labels
1.2 BB shall contain all visible parts of the object, with an error lower or equal to 3 pixels	2.2 BB shall contain all visible parts of the object, with an error lower or equal to 3 pixels	3.2 Fully occluded BBs are removed from original KITTI MoSeg labels
1.3 BB shall not include any estimated or occluded parts of the object, unless criteria 1.2 is applicable	2.3 BB shall not include any estimated or occluded parts of the object, unless criteria 2.2 is applicable	
1.4 BB must be added when more than 20% of one side of the object is visible		

*Incorrect stands for BBs that clearly do not belong to any target objects

In Criteria 2 (C2), the labels were created to the best of the annotator ability to identify vehicles in each frame but still keeping a small set of guidelines. The annotator was one of the authors. Object identified according to C2 are expected to be closer to real world, where all the vehicles should be identified by the DNN.

Finally, in Criteria 3 (C3), the annotator did not resize nor add any BBs, but only removed the BBs not belonging to real objects; in the case of multiple BBs encompassing the same vehicle, the annotator kept only the BB deemed as the best fit for that vehicle.

3.3. Neural networks

The four annotated dataset variations were used to train three different network architectures: Faster R-CNN (Ren et al., 2015), YOLOv5 (Jocher et al., 2022), and DETR (Carion et al., 2020). Faster R-CNN is an example of two-stage detector, which predicts BBs from region proposals. On the other hand, YOLO is a single-stage detector that does BB regression from anchors. Finally, DETR directly performs BB prediction with respect to the input image. For each network model described in the list below, the training process was performed four times from the base model (one for each dataset variation based on the annotation criteria, *that is* Original, C1, C2, and C3) on the whole original or relabeled training image set, leading to four different trained DNNs per architecture.

- *Faster R-CNN.* The network of choice consists of a ResNet-50 backbone with FPN (Feature Pyramid Network) feature extractor from the torchvision library (He et al., 2016; Lin et al., 2017). It was originally pretrained on COCO and then fine-tuned over the dataset of interest. The training was performed using the AdamW optimizer with a learning rate set to 10^{-3} and weight decay set to 0.2. In addition, learning rate scheduler with gamma 0.9 and step size 25 was used.
- *YOLOv5.* The training was performed starting from the pre-trained `yolov5m` with most parameters left to their default value. The backbone was frozen, image size set to 640 px, and optimiser set to AdamW.
- *DETR.* The network of choice consists of a ResNet-50 convolutional feature extractor from the hugging face library (He et al., 2016). It was originally pretrained on COCO and then fine-tuned over the dataset of interest. The training was performed using the AdamW optimiser with learning rate set to 10^{-5} for the backbone and 10^{-4} for the other layers and weight decay set to 10^{-4}.

https://doi.org/10.1017/dce.2024.39 Published online by Cambridge University Press

3.4. Evaluation metrics

The trained networks were all evaluated and compared using mAP_{50}, due to the presence of small size objects in the dataset. In fact, minor location offsets and size errors for predicted boxes can result in a much lower Intersection over Union (IoU) for small objects. For safety critical functions in automated vehicle, it is still important to detect the object, even if the location/size may be slightly off. The YOLOv5 repository provides this evaluation metric, and the `torchmetrics` library was used to compute mAP_{50} for Faster R-CNN and DETR implementations. Thus, the three reannotation criteria are compared using the three main types of neural network-based object detectors.

4. Results and discussion

The number of ground truth BBs was computed for each relabeled dataset, see Table 2. According to C3, which seeks to remove the *incorrect* or fully occluded BBs, there are 612 BBs (8.2%) which were deemed to be incorrect in the original dataset. C2 produced the largest amount of BBs based on what the human annotator could identify. Many of C2 BBs differ from BB identified via C1, due to objects being too small, or occlusion too high.

