SHARD: Safety and Human Performance Analysis for Requirements in Detection

Ken T. Mori and Steven Peters

Abstract—**Automated driving requires reliable perception of the environment to ensure the safety of the driving task. One common perception task is 3D object detection, which aims at perceiving location and attributes of dynamic objects. This task is typically evaluated on different benchmark datasets, which each propose different metrics. However, these different metrics generally lack consistency and bear no relation to safety. Most notably, there is a lack of consistent definitions of pass/fail criteria for any given detection metric. In this work, the issue is addressed by systematically considering safety and human performance across different aspects of the object detection task. This approach yields interpretable detection metrics as well as thresholds for pass/fail criteria. Furthermore, a validation approach leveraging a prediction network is introduced and successfully applied to the requirements. A comparison of existing detectors shows that current perception algorithms exhibit failures for a majority of objects on the nuScenes dataset. Therefore, the results indicate the necessity of explicit safety consideration in the development of perception algorithms for the automated driving task.**

Index Terms—**Environment perception, object detection, requirements, testing.**

I. Introduction

Recently, there has been considerable interest in the field of automated driving (AD) [41]. However, the introduction of AD requires a safety assurance which demonstrates a positive balance of risk [24]. This safety assurance of AD is commonly considered in terms of safety outcomes of driving such as the frequency of accidents [40], [67].

While this approach is applicable to test the entire AD system, it is not applicable to all components of a modular architecture. However, testing an AD system with respect to its internal structure is beneficial to understand performance bounds [89]. This is in line with the demands of the safety of the intended functionality (SOTIF) to test a perception subsystem separately [37]. We assume the simple and common system architecture of a functional decomposition into Sense-Plan-Act, where the sense module includes the task of perception [2]. The interface between perception and planning is commonly represented as an object list [36]. The corresponding perception task of classifying and localizing objects is object detection, for which deep neural network (DNN) have emerged as successful solution [27].

Both the training and testing of detectors is typically performed on datasets [4]. These evaluate detectors in an offline and open-loop setting separately from the driving function. Typical benchmark datasets for 3D object detection in automotive context include [12], [29], [87]. Such datasets generally define the perception task as well as their own metrics for the evaluation. In doing so, datasets have previously driven research in the field of computer vision including the task of object detection [52].

However, common detection evaluation metrics neglect the aspect of safety for the driving task [95]. This raises the concern that perception systems may be underspecified with regard to safety [21]. While proposals such as [69], [95] have been made to incorporate safety, the metrics remain inconsistent and have not been widely adopted. In addition, human performance on these metrics is currently largely unknown [74] and not considered. Finally, previous efforts have focused on the definition of metrics without defining clear requirements in the form of pass/fail criteria.

Therefore, the objective of this work is to incorporate safety into 3D object detection metrics. Specifically, we present three contributions. Firstly, we propose and apply a principled methodology to elicit detection requirements. Secondly, we provide a validation methodology and validation results for these requirements. Finally, the proposed requirements are applied to baseline object detectors to evaluate their performance.

The methodology reconsiders the entire object detection pipeline by decomposing it into interpretable aspects. For each aspect of tracking, association, localization and velocity, human detection performance is estimated. The human errors are paired with conservative estimates to include safety. This yields quantitative requirements for each aspect, which are substantiated by an argumentation. The validation applies a recent validation methodology which leverages a DNN prediction component. Requirements are considered valid if they do not affect the prediction component trained on human behavior. This successfully reconciles the requirements with context-aware DNN, ensuring interpretability, simplicity and validity.

Manuscript received 8 August 2023; revised 28 August 2023 and 14 September 2023; accepted 22 September 2023. Date of publication 28 September 2023; date of current version 23 February 2024. This work was supported in part by *VIVID*, promoted by the German Federal Ministry of Education and Research, based on a decision of the Deutsche Bundestag under Grant 16ME0173, and in part by the Open Access Monographs Fund of the University and State Library Darmstadt. *(Corresponding author: Ken T. Mori.)*

The authors are with the Institute of Automotive Engineering, Technische Universität Darmstadt, 64287 Darmstadt, Germany (e-mail: ken.mori@tu-darmstadt.de; steven.peters@tu-darmstadt.de).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2023.3320395.

Digital Object Identifier 10.1109/TIV.2023.3320395

II. Related Work

This section first focuses on the aspect of safety and reliability as defined in literature. Subsequently, prevalent perception metrics on benchmark datasets and recent safety aware metrics are discussed.

A. Reliability and Safety

Generally, the concept of reliability deals with the concept of failures, while safety deals with the consequences. Safety is the capacity of the system to not endanger persons for specified time and conditions. Reliability is the ability or probability of a system to perform its functions as specified by the requirements without failure [94]. The probability expressed as failure rate can be used for the estimation of the reliability [38]. A risk is understood as the severity of a failure multiplied with its probability expressed as a failure rate [84].

Within the context of automated driving, accidents have been used for the risk assessment of automated driving functions [97]. An alternative is to use criticality metrics as surrogate metrics to quantify the risk of actors [101]. Specifically for the context of perception, the SOTIF considers performance limitations of the systems which may lead to hazardous behavior. The SOTIF explicitly lists examples for perception limitations such as incorrect classification, incorrect measurements, incorrect tracking or misdetection [37]. However, no metrics for quantifying these aspects are provided.

B. Benchmark Dataset Evaluation Metrics

Perception evaluation is typically performed on benchmark datasets, which aim to produce a ranking between different methods. The object detection and the tracking task have distinct metrics over which a brief overview is provided.

The first step in computing a metric is to associate the estimation with the ground truth based on a distance metric. Positive and negative samples are defined based on arbitrary thresholds [36]. The common object detection metric based on this concept is average precision (AP) [65], first introduced by the Pascal Visual Object Classes (VOC) challenge. The AP is obtained by using a detectors ranked output and averaging precisions over multiple recall levels [25]. Common modifications such as mean average precision (mAP) include averaging over different matching thresholds and classes [20], [53].

The multiple object tracking accuracy (MOTA) metric [7] is the most commonly applied metric for tracking [57]. In addition to false positive (FP) and false negative (FN) samples considered in detection, mismatches are also included in a sum which is then divided by the number of ground truth (GT) objects [7]. Similar to detection, the MOTA can be modified by integrating over multiple recall values or scaling with the recall [57]. The higher order tracking accuracy (HOTA) metric proposes to provide and re-weight different interpretable components for tracking [56].

True positive (TP) accuracy such as localization accuracy can be indirectly integrated by averaging over different matching thresholds. However, more direct consideration is given by metrics such as multiple object tracking precision (MOTP), which assesses the localization precision of matched objects [7].

Automotive datasets for 3D object detection and tracking are heavily influenced by these prevalent metrics originating from 2D computer vision. AP has become the de facto standard adopted directly by KITTI [29] and A3D [68]. Variants of AP have been proposed by modifying the matching procedure [12], [87] or emphasizing object orientation [59], [87]. As with the detection task, variants of MOTA metrics are adopted to the automotive domain with only slight modifications [12], [15]. One notable exception is nuScenes [12], which incorporates true positive (TP) metrics into a weighted average named nuScenes detection score as its main metric.

Overall, typical perception metrics emphasize the average performance on a single measure without consideration of safety [95], [102].

C. Safety Aware Metrics

While not typically included in benchmarks, other metrics considering the driving task and its safety have been proposed. While they are not within the scope of this work, several works such as [6], [71], [76], [91], [95] attempt to focus the perception evaluation on objects relevant for collision avoidance.

Other works attempt to consider safety aspects by using heuristics. Examples include weighting a safety metric with the perception time [95]. Other options include directly evaluating the time until the first detection or between two detections of the same object [12]. Other works additionally consider safety by including a time to collision (TTC) either to weight objects [103] or for visual comparison [58]. The question of metric thresholds is addressed in the context of associating detected and GT objects. Different thresholds on egocentric distance [5] or longitudinal and lateral distance to the ego [23] have been proposed. However, even if thresholds are provided, they remain arbitrary.

