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

Pak Hung Chan\textsuperscript{1}, Boda Li\textsuperscript{1}, Gabriele Baris\textsuperscript{1}, Qasim Sadiq\textsuperscript{1}, and Valentina Donzella\textsuperscript{3}

\textsuperscript{1}WMG, University of Warwick

January 26, 2024
The inconvenient truth of ground truth errors in automotive datasets and DNN-based detection

Pak Hung Chan¹, Boda Li¹, Gabriele Baris¹, Qasim Sadiq¹ and Valentina Donzella¹

¹WMG, University of Warwick, Coventry, CV4 7AL, West Midlands, United Kingdom.
*Pak Hung Chan, Boda Li and Gabriele Baris all contributed equally.

Keywords: Machine Learning, Automated Vehicles, Automotive Dataset, Labelling

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, colours, etc. Supervised learning needs annotated datasets, and several datasets are available for developing automotive functions. However, these datasets are tremendous in volume, and labelling 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 impact negatively vehicle safety when these neural networks are deployed. This work investigates 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 paper 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; (iii) remove clear erroneous bounding boxes. 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 relabelled datasets still present some cases with “arbitrary”/“controversial” annotations, and therefore this work concludes with some guidelines related to dataset annotation, metadata/sub-labels, 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, and 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 labelling in the automotive community.

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

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 [9, 31, 28]. Compared to traditional algorithms, machine learning algorithms
can provide a 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 which 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 [5].

### 1.1. Assisted and Automated Driving and Deep Neural Networks

Assisted driving functions are currently deployed in commercial vehicles, and automated driving is a major research area in the vehicle industry and in academia[17, 21, 23, 10]. 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 Level 0 to Level 5 [32]. 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 [20, 22, 7].

There are many datasets collected for developing, testing, and benchmarking some of the perception and prediction functions used to support assisted and automated driving tasks. They present annotated and labelled ground truth data, appropriate to the different perception/prediction tasks (i.e. bounding boxes (BBs) for object detection, pixel masks for segmentation) [14]. Commonly used and established benchmarking datasets are the KITTI [13], Berkeley DeepDrive (BDD) [40], nuScene, [3] and CityScapes [6]. These datasets publish and update regularly leaderboard tables to compare performance of novel neural networks (for different tasks), and all developers can submit their results. However, for automotive datasets, the improvement of DNN evaluation metrics is saturating, and it seems impossible to achieve “perfect” performance, e.g. as close as possible to 100% accuracy [37]. Any inaccuracies in the DNN outputs is an indication that the safety of the intended function might be compromised, leading to potentially hazardous situations. In the ML learning community is uncertain if a clear step forward towards improved performance will be achievable; in this context, this work aims to understand the role of ground truth and specifically if well-specified and accurate ground truth can contribute to a further improvement in DNN performance. This hypothesis is based on the discovery of several errors and inconsistencies in the above mentioned benchmarking datasets, which are discussed in Sec. 3.2.

### 1.2. Contributions

This paper discusses, analyses and classifies the annotation errors in widely used automotive datasets, namely KITTI and nuScene [34, 13, 3]. 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 set 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:

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 are dependent on the DNN architecture and the annotation criteria, but overall training with more accurate labels seems beneficial;
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 transformer;
4. proposing different sets of labelling 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 bounding boxes 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 bounding boxes labels may also hinder the maximum potential of the DNN, affecting the training loss and the adjustments of weights.