During the reannotation process, there were some difficulties in adhering strictly to the identified criteria, and some labels can be subjective or open to interpretation, roughly representing 10% of the labels. Some of these cases have been listed in Table 3 and visually demonstrated in Figure 2. For example, 3 pixel error was selected to allow minor flexibility for the annotator and considering situations with low

Table 2. *Number of ground truth BBs for each set of annotations, with C# denoting the number of the set of criteria (1–3), and MoSeg denoting the original labels. The total number of frames (and their split into training and testing parts) remains the same throughout all the experiments*

	MoSeg	C1	C2	C3
Training	4302	4384	6025	4047
Validation	509	483	653	467
Testing	2648	2889	2852	2315
Total	7459	7756	9530	6829

Table 3. *Examples of ambiguous situations when applying the proposed criteria, related visual examples are given in Figure 2*

Case No.	Description	Criterion applied
1	Small but obvious objects to a human are not labeled	1.1
2	For small objects, 3 pixel error can contribute significantly to the dimension and position uncertainty of the BB (up to 20% error in width or height)	1.2
3	Objects meeting this criterion can have most of their surface occluded	1.3
4	Judging the percentage of occlusion is subjective	1.4
5	Small object annotation by humans is influenced by the annotator understanding and interpretation of the scene	2.1
6	In low contrast situations the judgment of object boundaries becomes arbitrary	2.2
7	Object through windows can be annotator dependent, and BB may include distortion and extensive occluded areas	2.3
8	It can be subjective to judge which BB is "correct" or "incorrect"	3.1
9	Annotated obstructed objects may not have any feature useful for recognition	3.2

https://doi.org/10.1017/dce.2024.39 Published online by Cambridge University Press