Attempts to avoid heuristics are made by directly considering the downstream task of planning. The Planning Kullback-Leibler divergence (PKL) metric is proposed to consider the effect of detection errors on a planner to judge their severity [69]. It has since been adopted by the popular nuScenes detection benchmark [64]. Similar concepts have also been applied to study the sensitivity of a specific planner to perturbed perception results [33], [113]. Alternatively, this process can be used to derive acceptable perception perturbations for a single scenario [70]. However, this approach is limited by the availability of the planner [103]. Furthermore, it suffers from ambiguities and challenges present in the planning task [31]. In addition, the validity is limited to the specific implementation of the planner [70]. While safety is considered in these metrics, no generally applicable pass/fail criteria for perception are identified.

III. Method

First, the objectives and general assumptions are discussed. Next, perception requirements are elicited for tracking and localization as well as velocity estimation which are relevant to collision avoidance.

A. Method Overview

In this section, a brief overview over the method is provided. Before the actual methodology, general considerations such as the objectives and the principles applied are presented. This is followed by the method proposed in this work. First, different aspects related to the tracking of objects are incorporated. Secondly, the association procedure between GT objects and detections is considered in detail. Finally, the requirements for the collision relevant attributes of an object are derived. Hereby, different attributes such as localization and velocity are considered separately.

B. General Considerations

This work assumes an offline evaluation of a 3D object list with bounding boxes. Detection confidence is not required, as failures are determined for a single object list. If confidence scores are present, the confidence threshold is optimized with regard to the final metric as in [65], [69].

Following ideas from localization recall precision (LRP) [65] and nuScenes TP metrics [12], each failure is designed to be interpretable. To allow interpretability, each attribute is considered separately. The objective is to identify metrics with thresholds defining pass/fail criteria for each attribute. All errors and corresponding requirements indicate standard L1 distances for the physical attributes in one-dimensional metric space. For brevity, only attributes which are directly relevant to collision avoidance are presented. More specifically, the attributes considered are localization and velocity.

To incorporate safety, conservative estimates are applied in this work. As noted by [5], the direction of an error may not be symmetrical regarding the safety consequences. For instance, overestimation of the distance to an object is more dangerous than underestimation. To account for this fact, this work only allows errors in direction of the conservative estimate which is also defined for each criterion.

The human baseline has previously been considered for driving performance [67] and as reference to plausibilize perception metrics [69]. In this work, we leverage the human performance to distinguish acceptable from unacceptable errors and thus define pass/fail criteria. In order to guarantee a positive risk balance, the human performance may be overestimated, but never underestimated. Therefore, the more accurate estimate and representation are selected if different values are available in literature. While both random and systematic errors may contribute, this work chooses the maximum of the two instead of a sum. This underestimates the human error and thus provides a conservative overestimation of the human performance.

C. Tracking

This section specifically considers aspects regarding tracked objects list and temporal aspects.

1) Identifier Switches: One attribute commonly evaluated in tracking metrics such as [7], [82] is the identifier switch, where a wrong identifier is assigned to a correct detection [108].

For human perception, it has been shown that large changes can go unnoticed if an interruption occurs. This phenomenon was demonstrated for saccades, blank images, mud splashes and cuts or pans in motion pictures [78]. Change blindness even occurs in real-world settings when the subject is paying attention [79]. These findings indicate that unique identification is not performed by humans and thus not required for the task of driving. Therefore, this work will not evaluate identifier switches.

2) Tracking Accuracy: The popular MOTA metric neglects the order and temporal distribution of failures [7]. However, this aspect is relevant to the task of driving since it affects the available reaction time [95]. Temporal requirements can be considered explicitly by re-weighting performance [44], [95], by direct evaluation [12] or implicitly by considering the completeness of a track [50].

Human performance for perception times is directly accessible by measuring event-related potentials during perceptual tasks. For the presence of natural and artificial object categories, perception times of approximately 150 ms are obtained [90], [93]. Changes in geometric constellations are detected at approximately 200 ms [45]. Motion detection yields perception times of 160–200 ms in various studies depending on the type of motion [46].

Overall, converging evidence from human perception shows the possibility of object or motion detection starting from 150 ms. This work therefore neglects false negatives within the first 150 ms after initiating a track and false positives within 150 ms of terminating a track. Perception errors within a track are considered fully regardless of duration, since no contrary evidence from human perception is found. The number of objects humans can simultaneously track is limited to single digits [1], [72]. However, this aspect is neglected in this work, since it is unclear which objects a human tracks in a given traffic situation.

D. Association

Evaluating perceived object detection requires an association with the GT to define FP and FN [36]. Given a pairwise matching, any ground truth object with a matched perceived object is considered a TP while a ground truth object without a matched perceived object is considered a false negative. Similarly, a perceived object without matched ground truth object is considered a false positive [12], [29].

1) Association and Classification: The most common evaluation procedures perform association for each class [53]. However, the requirement of obstacle detection supports a class-agnostic detection [39], which also aligns with human perceptual mechanisms [96]. Additionally, unknown classes and fuzzy borders between classes may occur [14]. The concept of class-neutral objectness receives additional support from its successful application by object detectors [75], [80], [105], [111], [112]. Therefore, this work performs association irrespective of class.

2) Reference Point: While object centers have been used as reference points for object location [12], other reference points are possible. Consider the following scenario in Fig. 1, where the difference between using the closest points and the center

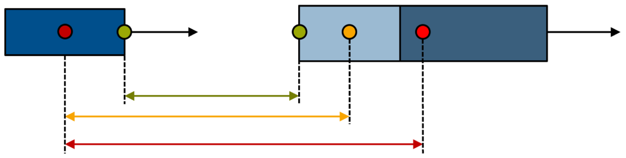


Fig. 1. Difference between using closest point (green) and center point (yellow/red) as reference.

points of two objects is visualized. Yellow and red indicate the center points for two possible lengths of the front vehicle. The relevant attribute for safety is the minimum distance to the other vehicle [113] shown in green. This corresponds to the available space for an emergency brake. Notably, the relevant distance between the closest points is independent of the object size.

Since the size estimation can prove difficult [99], any dependency between location and size is undesired. Sensor data does not include center points of objects, but only the surfaces [17], [73], [99]. Therefore, centers of object faces [112] or closest corners [110] are better suited for object localization. Overall, closets points simplify detection and correspond to the available space. Therefore, the point of the object closest to the ego vehicle is used as reference point in this work.

3) Association Procedure: Human perception displays behavior indicating an object-centric visual working memory [55] with features bound together by a location map [96]. While erroneous recombination of features occur [11], [92], these binding problems are limited to specific preconditions such as very brief presentations [104]. This object-centric approach using pairwise association with distance metrics is also common practice on benchmark datasets [12], [29], [87] and therefore adopted for this work.

The two common distance metrics are intersection over union (IoU) [29] and the distance between center points [12]. Since IoU couples localization, size and orientation of the object [12], it does not fulfil the requirement of interpretability. Additionally, all objects without overlap receive an equal IoU of zero [114]. In these cases, differences in performance cannot be distinguished. This effect is particularly pronounced for small objects [12], [107]. Therefore, a point distance is used in this work. However, as elaborated in the previous section, the closest point is used as reference instead of the center. From this point, the closest distance to the perceived object d_M is evaluated as shown in Fig. 2.

Different association methods such as greedy matching [12] and bipartite matching with the Hungarian algorithm [87] are available. Since the outcome is similar for both strategies, the greedy matching is applied for simplicity.

4) Association Threshold: In addition to the metric, it is also common to set a maximum threshold for association. This border between existence or matching and localization is generally fuzzy [36]. However, it corresponds to existing object detection pipelines [115]. Since it also increases interpretability, we propose to follow the approach of distinguishing matching and localization failures.