2. Background

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

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

2.1. Datasets and Annotation Criteria

Datasets have moved from simple classification tasks of the MNIST dataset, where each sample contains one class, to current big curated datasets which include frames with multiple objects and different annotations and classes [8, 39]. Existing works on dataset annotation cover different issues and they 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 [36], RADIATE [33] or nuScenes [3]. nuScenes has also published instructions for their labelling, however these labels are open to interpretation and produce errors similar to the ones identified into 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 [1, 18, 26]. These methods of annotation are otherwise known as semi-automatic. Further techniques include the development of new annotation tools to adapt current datasets to specific use cases in AAD [2, 38]. 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 labelling [11]. The guidelines provide guidance on which images to label and how they should be categorised. Examples include: categorising an object as occluded if more than 5% of the object within a bounding box is
occluded, and there are images considered too difficult to segment and left unlabelled 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. There are recent works 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 re-annotated the MS COCO and Google Open Images Datasets by providing a description of the object as well as providing some common examples of positive or negative cases which are open to the interpretation of the annotators [25]. They have trained 5 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 re-annotation, 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 machine learning to understand and categorise the errors in labelling [27]. 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 [35]. The original ImageNet dataset provided only one class per image, however, each image may contain several secondary objects or clutter. Tsipras et al. relabelled a subset of the ImageNet database (10 images per class, 10000 images in total) and provided multi class annotations where appropriate in their subset. The neural network performed worse on images which 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 (i.e. identification of dog breeds by non specialists).

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 [12, 29]. Object detectors can be broadly divided into two classes, moreover recently there have been several implementations of object detectors using transformers [16, 4]; a comprehensive review of object detectors is given in [42]. Vision transformers are becoming very popular due to their promising performance, however they usually require more epochs to converge and usually do not perform well on small objects [41]. 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 with lower accuracy with respect to 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 re-annotating the datasets, are common across different architectures for the same perception task, i.e. 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 re-annotated using three different sets of proposed criteria by the authors, i.e. C1, C2, and C3, see Section 3.2. Three DNNs were fine tuned on the dataset
Table 1: The three sets of criteria for the re-annotation of the KITTI MoSeg dataset for object detection.
* Incorrect stands for BBs that clearly do not belong to any target objects

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

four 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 (i.e. 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 [34, 13]. Compared to the KITTI dataset, KITTI MoSeg provides chronological sequences of frames, totalling to 1449 frames [34]. The KITTI MoSeg dataset provides 2D bounding box 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, the original annotations were expanded by using the information available in the original dataset (3D bounding box, odometry information), where 3D ground truth bounding boxes were converted into 2D bounding boxes, associating bounding boxes across chronological frames to obtain estimated velocity per bounding box (BB). Finally, a filtering process was applied to keep objects consistently identified across frames [34]. As the experiments proposed in this paper entail manually re-annotating 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.

Table 2: Number of ground truth bounding boxes for each set of annotations, with C# denoting the number of the set of criteria (1-3), and MoSeg denoting the original labels

<table>
<thead>
<tr>
<th></th>
<th>MoSeg</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>4302</td>
<td>4384</td>
<td>6025</td>
<td>4047</td>
</tr>
<tr>
<td>Validation</td>
<td>509</td>
<td>483</td>
<td>653</td>
<td>467</td>
</tr>
<tr>
<td>Testing</td>
<td>2648</td>
<td>2889</td>
<td>2852</td>
<td>2315</td>
</tr>
<tr>
<td>Total</td>
<td>7459</td>
<td>7756</td>
<td>9530</td>
<td>6829</td>
</tr>
</tbody>
</table>
Table 3: Examples of ambiguous situations when applying the proposed criteria, related visual examples are given in Figure 2.

<table>
<thead>
<tr>
<th>Case No.</th>
<th>Description</th>
<th>Criterion applied</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Small but obvious objects to a human are not labelled</td>
<td>1.1</td>
</tr>
<tr>
<td>2</td>
<td>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)</td>
<td>1.2</td>
</tr>
<tr>
<td>3</td>
<td>Objects meeting this criterion can have most of their surface occluded</td>
<td>1.3</td>
</tr>
<tr>
<td>4</td>
<td>Judging the percentage of occlusion is subjective</td>
<td>1.4</td>
</tr>
<tr>
<td>5</td>
<td>Small object annotation by humans is influenced by the annotator understanding and interpretation of the scene</td>
<td>2.1</td>
</tr>
<tr>
<td>6</td>
<td>In low contrast situations the judgment of object boundaries becomes arbitrary</td>
<td>2.2</td>
</tr>
<tr>
<td>7</td>
<td>Object through windows can be annotator dependent, and BB may include distortion and extensive occluded areas</td>
<td>2.3</td>
</tr>
<tr>
<td>8</td>
<td>It can be subjective to judge which BB is “correct” or “incorrect”</td>
<td>3.1</td>
</tr>
<tr>
<td>9</td>
<td>Annotated obstructed objects may not have any feature useful for recognition</td>
<td>3.2</td>
</tr>
</tbody>
</table>