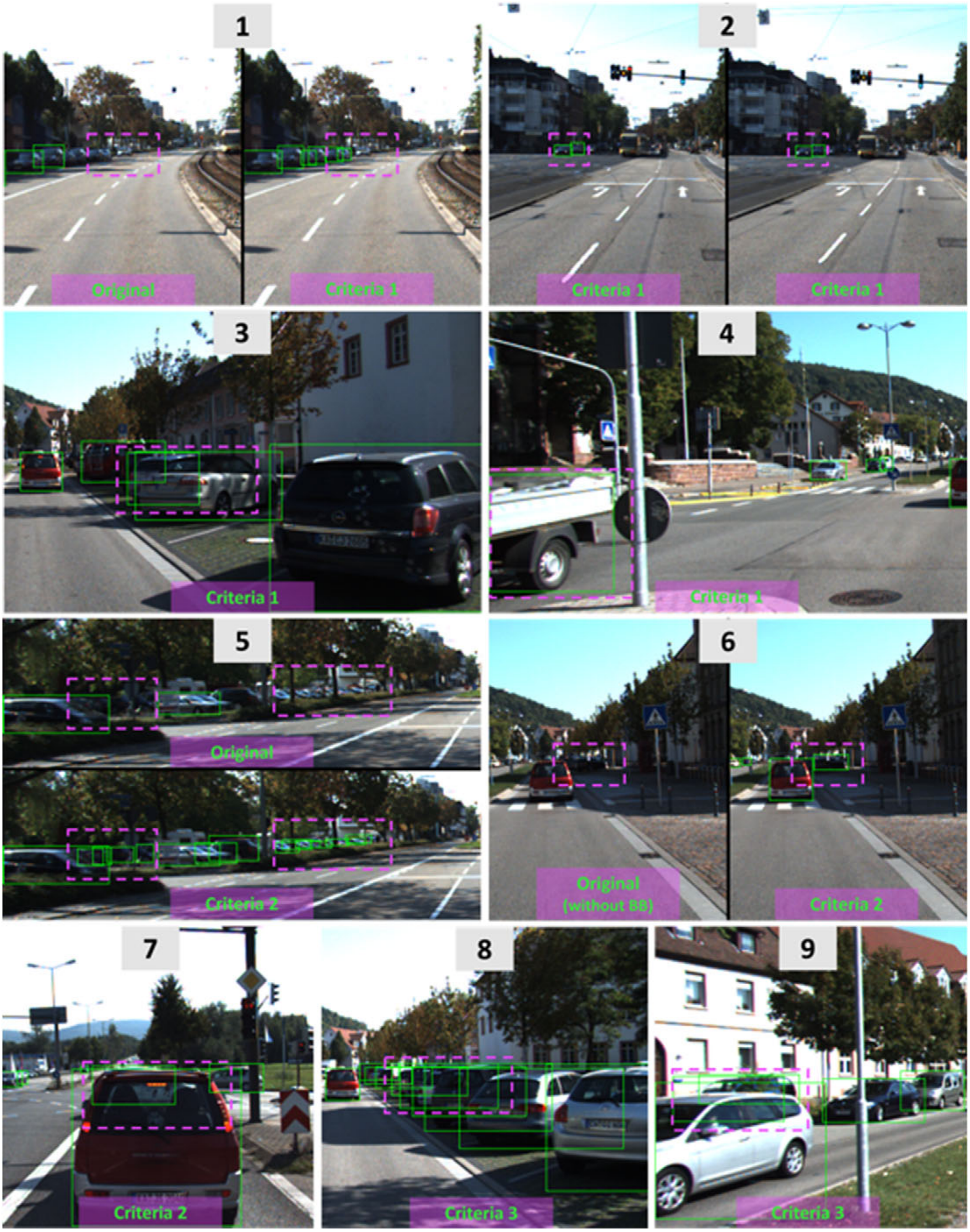
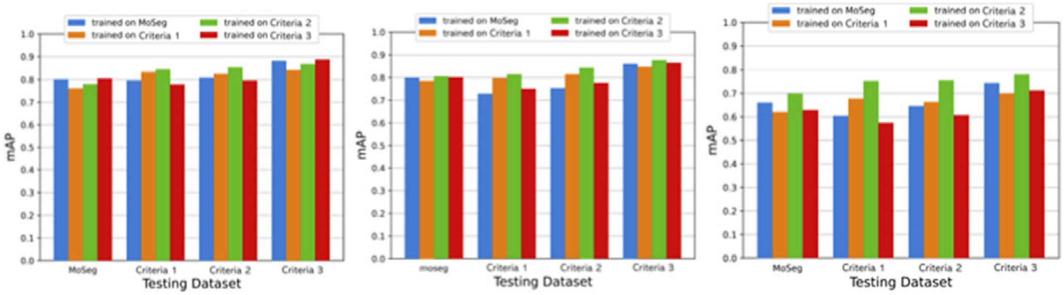


Figure 2. *Examples of cases that are ambiguous or open to interpretation cases depending on the criterion applied*

contrast (case 6). However, if the vehicle has a minimum size BB of 15 by 15 pixels (C1.1), a 3 pixel error results in a BB that is 44% larger than the 'real' size (see case 2, Table 3). Moreover, if the BB was drawn even 1 pixel smaller, it would not meet the criterion 1.1 anymore. This situation is common in the dataset, particularly for parked vehicles.

https://doi.org/10.1017/dce.2024.39 Published online by Cambridge University Press



(a) Faster R-CNN (b) YOLOv5 (c) DETR

Figure 3. *Calculated mAP50 for a) Faster R-CNN, b) YOLOv3, c) DETR; on the x-axis, there are the annotations used for the testing datasets, and the different colors stand for the different labels of the training and validation sets*

Table 4. *mAP_{50} of the three network models trained on the different criteria (rows) and tested on the different criteria (columns)*