For simplicity, a circular distance threshold is applied as visualized in Fig. 2. The threshold is set in accordance with the most lenient localization accuracy criterion to allow distinguishing association and localization errors. As shown later in section III-E, this is the accuracy of the distance d between the GT object and the ego vehicle. Here, only the final result for the corresponding distance error $\Delta d \leq 0.15 \cdot d$ is applied. The process of obtaining this requirement is elaborated in Section III-E1. In addition, a minimum permissible offset is proposed to counteract the otherwise unreasonably small thresholds at small distances. The radius is chosen to be 2 m in accordance with previous work [12], [28]. This leads to a radius of $r_\mathrm{M} = \max(0.15 \cdot d, 2\ \mathrm{m})$ with d being the distance of the GT object to the ego vehicle. A perceived object is considered matched with the GT object if the matching distance $d_\mathrm{M} < r_\mathrm{M}$. Note that the exact distinction between matching and localization failures is inconsequential in this work, since both are equally considered in the final evaluation.

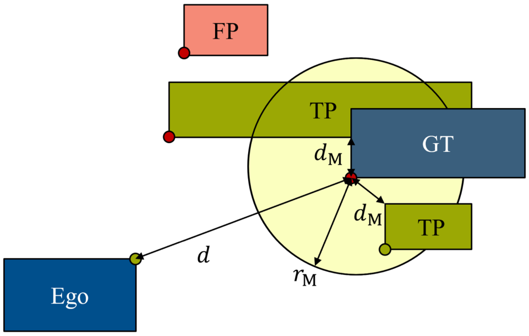


Fig. 2. Association based on circular distance threshold r_M around the closest point of GT. Green objects are potential matching candidates while red objects are beyond the matching threshold.

E. Localization

This section substantiates the localization requirements applied for the association. Distance to the ego vehicle and angular positions are considered separately due to correspondence with human perception literature.

1) Distance: As argued in previous sections, the closest point distances are most relevant for the safety of the driving task. Distance estimation errors generally include random and systematic errors which depend on the distance [22]. However, the relative error is $> 15\%$ for interobject distances [49] as well as for egocentric distance estimation in open terrain [22] or in a road environment [86]. At distances below 20 m, the context of the car may introduce additional bias with errors of approximately 40% [62]. For the task of direct depth labeling on monocular images, relative errors lie above 20% [106].

The conservative estimate to ensure safety prohibits overestimating the distance while underestimation is permissible. In accordance with human perception, the maximum permissible underestimation is defined as:

$$\Delta d = d_\mathrm{GT} - d_\mathrm{PRED} \leq 0.15 \cdot d \tag{1}$$

Note that the permissible error is always larger than zero if a distance of zero corresponding to an accident is avoided.

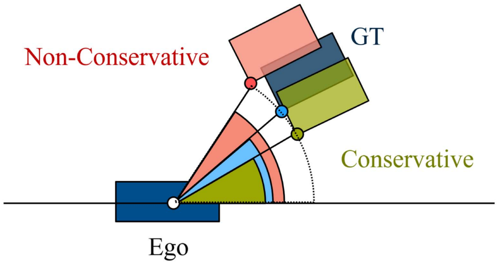


Fig. 3. Angular position referencing ego heading and conservative estimates.

2) Angular Position: One option is to consider the lateral distances of objects [77], [83]. However, human performance depends on the visual angle [49] which motivates evaluating angles in this work. Human estimation accuracy is consistent across viewing conditions such as fixed gaze, fixed head or free head [51]. For simplicity, this work therefore chooses the heading of the car as reference. The object point with smallest angle is used as conservative reference point which may differ from the point with closest distance. This corresponds to the intuitive notion that an object in front of the ego is more critical than further to the side. While systematic and random errors depending on the experimental setup exist, multiple studies show random errors of $> 5^\circ$ across a range of visual angles in both horizontal and vertical direction [34], [51].

For simplicity, a constant permissible error of 5° is chosen across the whole value range of visual angles for both azimuth and elevation. Conservative estimates demand that the smallest angle between object and ego heading in front or back may be underestimated, but never overestimated. A visualization is provided in Fig. 3.

F. Velocity

In this work, relative velocities are evaluated since they are more informative than absolute velocities regarding potential collisions. Angular and radial velocity correspond to the distinct aspects of possibility and timing of collisions and are therefore evaluated separately.

1) Radial Velocity: The radial velocity is considered as a TTC since humans estimate TTC directly based on the expansion rate of the image τ rather than by estimating speed [109]. For interceptive movements, humans can achieve high temporal precision by making continuous adjustments to the movement and the estimate. Thus, the accuracy increases as the distance decreases over time [8], [9]. However, this accuracy does not transfer to the driving task where collisions are avoided. For vehicle following scenarios, the TTC is typically underestimated by 30% or more for various settings [13], [43]. The standard deviation is substantial at more than 10% across different TTC values [35]. The inverse TTC given by $1/\text{TTC} = \text{iTTC} = v/d$ is applied to avoid unbounded values for the TTC if the velocity approaches zero. If accidents are avoided, the distance never reaches zero and the inverse time to collision (iTTC) thus remains bounded. In the following, the radial velocity is positive when an object is moving towards the vehicle. For evaluation, the ground truth distance is used to disentangle velocity and distance estimation. Converting the TTC data from literature [13], [35] to iTTC yields at least 10% relative error across different experimental settings.

Conservative estimates demand that the radial velocity towards the vehicle and therefore also the iTTC must never be underestimated. It should be noted that this requirement applies for positive and negative radial velocities. As the velocity approaches zero, humans show a perception threshold under naturalistic conditions such as braking given by $\text{iTTC}_{\text{low}} = 0.2\frac{1}{\text{s}}$ [60]. This threshold value is added to the permissible error leading to an overall permissible error of:

$$\Delta\text{iTTC} = 10\% \cdot \text{iTTC} + \text{iTTC}_{\text{low}} \tag{2}$$

2) Angular Velocity: Generally, performance differences regarding angular velocity estimation can be observed depending on various factors such as luminance [88] or velocity [10]. However, the error lies above a threshold of 5% across different settings [10], [18], [32], [88].

When considering relative velocities, a low tangential velocity is required for a collision. This means that low angular velocities may lead to a collision when closing in on an object, while high angular velocities mean the object will pass. Therefore, underestimation of the absolute angular velocity is permissible, while an overestimation is not. As the angular velocity approaches zero, humans exhibit a motion detection threshold which lies above $\dot{\Theta}_{\text{low}} = 0.03^\circ/s$ for different settings [63], [81]. Adding this threshold to the permissible error leads to:

$$\Delta\dot{\Theta} = 5\% \cdot \dot{\Theta} + \dot{\Theta}_{\text{low}} \tag{3}$$

IV. Validation

In the previous section, quantitative requirements are developed for different aspects of the driving task. While the requirements are all based on an argumentation, further validation is required.

To provide this validation, we apply the methodology and the implementation of SURE-Val [85]. This approach is based on a motion prediction network pre-trained on human trajectories, which is applied to two types of input. By contrasting the unmodified GT input with a perturbed object list as input, the effect of the perturbations is visible. Note that despite the similarities, the objective is not to directly use a downstream task as done by PKL [69]. Rather, the prediction network acts as a proxy for human behavior. If the perturbation of the object list has a discernible change in prediction performance, the corresponding requirement is considered invalid. Following SURE-Val [85], the change in prediction performance is assessed by calculating the p-values of testing for equality of distributions with the Cramer-von Mises [3] test. Small values indicate that the input perturbation leads to a change in prediction performance and is therefore not valid.