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 categorised using the definitions below:

- **Missing** - There are clear and obvious vehicles in the frames which are not labelled.
- **Incorrect** - There is not an object of the labelled class in the bounding box.
- **Bad Fit** - The bounding box size and/or position are not appropriate for the identified object.
- **Occlusion** - The bounding box 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 Figure 1. These sets of criteria were implemented one at a time to completely re-annotate the KITTI MoSeg manually, using an ad hoc Matlab app developed by the authors to re-annotate the dataset. Although human labelling error can still exist, the some of the criteria can be coded to ensure the drawn bounding box 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, 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 (i.e. wheels, shapes, etc.) C1.1 specifies the minimum dimension of a bounding box based on an analysis of box sizes in the dataset. 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 human inaccuracies to pick out the exact edge of an object. 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 [3].

In Criteria 2, 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
Figure 2: Examples of cases that are ambiguous or open to interpretation cases depending on the criterion applied.

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, the annotator did not resize nor add any BBs, but only removed the BBs not belonging to real objects; in the case of multiple bounding boxes encompassing the same vehicle, the annotator kept only the bounding box 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 [30], YOLOv5 [19], and DETR [4]. Faster R-CNN is an example of two-stage detector, which predict bounding boxes from region proposals. On the other hand, YOLO is a single-stage detector, that does bounding box regression from anchors. Finally, DETR directly performs bounding box prediction with respect to the input image. For each model, described in the list below, the training process was performed four times from the base model (one for each dataset variation), 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 [15, 24]. It was originally pre-trained on COCO and then fine-tuned over the dataset of interest. The training was performed using the AdamW optimiser with learning rate set to $10^{-3}$ 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 [15]. It was originally pre-trained 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 $10^{-4}$.

### 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 re-annotation criteria are compared using the three main types of neural network based object detectors.

### 4. Results and Discussion

The number of ground truth bounding boxes was computed for each relabelled dataset, see Table 2. According to Criteria 3, which seeks to remove the incorrect or fully occluded bounding boxes, there are 612 bounding boxed (8.2%) which were deemed to be incorrect in the original dataset. Criteria 2 produced the largest amount of bounding boxes based on what the human annotator could identify. Many of C2 bounding boxes differ from BB identified via C1, due to objects being too small, or occlusion too high.

During the re-annotation process, there were some difficulties in adhering strictly to the identified criteria and some labels can be subjective or open to interpretation. 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 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 bounding box that is 44% larger than the ‘real’ size (see case 2, Table 3). Moreover, if the bounding box 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.

Another common ambiguous scenario is occluded objects, in particular, cases 4, 7 and 9 in Table 3. For case 4, the object is labelled 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 Criteria 2, 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 present the mAP of the three different types of DNNs trained and evaluated on the datasets with the 4 different sets of annotations (different colours stands for the criteria used for the annotations of the training dataset), the values for the 48 combinations are reported in Table 4. When
Figure 3: Calculated mAP50 for a) Faster R-CNN, b) YOLOv5, c) DETR; on the x-axis there are the annotations used for the testing datasets and the different colours stand for the different labels of the training and validation sets.

comparing testing on original MoSeg labels with respect to testing on Criteria 3 labels, all the networks performed better on C3. As the erroneous ground truth BBs are removed in C3 annotations, this BB 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 Criteria 3 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 overall 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, Criteria 2 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 Criteria 1, where manual labelling has a few stricter requirements. These requirements can be the basis for a semi-automated labelling procedure in the future. In all of the results when testing for “real world” performance (i.e. testing with Criteria 1 and 2), training with labels based on Criteria 2 performed the best, followed by Criteria 1. Compared to Criteria 1, Criteria 2 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 Criteria 2 labels and tested in the case of improved labels (C1 and C2 for the testing set) the performance are 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 Criteria 2, and secondly much more smaller (but accurate) bounding boxes to train from.