		Testing dataset				
		C1	C2	C3	MoSeg	
Training dataset	C1	0.833	0.825	0.842	0.761	F. R-CNN
	C2	0.844	0.853	0.867	0.780	
	C3	0.779	0.796	0.888	0.806	
	MoSeg	0.796	0.809	0.882	0.802	
	C1	0.797	0.753	0.861	0.799	YOLOv5
	C2	0.814	0.844	0.876	0.806	
	C3	0.750	0.776	0.865	0.801	
	MoSeg	0.728	0.753	0.861	0.799	
	C1	0.678	0.664	0.699	0.620	DETR
	C2	0.752	0.755	0.781	0.700	
	C3	0.575	0.607	0.712	0.630	
	MoSeg	0.605	0.646	0.743	0.661	

Another common ambiguous scenario is occluded objects, in particular, cases 4, 7, and 9 in Table 3. For case 4, the object is labeled based on criterion 1.4. The BB may be highly subjective, as the visible part of the vehicle in the frame prevents the annotator to know the true object size and may require understanding of vehicle model and pose to truly estimate the occlusion amount. Additionally, case 9 identifies that the occlusion may hide key features of an object that neural networks may look for. Another peculiar case in the automotive field is that vehicles have windows that are transparent or translucent (case 7); hence, it is possible to see parts of an occluded object through the vehicle's window. In C2, the annotator handled these situations by including the area occluded but seen through the window; these BBs are again subjective, and the annotated objects might not have any visible features relevant to DNNs.

Figures 3a to 3c presents the mAP of the three different types of DNNs trained and evaluated on the datasets with the 4 different sets of annotations (different colors stand for the criteria used for the annotations of the training dataset). The values for the 48 combinations are reported in Table 4. When comparing testing on original MoSeg labels with respect to testing on C3 labels, all the networks performed better on C3. As the erroneous ground truth BBs are removed in C3 annotations, this BB

https://doi.org/10.1017/dce.2024.39 Published online by Cambridge University Press

reduction will also decrease the number of false negatives. However, due to the large difference in number of ground truth BBs in the test set between C3 and the other sets of annotations (~13% lower than the original MoSeg labels, Table 2), a bias can be created in the results and falsely indicate that the DNN is performing better in the real-world situation. An important finding is that removing the clearly wrong or redundant BBs improves the DNN computed performance metrics, improving the public confidence in automated vehicles' technology.

Based on the method of annotation, C2 ground truth reflects more accurately the real-world performance, as the annotator is identifying almost all the vehicles in each frame. The "realism" of C2 is then followed by C1, where manual labeling has a few stricter requirements. These requirements can be the basis for a semiautomated labeling procedure in the future. In all of the results when testing for "real world" performance (that is testing with C1 and C2), training with labels based on C2 performed the best, followed by C1. Compared to C1, C2 provides ground truth BBs for smaller objects and higher degrees of occlusion that would otherwise be filtered out. For all the network architectures, trained with C2 labels and tested in the case of *improved* labels (C1 and C2 for the testing set), the performance is always better than training with MoSeg labels. There are two factors likely providing the better performance: firstly, the higher number of training ground truth BBs in C2, and secondly, much more smaller (but accurate) BBs to train from.

For YOLOv5 and Faster R-CNN, the difference in performance between networks trained on MoSeg labels compared to C3 training the DETR DNN, MoSeg labels always perform better, independently of the labels used for testing. In fact, for the DETR model, see Figure 3c, training with C3 performed around 0.03–0.04 worse in terms of mAP than training with MoSeg labels. The DETR network requires a significant amount of training data, and overall, the selected dataset was small enough to enable the manual relabeling process (Carion et al., 2020). In fact, it is noticeable that DETR performs better with the relabeled dataset with the highest number of ground truth BBs (C2) and worst with the datset with fewer BBs (C3). It is expected that its overall performance can be improved with bigger datasets and with hyperparameter tuning, but the trends will remain the same. This reduced performance could be due to the lower number of training annotations in C3 dataset, or, given the different learning process in transformers, erroneous BBs may actually help the generalization of the DETR network.