Originally, the SURE-VAl methodology is intended to evaluate the relevance of objects. In this work, it is modified by perturbing the object list with location offsets instead of removing objects. For any given localization requirement, the position of all objects in the object list is modified in accordance with the permissible error. Localization requirements are emphasized in

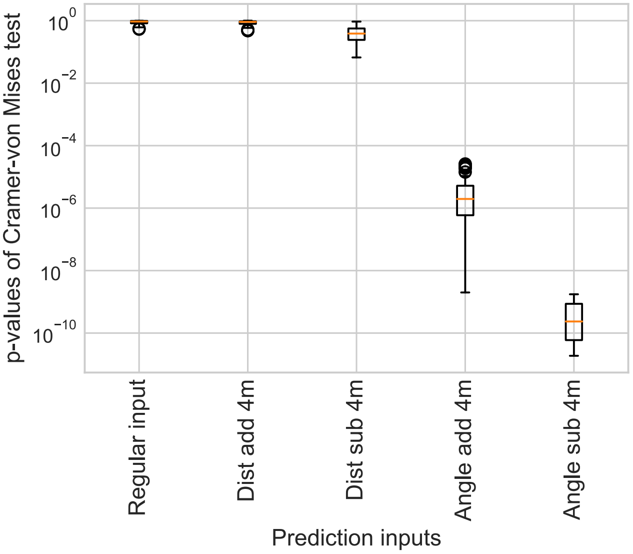


Fig. 4. Boxplot of log-scaled p-values for comparing prediction error distributions of inputs with errors in different directions.

this work because the manipulation of locations is comparatively simple. In this case, a change in prediction performance indicates that the perturbation is invalid. Since the perturbation is chosen according to a localization requirement, the corresponding requirement is also invalid. While results are not shown for brevity, the validation method is verified by analyzing the impact of halving and doubling the distance. In these cases, the validation method successfully identifies the verification inputs as invalid.

The remainder of this section separately validates different aspects proposed in this work. First, the distinction of radial distance and azimuth is evaluated. This is followed by considering the direction and the magnitude of the errors. The discussion of the validation is postponed for the later discussion following the results.

A. Types of Error

The first question is what influence the type of the location error exhibits. For this purpose, the effect of perturbing object distances and azimuth angles are compared. The comparison shows results for the largest nuScenes distance threshold of 4 m in Fig. 4. This value is either added or subtracted from the distance or it is added to increase or decrease the azimuth angle. It is observed that the results for the p-value clearly differ for distance errors and angular errors. This indicates that the effect of same error magnitude differs depending on the type of error.

B. Direction of Error

The next question is if positive or negative errors exhibit different influences on the results. For this purpose, we again compare the results in Fig. 4. The sign makes a difference both for the distance and the azimuth angle. Therefore, these results indicate that different error directions require separate consideration. This observation occurs despite the fact that the evaluation metric average distance error (ADE) does not consider the direction of the error. Overestimating the distance shows similar p-values as the regular input. Therefore, it appears that conservative estimates as proposed in this work are not supported.

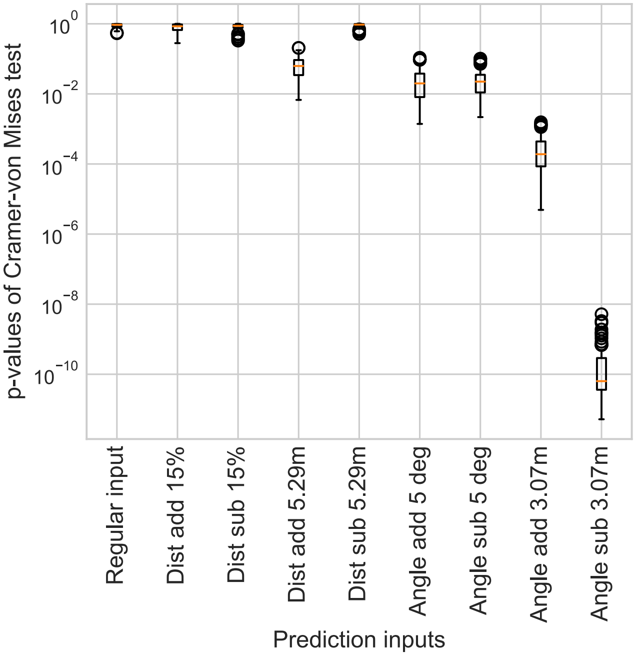


Fig. 5. Boxplot of log-scaled p-values for comparing prediction error distributions of inputs with different error scalings.

C. Error Scale and Magnitude

Another question is whether fixed errors or errors scaled by the distance to the ego are more appropriate. For this purpose, both variants are compared with thresholds constructed to yield an equal average error. For the distance error of 15%, the average distance error is 5.29 m. For the azimuth error of 5°, the average location error is 3.07 m. Results of the comparison are depicted in Fig. 5. For the distance errors, similar p-values to the regular input are observed. An exception is subtracting a constant distance, which shows lower values. This indicates that the proposed criteria are supported, while subtracting constant distance errors is invalid. All angular errors show lower p-values than for regular inputs, indicating they are not valid. However, the constant location errors in angular direction show p-values orders of magnitude lower than the angular errors which effectively scale with distance. This indicates that distance scaled angular errors affect the prediction component less. While the angular error threshold is invalid, the results therefore favor angular errors which scale with distance.

V. Results

The previous sections define and validate different requirements for different aspects of the detection task. Quantitative pass/fail criteria are developed for each aspect such as tracking, association, localization and velocity. In this section, the

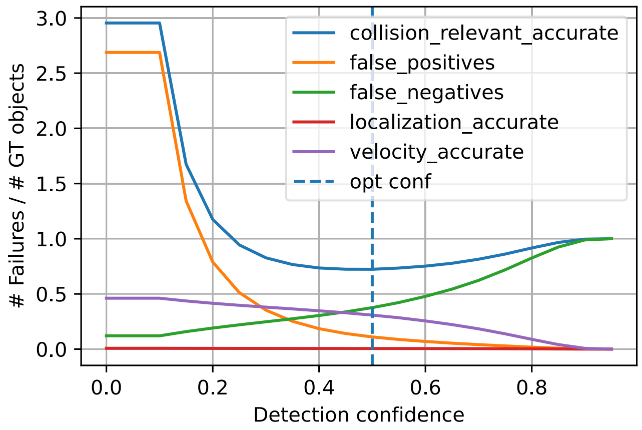


Fig. 6. Frequency and optimum of failures of the CenterPoint detector normalized with the number of GT objects. Unlike metrics based on recall and precision, lower is better.

proposed detection requirements are applied to a set of baseline detectors to evaluate their performance on an exemplary dataset. The evaluation is performed by counting the frequency of failures. While the number of failures is normalized with the number of GT objects for interpretability, no weighting of different failures is performed.

A. Implementation

Since labeled data is required to evaluate object detection, the validation split of the nuScenes dataset [12] is selected. For nuScenes, different detection baselines are readily available. All baselines are applied as implemented by the mmDetection3D [61] framework using pre-trained weights and default settings. Exemplary results for the popular lidar based detectors PointPillars [48] and CenterPoint [112] as well as for the camera based detector FCOS3D [98] are shown. Class-specific distance based filtering as in the nuScenes detection evaluation [64] is applied to objects prior to evaluation.

B. Prevalence of Failure Types

While this section is presented only for the CenterPoint [112] detection baseline, the general trends discussed here are similar for different detectors.

The first noteworthy result is that conservativity is only given for approximately half of the objects. To avoid skewing results, the following evaluation results therefore focus exclusively on accuracy. Fig. 6 shows different types of failures for the detector. Even for the optimal confidence threshold, an average of 0.72 collision relevant failures per GT box is observed. It is observed that the collision relevant failures are dominated by the association failures with 0.49 failures per GT box. Velocity and localization failures show 0.31 and 0.01 failures per GT box, respectively.

C. Different Detectors

While the previous section focused on general results that are similar for different detectors, this section emphasizes differences between detectors. For this purpose, two lidar and one camera baseline are compared in Fig. 7. Only matching and velocity failures are evaluated since they are shown to be the dominant failure types.