For YOLOv5 and Faster R-CNN, the difference in performance between networks trained on MoSeg labels compared to Criteria 3 is low, and in Faster R-CNN, training with Criteria 3 performs worse compared to training with MoSeg labels. When considering original MoSeg labels versus Criteria 3 ones (only removal of erroneous labels) for 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 Criteria 3 performed around 0.03- 0.04 worse in terms of mAP than training with MoSeg labels. This effect could be due to the lower number of training annotations in Criteria 3 dataset, or, given the different learning process in transformers, erroneous BBs actually help the generalisation 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 are the best for all testing labels. In addition, training and testing on original MoSeg is never the best combination for all the architectures.
Table 4: $mAP_{50}$ of the three network models trained on the different criteria (rows) and tested on the different criteria (columns).

<table>
<thead>
<tr>
<th>Testing dataset</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>MoSeg</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0.833</td>
<td>0.825</td>
<td>0.842</td>
<td>0.761</td>
</tr>
<tr>
<td>C2</td>
<td>0.844</td>
<td>0.853</td>
<td>0.867</td>
<td>0.780</td>
</tr>
<tr>
<td>C3</td>
<td>0.779</td>
<td>0.796</td>
<td>0.888</td>
<td>0.806</td>
</tr>
<tr>
<td>MoSeg</td>
<td>0.796</td>
<td>0.809</td>
<td>0.882</td>
<td>0.802</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training dataset</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>MoSeg</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0.797</td>
<td>0.753</td>
<td>0.861</td>
<td>0.799</td>
</tr>
<tr>
<td>C2</td>
<td>0.814</td>
<td>0.844</td>
<td>0.876</td>
<td>0.806</td>
</tr>
<tr>
<td>C3</td>
<td>0.750</td>
<td>0.776</td>
<td>0.865</td>
<td>0.801</td>
</tr>
<tr>
<td>MoSeg</td>
<td>0.728</td>
<td>0.753</td>
<td>0.861</td>
<td>0.799</td>
</tr>
</tbody>
</table>

| C1               | 0.678 | 0.664 | 0.699 | 0.620 |
| 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 |

5. Conclusion

This work investigated the quality of the annotations in automotive datasets. Three sets of re-annotation criteria are proposed based on the categories of errors identified by visual inspection of commonly used automotive datasets and their original bounding boxes. The newly generated labels are then used to train different neural network architectures, in turn used to evaluate the four different set of labels (i.e. original and the three proposed criteria). Between the original labels and the re-annotations, Criteria 2 (i.e. labelling 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 labelling errors were removed (Criteria 3), perform worse than the Criteria 1 and 2, 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 in both the training and evaluation of the networks. The proposed criteria can be adopted and applied to any automotive datasets, and can be automatically enforced.

This work highlights that, in specific fields, it is imperative to ensure that the ground truth annotations are appropriately labelled for the specific use case, especially getting it right the first time using clear and enforceable criteria to avoid re-annotation. In the case of assisted and automated driving, safety of the vehicle and the control decisions is an important concern. In this use case, all objects which may compromise the safety should be appropriately labelled 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 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/sub-labels to improve label quality for AAD. For example, in this work, by providing some metadata for Criteria 2, 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 Criteria 1. Metadata can also allow a further understanding into 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.
Acknowledgments. The Authors acknowledge Lucy Inett for her work on the labelling. 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.

Funding Statement. This work is partially funded by the European Union (grant no. 101069576). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them. The UK participant (WMG) in this project is supported by Innovate UK (contract no. 10045139).

Competing Interests. None

Data Availability Statement. The re-annotated labels for the three sets of criteria used in this work are released and can be accessed at: \url{https://github.com/WMG-IV-Sensors/WMG_NoSeg}.

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

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

References