Finally, overall, the trends and performance in the plots are similar for the one-stage and two-stage detectors, whereas, as mentioned before, they are different for the DETR. That again highlights that the learning process has an impact on the overall outcome. Moreover, even though transformers are supposed to have worst performance with small objects, when DETR is trained with C2 (including the smallest vehicle BBs), the performance is the best for all testing labels. In addition, training and testing on original MoSeg is never the best combination for all the architectures.

5. Conclusion

This work investigated the quality of the annotations in automotive datasets. Three sets of reannotation criteria are proposed based on the categories of errors identified by visual inspection of commonly used automotive datasets and their original BBs. The newly generated labels are then used to train different neural network architectures, in turn used to evaluate the four different set of labels (that is original and the three proposed criteria). Between the original labels and the reannotations, Criteria 2 (that is labeling carried out to the best of a human annotator) is likely the most representative of vehicles identification by a human driver. Overall, when predicting real-world situations, networks trained based on the original labels and also a set of labels where the obvious labeling errors were removed (Criteria 3), perform worse than the C1 and C2, set out more formally for the manual annotation. All three different architectures types of DNNs used in this work performed better with the stricter labels compared to the original labels. Setting some strict criteria and guidelines in the annotation process will have a positive effect on both the training and evaluation of the networks. However, the selection of the criteria parameters will affect the number of BBs in the annotations and the performance of the DNNs. The proposed criteria can be adopted and applied to any automotive datasets, and nonsubjective criteria can be automatically enforced.

https://doi.org/10.1017/dce.2024.39 Published online by Cambridge University Press

This work highlights that, in specific fields, it is imperative to ensure that the ground truth annotations are appropriately labeled for the specific use case, especially getting it right the first time using clear and enforceable criteria to avoid reannotation. In the case of AAD, the safety of the vehicle and the control decisions are important concerns. In this use case, all objects which may compromise the safety should be appropriately labeled in benchmarking datasets. However, the use case can vary from vehicle to vehicle, depending on the specific AAD function, the operational design domain, and the selected scenarios, and therefore it has an implication on which objects *have to* be detected. It may be unreasonable to create new datasets per use case, and hence, this work stresses the importance of clear criteria for annotations and the possibility to add metadata/sublabels to improve label quality for AAD. For example, in this work, by providing some metadata for C2, such as occlusion, percentage of occlusion, truncation, etc., an automated filtering process can be used to produce labels aligned to a set of criteria similar to C1. Metadata can also allow a further understanding of the learning process and the key features for learning and support a more automated process for producing use case-specific labels. Additionally, metadata can allow for an understanding of which types of objects in a class are less likely to be identified, and therefore, ad hoc data augmentation can be carried out for these cases.

Data availability statement. The reannotated labels for the three sets of criteria used in this work are released and can be accessed at: https://github.com/WMG-IV-Sensors/WMG_MoSeg.

Acknowledgements. The authors acknowledge Lucy Inett for her work on the labeling. Dr Donzella acknowledges that this work was supported by the Royal Academy of Engineering under the Industrial Fellowships scheme. The authors wish to acknowledge the support of High Value Manufacturing CATAPULT.

Author contribution. Please provide an author contributions statement using the CRediT taxonomy roles as a guide \ https://www.casrai.org/credit.html. Conceptualization: A.A. and A.B. Methodology: A.A. and A.B. Data curation: A.C. Data visualization: A.C. Writing original draft: A.A. and A.B. All authors approved the final submitted draft.

Funding statement. Cofunded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them. Project grant no. 101069576. UK participants in this project are cofunded by Innovate UK under contract no.10045139.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Angus, M, ElBalkini, M, Khan, S, Harakeh, A, Andrienko, O, Reading, C, Waslander, S and Czarnecki, K (2018) Unlimited road-scene synthetic annotation (URSA) dataset. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 985–992.