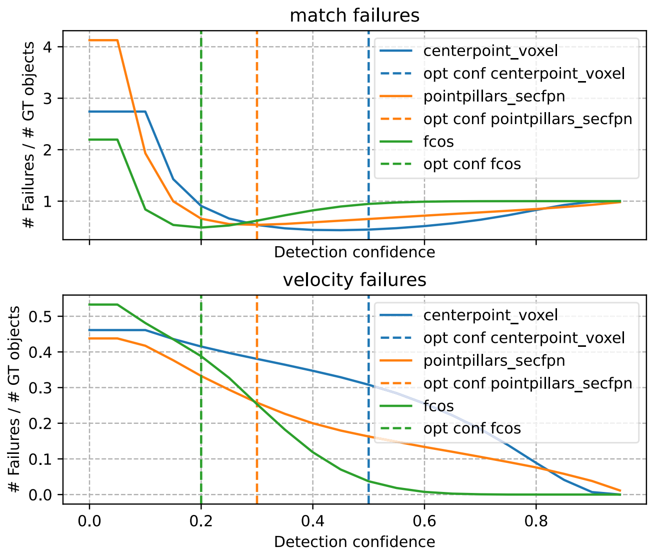


Fig. 7. Frequency of and optimum of matching and velocity failures for different detectors normalized with the number of GT objects.

Different detectors display different optimal confidence thresholds. While the qualitative distribution is similar, the exact number of failures and their distribution over confidence scores differ depending on the detector.

D. Ideal Fusion and Uncorrelated Fusion

One common assumption is that fusing multiple detectors and sensor modalities improves accuracy. To test this hypothesis, fusion is evaluated regarding the matching and localization failures as in the previous section.

To estimate the upper bounds of performance achievable with fusion, an ideal detector is constructed. False negatives are only counted if the object is missing in both modalities, while false positives are only counted if they are present in both modalities. Since this fusion procedure assumes GT knowledge, it overestimates practically achievable fusion performance. Another reference is obtained by calculating failure rates assuming that two modalities or detectors have no correlation. In this uncorrelated case, the failure likelihood is obtained by simply multiplying the failure likelihoods of the two detectors.

A comparison of ideal and uncorrelated fusion along with the baseline results is presented in Fig. 8. The ideal fusion improves the baselines especially regarding matching. However, it fails to achieve large gains for matching or velocity if three detectors are fused. In these cases, the failure rates remain substantially higher than for the case where no correlation of failures is present.

VI. Discussion

In this section, a discussion of the perception criteria proposed in this work is presented. Finally, the detection and fusion performance as well as their implications are discussed.

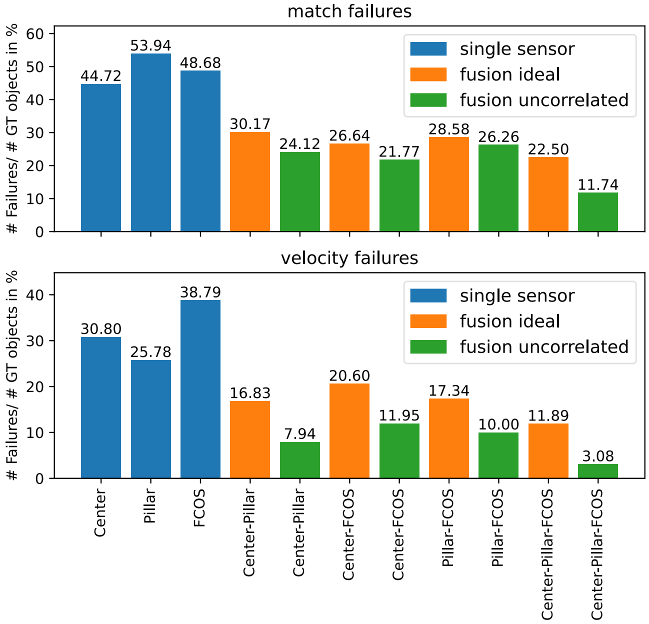


Fig. 8. Frequency of matching and velocity failures normalized the number of GT objects. Optimum thresholds of three detectors are compared with an ideal fusion and a hypothetical uncorrelated fusion.

A. *Comparison With Existing Metrics*

Contrary to the common tracking task, the unique identification of objects is not supported by human perception literature. This is in agreement with the typical 3D perception pipeline which primarily considers detection in the track by detection paradigm [66].

The temporal requirements are comparatively strict, demanding detection within 150 ms. This indicates that annotation frequencies such as in the nuScenes dataset with 500 ms between frames [12] are insufficient. Additionally, the typical lidar and radar scan rates are approximately 100 ms [29], [87]. Common strategies as aggregating information from multiple time frames for detection [16], [47], [99] or tracking [19], [66], [100] may not be applicable for sudden object appearances. This may strengthen the case for using sensors with faster frame rates.

Matching and localization are distinguished for interpretability and consistency with existing pipelines. However, the mistakes are considered equally critical, thus diminishing the importance of the distinction. It is shown that the majority of failures originates from matching failures. This indicates that even if no fine-grained localization is required, determining the existence of objects remains challenging. Perception results may benefit from explicit labeling of groups as in the Microsoft COCO dataset [52] since matching errors constitute a common failure.

Classification is less relevant compared to common perception metrics such as mAP, which consider it a prerequisite for matching. It should be noted that classification may be required by the downstream task depending on the implementation. However, we show that even when neglecting classification requirements, perception errors remain common. While this question requires further attention, a possible reason is the fact that current detectors are designed for standard evaluation pipelines.

The requirements of this work differ from common metrics, since they do not provide a weighting of different error types. Since no averaging between different attributes is performed, the requirements are more sensitive to individual attribute errors. Velocity failures are present for approximately half of the GT boxes for which collision relevant failures occur. This indicates that existing metrics may insufficiently consider attributes other than localization. At the same time, the proposed distance errors are more lenient than for common detection metrics.

B. *Perception Criteria*

This work successfully defines pass/fail criteria for different aspects of the object detection task. The criteria are firmly grounded in human performance and allow fair comparison regardless of the sensor modality. Each failure is interpretable and can be treated independently from other failure types.

The validation procedure from SURE-Val [85] is applied to the perception criteria. Results show that for instance the largest distance threshold of 4 m as used by nuScenes [12] is invalid. In addition, the validation falsifies the common assumption that different directions of location errors should be treated equally. Furthermore, fixed error thresholds obtain lower p-values than errors which scale with distance for equal error budget. This indicates that fixed error thresholds for localization are invalid, favouring scaling with the distance. The validation also justifies treating positive and negative errors differently. The distance error threshold of 15% proposed in this work is supported by the validation, since the p-values are in a similar range as for unmodified inputs. While the nuScenes thresholds are valid when applied to distance, the requirements proposed in this work are favoured since they are more permissive. However, the angular threshold of 5° is falsified by the validation. This indicates that requirements for the azimuth angle are more restrictive than the distance requirements. One potential reason is that humans do not utilize visual angles but instead rely upon distance estimation to the lane or the planned trajectory. While the validation confirms distinguishing positive and negative errors, conservativity if not confirmed. Since the validation metric uses a symmetric distance metric, safety outcomes are not considered. In addition, the dataset may not include critical driving situation where any of the participants is close to an accident.

Overall, the validation results show the deficiency of current perception metrics and largely support the results of this work. Importantly, the approach of obtaining interpretable analytic criteria from human perception is shown to be feasible. However, the exact threshold for the angular error and the conservative estimates are not confirmed.

C. *Validation Method*

The validation procedure is verified and successfully applied to the perception criteria proposed in this work. It is shown that reconciling the interpretability and simplicity of analytic criteria with the context-awareness of neural networks is possible.

The validation results largely support the assumptions of this work while falsifying the assumptions of common dataset metrics. It is shown that the type and the direction of a location error influence the prediction result. Furthermore, errors which scale with the distance of an object are favored. While the distance threshold is supported, this is not the case for the angular threshold. Contrary to the assumptions of this work, both conservative and non-conservative errors are supported.

Nevertheless, there are limitations to its application. Firstly, the validation procedure is currently only applicable to location errors. For velocity errors, it is currently unclear how to appropriately modify the prediction input to consistently manipulate velocity and location over time. Other attributes such as size or orientation are not used as input by the implemented prediction network. Accordingly, their errors cannot influence the prediction result. In addition, the current symmetric distance metric used for validation does not consider safety outcomes. However, evaluating safety outcomes may be required to adequately consider the effects of conservative estimates.