Arief, HA, Arief, M, Zhang, G, Liu, Z, Bhat, M, Indahl, UG, Tveite, H and Zhao, D (2020) Sane: smart annotation and evaluation tools for point cloud data. *IEEE Access 8*, 131848–131858.

Caesar, H, Bankiti, V, Lang, AH, Vora, S, Liong, VE, Xu, Q, Krishnan, A, Pan, Y, Baldan, G and Beijbom, O (2020) Nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Carion, N, Massa, F, Synnaeve, G, Usunier, N, Kirillov, A and Zagoruyko, S (2020) *End-to-End Object Detection with Transformers*. Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-030-58452-8_13

Chen, Y, Liu, F and Pei, K (2021) Cross-modal matching cnn for autonomous driving sensor data monitoring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3110–3119.

Cordts, M, Omran, M, Ramos, S, Rehfeld, T, Enzweiler, M, Benenson, R, Franke, U, Roth, S and Schiele, B (2016) The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Cui, Y, Chen, R, Chu, W, Chen, L, Tian, D, Li, Y and Cao, D (2021) Deep learning for image and point cloud fusion in autonomous driving: a review. *IEEE Transactions on Intelligent Transportation Systems 23*(2), 722–739.

Deng, L (2012) The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine 29* (6), 141–142.

Dokic, K, Blaskovic, L and Mandusic, D (2020) From machine learning to deep learning in agriculture – the quantitative review of trends. *IOP Conference Series: Earth and Environmental Science 614*(1), 012138.

Du, M (2023) Overview of autonomous vehicle. In *Autonomous Vehicle Technology*, New York: Springer, 1–15.

https://doi.org/10.1017/dce.2024.39 Published online by Cambridge University Press

Everingham, M, **Van Gool, L**, **Williams, CK**, **Winn, J and Zisserman, A** (2009) The pascal visual object classes (voc) challenge. *International Journal of Computer Vision 88*, 303–308.

Feng, D, **Harakeh, A**, **Waslander, SL and Dietmayer, K** (2021) A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems 23*(8), 9961–9980.

Geiger, A, **Lenz, P and Urtasun, R** (2012) Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR).*

Guo, J, **Kurup, U and Shah, M** (2020) Is it safe to drive? An overview of factors, metrics, and datasets for driveability assessment in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems 21*(8), 3135–3151.

He, K, **Zhang, X**, **Ren, S and Sun, J** (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

He, L, **Zhou, Q**, **Li, X**, **Niu, L**, **Cheng, G**, **Li, X**, **Liu, W**, **Tong, Y**, **Ma, L and Zhang, L** (2021) End-to-end video object detection with spatial-temporal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1507–1516.

Ingle, S and Phute, M (2016) Tesla autopilot: semi autonomous driving, an uptick for future autonomy. *International Research Journal of Engineering and Technology 3*(9), 369–372.

Janosovits, J (2022) Cityscapes tl++: Semantic traffic light annotations for the cityscapes dataset. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2569–2575.

Jocher, G, **Chaurasia, A**, **Stoken, A**, **Borovec, J**, **NanoCode012**, **Kwon, Y**, **TaoXie**, **Michael, K**, **Fang, J**, **imyhxy**, **Lorna**, **Wong, C**, **Yifu**, **ZVA**, **Montes, D**, **Wang, Z**, **Fati, C**, **Nadar, J**, **Laughing**, **UnglvKitDe**, **tkianai**, **yxNONG**, **Skalski, P**, **Hogan, A**, **Strobel, M**, **Jain, M**, **Mammana, Land xylieong** (2022) ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations.

Koopman, P and Wagner, M (2016) Challenges in autonomous vehicle testing and validation. *SAE International Journal of Transportation Safety 4*(1), 15–24.

Kukkala, VK, **Tunnell, J**, **Pasricha, S and Bradley, T** (2018) Advanced driver-assistance systems: A path toward autonomous vehicles. *IEEE Consumer Electronics Magazine 7*(5), 18–25.