D. Detection Performance

Overall, the analyzed detection architectures show high numbers of failures with FCOS3D at 0.82, PointPillars at 0.76 and CenterPoint at 0.72 collision relevant failures per GT box on average. Note that this work analyzes the failure likelihood, meaning that lower scores are better with the ideal score being zero. The prevalence of matching failures occurs despite the fact that the location thresholds are more lenient than for standard AP. In addition, half of the estimates are not conservative for each attribute, which may pose safety risks. This means current perception algorithms frequently fail to meet requirements, providing clear evidence that safety is insufficiently considered. Future algorithm evaluation and development should explicitly consider safety requirements.

Performances on mAP are 32% for FCOS3D, 34% for PointPillars and 56% for CenterPoint [61]. While the ranking remains the same when using the metrics in this work, the discrepancy between methods is different. The performance gap between PointPillars and CenterPoint is smaller in this work than based on mAP when assuming linear scaling. This may indicate that the discrepancy on standard metrics results of the capability of CenterPoint for fine-grained localization. Conversely, the wider gap between PointPillars and FCOS3D in this work may be attributed to the fact that classification is not considered. The superior ability of camera images to capture semantic information [26], [54] may improve AP without significantly affecting the results in this work.

Another aspect visible from the results is that the optimum occurs at different confidence scores for each detector. This in agreement with previous observations in literature that confidence scores are insufficiently calibrated [42], [102]. However, the defined evaluation procedure does not require confidence scores. Developing detectors which leverage this fact as well as other specifics of the evaluation procedure is left for future work.

E. Fusion Performance

The investigation reveals that an ideal fusion with access to GT information for the fusion procedure shows large potential for improving the baselines. However, even this ideal fusion has higher failure likelihoods than the uncorrelated case at failure likelihoods of approximately 10% and below. In this case, common causes [84] such as occlusion or small size may cause correlation among errors, regardless of the modality or architecture. Additionally, the correlation may be even stronger under challenging conditions such as adverse weather. Overall, the naive assumption of non-existent correlation between sensor modalities is likely insufficient. This also agrees with prior work showing that different architectures and modalities exhibit strong correlation in a more theoretical setting [30].

It also appears that the ideal fusion performance does not differ if two detectors are combined for two different modalities or for the same modality. However, only limited conclusions can be drawn from the limited number of detectors studied. Further investigation into reduction of correlations as well as including other sensor types such as FMCW lidar, high-resolution radar or thermal cameras is warranted.

F. Transfer of Results

In this section, the transfer of the methods and results obtained in this work to other tasks and domains is discussed.

The first question is whether the methodology is also applicable to other perceptual tasks such as semantic segmentation. Generally, the proposed methodology is limited to tasks for which a human perceptual equivalent exists. Human perception was shown to be object-centric in this work. Further study may be required to identify other representations in human perception. However, since it is unknown if any equivalent to semantic segmentation exists, the method of this work may be inapplicable. The validation method is however applicable if any such requirements are identified in the future. Applying the validation method in this case requires a motion prediction network which ingests semantic segmentation.

Another question is the transfer of the results to other domains and datasets. Regarding the perception requirements, no assumptions regarding domain or sensor setup are incorporated. We therefore believe that the requirements apply generally for 3D object detection. The validity of the requirements is at present only confirmed on the nuScenes dataset. However, the validation only utilizes object list data of surrounding traffic participants. Therefore, characteristics of the dataset are only weakly expressed. A transfer of the requirements and their validity to other datasets thus seems plausible.

For the detection performance, the transfer is more difficult to ascertain. It should be noted that object detection performance always differs between different datasets. Reasons include the differences in sensor setup as well as potentially different difficulties of the perception task. Nevertheless, CenterPoint reports an AP of 58% on nuScenes and 72% on Waymo [112]. We consider it unlikely that all of the failures observed in this work are accounted for by the specifics of the dataset. Therefore, we believe that the general results of high failure rates and correlation of sensors transfer to other datasets.

However, while a transfer may be plausible, explicit testing on other datasets is required. While further research in this direction is required, this is left for future work.

VII. Conclusion and Outlook

Within this work, new perception metrics as well as thresholds to define interpretable pass/fail criteria are developed. Human performance as well as conservative estimates are considered to develop interpretable analytic perception requirements linked to safety. A validation method based on a motion prediction network is introduced and applied. Results show lack of validity of current metrics while supporting the propositions put forth in this work.

The assessment of existing detectors shows that the requirements are not met for the majority of objects in a contemporary dataset. Despite localization criteria which are more lenient than common AP matching thresholds, matching failures dominate the results. The differences between different detectors as well as different modalities are moderate when using the metrics developed in this work. This indicates that fine-grained localization may be over-emphasized in current metrics. This shows the need for explicit consideration of safety in evaluating and developing perception algorithms.

Further investigation effort is required to understand the effects of dataset, modality, architecture and optimization goal on the failure rates. Regarding fusion of different detectors, preliminary results indicate substantial correlation between detectors and modalities. This indicates that multi-modal fusion may be insufficient to alleviate the failures that occur in contemporary object detection pipelines and requires further research.

Finally, the authors hope that these requirements can serve as basis for future algorithm evaluation and development.

References

[1] G. A. Alvarez and S. L. Franconeri, "How many objects can you track? Evidence for a resource-limited attentive tracking mechanism," *J. Vis.*, vol. 7, no. 13, pp. 14.1–14.10, 2007.

[2] C. T. Amersbach, "Functional decomposition approach - Reducing the safety validation effort for highly automated driving," Doctoral dissertation, Technische Universität Darmstadt, Darmstadt, 2020.

[3] T. W. Anderson, "On the distribution of the two-sample Cramer-von Mises criterion," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1148–1159, 1962.

[4] R. Ashmore, R. Calinescu, and C. Paterson, "Assuring the machine learning lifecycle: Desiderata, methods, and challenges," *ACM Comput. Surveys*, vol. 54, no. 5, pp. 1–39, 2022.

[5] A. Bansal et al., "Verifiable obstacle detection," in *Proc. IEEE 33rd Int. Symp. Softw. Rel. Eng.*, 2022, pp. 61–72.

[6] A. Bansal, J. Singh, M. Verucchi, M. Caccamo, and L. R. Sha, "Risk ranked recall: Collision safety metric for object detection systems in autonomous vehicles," in *Proc. 10th Mediterranean Conf. Embedded Comput.*, 2021, pp. 1–4.

[7] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, 2008.

[8] E. Brenner and J. B. J. Smeets, "Continuous visual control of interception," *Hum. Movement Sci.*, vol. 30, no. 3, pp. 475–494, 2011.

[9] E. Brenner and J. B. J. Smeets, "How people achieve their amazing temporal precision in interception," *J. Vis.*, vol. 15, no. 3, pp. 1–21, 2015.

[10] B. Bruyn and G. A. Orban, "Human velocity and direction discrimination measured with random dot patterns," *Vis. Res.*, vol. 28, no. 12, pp. 1323–1335, 1988.

[11] P. F. Bulakowski, K. Koldewyn, and D. Whitney, "Independent coding of object motion and position revealed by distinct contingent aftereffects," *Vis. Res.*, vol. 47, no. 6, pp. 810–817, 2007.

[12] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11618–11628.

[13] V. Cavallo and M. Laurent, "Visual information and skill level in time-to-collision estimation," *Perception*, vol. 17, no. 5, pp. 623–632, 1988.

[14] R. Chan et al., "Segmentmeifyoucan: A benchmark for anomaly segmentation," in *Proc. 35th Conf. Neural Inf. Process. Syst. (NeurIPS 2021) Track Datasets Benchmarks*, 2021, pp. 1–13.

[15] M. Chang et al., "Argoverse: 3D tracking and forecasting with rich maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8740–8749.

[16] Q. Chen, L. Sun, E. Cheung, and A. Yuille, "Every view counts: Cross-view consistency in 3D object detection with hybrid-cylindrical-spherical voxelization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–12.

[17] Q. Chen, L. Sun, Z. Wang, K. Jia, and A. Yuille, "Object as hotspots: An anchor-free 3D object detection approach via firing of hotspots," in *Proc. Eur. Conf. Comput. Vis.*, Cham, 2020, pp. 68–84.

[18] Y. Chen, H. E. Bedell, and L. J. Frishman, "The precision of velocity discrimination across spatial frequency," *Percep. Psychophys.*, vol. 60, no. 8, pp. 1329–1336, 1998.

[19] H. Chiu, J. Li, A. Prioletti, and J. Bohg, "Probabilistic 3D multi-object tracking for autonomous driving," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 14227–14233.

[20] *COCO Consortium*, "COCO common objects in context: Detection evaluation," 2015. [Online]. Available: https://cocodataset.org/#detection-eval

[21] A. D'Amour et al., "Underspecification presents challenges for credibility in modern machine learning," *J. Mach. Learn. Res.*, vol. 23, no. 226, pp. 1–61, 2022.

[22] O. S. Daum and H. Hecht, "Distance estimation in vista space," in *Attention Percep. Psychophys.*, vol. 71, no. 5, pp. 1127–1137, 2009.

[23] B. Deng, C. Qi, M. Najibi, T. A. Funkhouser, Y. Zhou, and D. Anguelov, "Revisiting 3D object detection from an egocentric perspective," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 26066–26079.

[24] *Ethics Comission*, "Automated and connected driving," 2017. [Online]. Available: https://bmdv.bund.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile

[25] M. Everingham, L. v. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[26] J. Fei, W. Chen, P. Heidenreich, S. Wirges, and C. Stiller, "SemanticVoxels: Sequential fusion for 3D pedestrian detection using LiDAR point cloud and semantic segmentation," in *Proc. IEEE Int. Conf. Multisensor Fusion Integration Intell. Syst.*, 2020, pp. 185–190.

[27] D. Feng et al., "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, Mar. 2021.

[28] R. Ge et al., "AFDet: Anchor free one stage 3D object detection," 2020, *arXiv:2006.12671*.

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

[30] H. Gottschalk, M. Rottmann, and M. Saltagic, "Does redundancy in AI perception systems help to test for super-human automated driving performance?," in *Deep Neural Networks and Data for Automated Driving: Robustness, Uncertainty Quantification, and Insights Towards Safety*, Cham, Switzerland:Springer International Publishing, 2022, pp. 81–106.

[31] Y. Guo, H. Caesar, O. Beijbom, J. Philion, and S. Fidler, "The efficacy of neural planning metrics: A meta-analysis of PKL on nuscenes," 2020, *arXiv:2010.09350*.

[32] T. Haarmeier and P. Thier, "Detection of speed changes during pursuit eye movements," *Exp. Brain Res.*, vol. 170, no. 3, pp. 345–357, 2006.

[33] F. Henze, D. Fabender, and C. Stiller, "Identifying admissible uncertainty bounds for the input of planning algorithms," *IEEE Trans. Intell. Veh.*, vol. 8, no. 4, pp. 3129–3143, Apr. 2023.

[34] A. Higashiyama, "Anisotropic perception of visual angle: Implications for the horizontal-vertical illusion, overconstancy of size, and the moon illusion," *Percep. Psychophys.*, vol. 51, no. 3, pp. 218–230, 1992.

[35] E. R. Hoffmann and R. G. Mortimer, "Drivers' estimates of time to collision," *Accident Anal. Prevention*, vol. 26, no. 4, pp. 511–520, 1994.

[36] M. Hoss, M. Scholtes, and L. Eckstein, "A review of testing object-based environment perception for safe automated driving," *Automot. Innov.*, vol. 5, no. 3, pp. 223–250, 2022.

[37] Technical Committee ISO/TC 22, Road Vehicles, Subcommittee SC 32, Electrical and Electronic Components and General System Aspects, *Road Vehicles - Safety of the Intended Functionality*, ISO/PAS 21448, International Organization for Standardization, Geneva, Switzerland, Jan. 2019.

[38] R. Isermann, *Fault-Diagnosis Systems: An Introduction From Fault Detection to Fault Tolerance*. Berlin Heidelberg:Springer-Verlag, 2006.

[39] A. Jaiswal, Y. Wu, P. Natarajan, and P. Natarajan, "Class-agnostic object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 918–927.

[40] P. Junietz, W. Wachenfeld, K. Klonecki, and H. Winner, "Evaluation of different approaches to address safety validation of automated driving," in *Proc. 21st Int. Conf. Intell. Transp. Syst.*, 2018, pp. 491–496.

[41] Y. Kang, H. Yin, and C. Berger, "Test your self-driving algorithm: An overview of publicly available driving datasets and virtual testing environments," *IEEE Trans. Intell. Veh.*, vol. 4, no. 2, pp. 171–185, Jun. 2019.

[42] Y. Kato and S. Kato, "A conditional confidence calibration method for 3D point cloud object detection," in *Proc. IEEE Intell. Veh. Symp.*, 2022, pp. 1835–1844.

[43] R. J. Kiefer, C. A. Flannagan, and C. J. Jerome, "Time-to-collision judgments under realistic driving conditions," *Hum. Factors*, vol. 48, no. 2, pp. 334–345, 2006.

[44] K.-Y. Kim, Y. Kim, J. Park, and Y.-S. Kim, "Real-time performance evaluation metrics for object detection and tracking of intelligent video surveillance systems," *Asia Pacific J. Contemporary Educ. Commun. Technol.*, vol. 2, pp. 173–179, 2016.

[45] M. Koivisto and A. Revonsuo, "An ERP study of change detection, change blindness, and visual awareness," *Psychophysiology*, vol. 40, no. 3, pp. 423–429, 2003.

[46] M. Kuba, Z. Kubová, J. Kremlácek, and J. Langrová, "Motion-onset VEPs: Characteristics, methods, and diagnostic use," *Vis. Res.*, vol. 47, no. 2, pp. 189–202, 2007.

[47] A. Laddha, S. Gautam, G. P. Meyer, and C. Vallespi-Gonzalez, "RV-FuseNet: Range view based fusion of time-series LiDAR data for joint 3D object detection and motion forecasting," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2021.

[48] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 12697–12705.

[49] C. A. Levin and R. N. Haber, "Visual angle as a determinant of perceived interobject distance," *Percep. Psychophys.*, vol. 54, no. 2, pp. 250–259, 1993.

[50] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2953–2960.

[51] Z. Li and F. H. Durgin, "Perceived azimuth direction is exaggerated: Converging evidence from explicit and implicit measures," *J. Vis.*, vol. 16, no. 1, pp. 1–19, 2016.

[52] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[53] L. Liu et al., "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020.

[54] Z. Liu et al., "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," 2022, *arXiv:2205.13542*.

[55] S. J. Luck and E. K. Vogel, "The capacity of visual working memory for features and conjunctions," *Nature*, vol. 390, no. 6657, pp. 279–281, 1997.

[56] J. Luiten et al., "HOTA: A higher order metric for evaluating multi-object tracking," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 548–578, 2021.

[57] W. Luo et al., "Multiple object tracking: A literature review," *Artif. Intell.*, vol. 293, no. 103448, pp. 1–21, 2021.

[58] M. Lyssenko, C. Gladisch, C. Heinzemann, M. Woehrle, and R. Triebel, "Towards safety-aware pedestrian detection in autonomous systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 293–300.

[59] J. Mao et al., "One million scenes for autonomous driving: Once dataset," in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, 2021, pp. 1–13.

[60] G. Markkula, J. Engström, J. Lodin, J. Bärgman, and T. Victor, "A farewell to brake reaction times? Kinematics-dependent brake response in naturalistic rear-end emergencies," in *Accident Anal. Prevention*, vol. 95, pp. 209–226, 2016.