Li, Y and Ibanez-Guzman, J (2020) Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Processing Magazine 37*(4), 50–61.

Li, Y and Shi, H (2022) *Advanced Driver Assistance Systems and Autonomous Vehicles: From Fundamentals to Applications*. New York: Springer Nature.

Lin, T-Y, **Dollár, P**, **Girshick, R**, **He, K**, **Hariharan, B and Belongie, S** (2017) Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125.

Ma, J, **Ushiku, Y and Sagara, M** (2022) The effect of improving annotation quality on object detection datasets: a preliminary study. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4849–4858.

Meyer, M and Kuschk, G (2019) Automotive radar dataset for deep learning based 3D object detection. In *2019 16th European Radar Conference (EuRAD)*, pages 129–132.

Northcutt, CG, **Athalye, A and Mueller, J** (2021) Pervasive label errors in test sets destabilize machine learning benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1).*

Park, H, **Park, K**, **Mo, S and Kim, J** (2021) Deep neural network based electrical impedance tomographic sensing methodology for large-area robotic tactile sensing. *IEEE Transactions on Robotics 37*(5), 1570–1583.

Rashed, H, **Mohamed, E**, **Sistu, G**, **Kumar, VR**, **Eising, C**, **El-Sallab, A and Yogamani, S** (2021) Generalized object detection on fisheye cameras for autonomous driving: Dataset, representations and baseline. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2272–2280.

Ren, S, **He, K**, **Girshick, R and Sun, J** (2015) Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, *39*(6), 1137–1149.

Retson, TA, **Besser, AH**, **Sall, S**, **Golden, D and Hsiao, A** (2019) Machine learning and deep neural networks in thoracic and cardiovascular imaging. *Journal of Thoracic Imaging 34*(3), 192.

SAE J3016_202104 (2021) *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. Warrendale (PA), USA: Standard, Sociaty of Automotive Engineers.

Sheeny, M, **De Pellegrin, E**, **Mukherjee, S**, **Ahrabian, A**, **Wang, S and Wallace, A** (2021) Radiate: a radar dataset for automotive perception in bad weather. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7.

Siam, M, **Mahgoub, H**, **Zahran, M**, **Yogamani, S**, **Jagersand, M and El-Sallab, A** (2017) MODNet: moving object detection network with motion and appearance for autonomous driving. *arXiv preprint arXiv:1709.04821.*

Tsipras, D, **Santurkar, S**, **Engstrom, L**, **Ilyas, A and Madry, A** (2020) From imagenet to image classification: Contextualizing progress on benchmarks. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Uřičář, M, **Křížek, P**, **Sistu, G and Yogamani, S** (2019) Soilingnet: Soiling detection on automotive surround-view cameras. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 67–72.

Valada, A (2022) Keynote: past, present, and future of scene understanding for automated driving. In *AutoSens Brussels 2022*. SenseMedia in Brussels, Belgium.

Wang, Z, **Acuna, D**, **Ling, H**, **Kar, A and Fidler, S** (2019) Object instance annotation with deep extreme level set evolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Xiao, H, **Rasul, K and Vollgraf, R** (2017) Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747.*

https://doi.org/10.1017/dce.2024.39 Published online by Cambridge University Press

Xu, H, **Gao, Y**, **Yu, F and Darrell, T** (2017) End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Yao, Z, **Ai, J**, **Li, B and Zhang, C** (2021) Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*.

Zaidi, SSA, **Ansari, MS**, **Aslam, A**, **Kanwal, N**, **Asghar, M and Lee, B** (2022) A survey of modern deep learning based object detection models. *Digital Signal Processing 126*, 103514.

Cite this article: Chan PH, Li B, Baris G, Sadiq Q and Donzella V (2024). The inconvenient truth of ground truth errors in automotive datasets and DNN-based detection. *Data-Centric Engineering*, 5, e34. doi:10.1017/dce.2024.39