[61] *MMDetection3D Contributors*, "Mmdetection3D: Openmmlab next-generation platform for general 3D object detection," 2020. [Online]. Available: https://github.com/open-mmlab/mmdetection3d

[62] B. Moeller, H. Zoppke, and C. Frings, "What a car does to your perception: Distance evaluations differ from within and outside of a car," *Psychon. Bull. Rev.*, vol. 23, no. 3, pp. 781–788, 2016.

[63] I. Murakami, "Correlations between fixation stability and visual motion sensitivity," *Vis. Res.*, vol. 44, no. 8, pp. 751–761, 2004.

[64] *nuScenes*, "nuScenes detection task: Leaderboard," 2020. [Online]. Available: https://www.nuscenes.org/object-detection?externalData=all&mapData=all&modalities=Any

[65] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, "One metric to measure them all: Localisation recall precision (LRP) for evaluating visual detection tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9446–9463, Dec. 2022.

[66] Z. Pang, Z. Li, and N. Wang, "SimpleTrack: Understanding and rethinking 3D multi-object tracking," in *Proc. Comput. Vis. - ECCV 2022 Workshops*, 2023, vol. 13801, pp. 680–696.

[67] PEGASUS Project, "Pegasus method: An overview," 2019. [Onl ine]. Available: https://www.pegasusprojekt.de/files/tmpl/Pegasus-Abschlussveranstaltung/PEGASUS-Gesamtmethode.pdf

[68] Q.-H. Pham et al., "A 3D dataset: Towards autonomous driving in challenging environments," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 2267–2273.

[69] J. Philion, A. Kar, and S. Fidler, "Learning to evaluate perception models using planner-centric metrics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14052–14061.

[70] R. Philipp, H. Qian, L. Hartjen, F. Schuldt, and F. Howar, "Simulation-based elicitation of accuracy requirements for the environmental perception of autonomous vehicles," in *Proc. Int. Symp. Leveraging Appl. Formal Methods, Verification Validation*, 2021, pp. 129–145.

[71] R. Philipp et al., "Systematization of relevant road users for the evaluation of autonomous vehicle perception," in *Proc. IEEE Int. Syst. Conf.*, 2022, pp. 1–8.

[72] Z. W. Pylyshyn and R. W. Storm, "Tracking multiple independent targets: Evidence for a parallel tracking mechanism," *Spatial Vis.*, vol. 3, no. 3, pp. 179–197, 1988.

[73] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3D object detection in point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9276–9285.

[74] C. R. Qi et al., "Offboard 3D object detection from point cloud sequences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2021, pp. 6130–6140.

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

[76] V. Schönemann, M. Duschek, and H. Winner, "Maneuver-based adaptive safety zone for infrastructure-supported automated valet parking," in *Proc. 5th Int. Conf. Veh. Technol. Intell. Transport Syst.*, 2019, pp. 343–351.

[77] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "On a formal model of safe and scalable self-driving cars," 2017, *arXiv:1708.06374*.

[78] D. J. Simons, "Current approaches to change blindness," *Vis. Cogn.*, vol. 7, no. 1/3, pp. 1–15, 2000.

[79] D. J. Simons and D. T. Levin, "Failure to detect changes to people during a real-world interaction," *Psychon. Bull. Rev.*, vol. 5, no. 4, pp. 644–649, 1998.

[80] B. Singh, H. Li, A. Sharma, and L. S. Davis, "R-FCN-3000 at 30FPS: Decoupling detection and classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1081–1090.

[81] R. J. Snowden and E. Kavanagh, "Motion perception in the ageing visual system: Minimum motion, motion coherence, and speed discrimination thresholds," *Perception*, vol. 35, no. 1, pp. 9–24, 2006.

[82] B. Song et al., "A stochastic graph evolution framework for robust multi-target tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 605–619.

[83] S. Sontges, M. Koschi, and M. Althoff, "Worst-case analysis of the time-to-react using reachable sets," in *Proc. IEEE Intell. Veh. Symp.*, 2018, pp. 1891–1897.

[84] R. F. Stapelberg, *Handbook of Reliability, Availability, Maintainability and Safety in Engineering Design*. London, U.K.:Springer, 2009.

[85] K. Storms, K. Mori, and S. Peters, "Sure-val: Safe urban relevance extension and validation," 2023, *arXiv:2308.02266*.

[86] M. Strauss and J. Carnahan, "Distance estimation error in a roadway setting," *Police J.*, vol. 82, no. 3, pp. 247–264, 2009.

[87] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2443–2451.

[88] T. Takeuchi and K. K. Valois, "Velocity discrimination in scotopic vision," *Vis. Res.*, vol. 40, no. 15, pp. 2011–2024, 2000.

[89] E. Thorn, S. Kimmel, and M. Chaka, “A framework for automated driving system testable cases and scenarios,” 2018. [Online]. Available: https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/13882-automateddrivingsystems_092618_v1a_tag.pdf

[90] S. Thorpe, D. Fize, and C. Marlot, “Speed of processing in the human visual system,” *Nature*, vol. 381, no. 6582, pp. 520–522, 1996.

[91] S. Topan et al., “Interaction-dynamics-aware perception zones for obstacle detection safety evaluation,” in *Proc. IEEE Intell. Veh. Symp.*, 2022, pp. 1201–1210.

[92] A. Treisman and H. Schmidt, “Illusory conjunctions in the perception of objects,” *Cogn. Psychol.*, vol. 14, no. 1, pp. 107–141, 1982.

[93] R. VanRullen and S. J. Thorpe, “The time course of visual processing: From early perception to decision-making,” *J. Cogn. Neurosci.*, vol. 13, no. 4, pp. 454–461, 2001.

[94] A. K. Verma, A. Srividya, and D. R. Karanki, *Reliability and Safety Engineering, Volume 0 of Springer Series in Reliability Engineering*. London, U.K.:Springer-Verlag, 2010.

[95] G. Volk, J. Gamerdinger, A. V. Betnuth, and O. Bringmann, “A comprehensive safety metric to evaluate perception in autonomous systems,” in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst.*, 2020, pp. 1–8.

[96] E. Vul, C. A. Rieth, T. F. Lew, and A. N. Rich, “The structure of illusory conjunctions reveals hierarchical binding of multipart objects,” in *Attention, Percep. Psychophys.*, vol. 82, no. 2, pp. 550–563, 2020.

[97] W. Wachenfeld and H. Winner, “The release of autonomous vehicles,” in *Autonomous Driving*. Berlin, Heidelberg:Springer, 2016, pp. 425–449.

[98] T. Wang, X. Zhu, J. Pang, and D. Lin, “FCOS3D: Fully convolutional one-stage monocular 3D object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 913–922.

[99] Y. Wang et al., “Train in Germany, test in the USA: Making 3D object detectors generalize,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11710–11720.

[100] X. Weng, J. Wang, D. Held, and K. Kitani, “3D multi-object tracking: A baseline and new evaluation metrics,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 10359–10366.

[101] L. Westhofen et al., “Criticality metrics for automated driving: A review and suitability analysis of the state of the art,” *Arch. Comput. Methods Eng.*, vol. 30, no. 1, pp. 1–35, 2023.

[102] O. Willers, S. Sudholt, S. Raafatnia, and S. Abrecht, “Safety concerns and mitigation approaches regarding the use of deep learning in safety-critical perception tasks,” in *Proc. Int. Conf. Comput. Safety Rel. Secur.*, 2020, pp. 336–350.

[103] M. Wolf, L. R. Douat, and M. Erz, “Safety-aware metric for people detection,” in *Proc. IEEE Int. Intell. Transp. Syst. Conf.*, 2021, pp. 2759–2765.

[104] J. M. Wolfe and K. R. Cave, “The psychophysical evidence for a binding problem in human vision,” *Neuron*, vol. 24, no. 1, pp. 11–17, 1999.

[105] X. Wu, D. Sahoo, and S. C. H. Hoi, “Recent advances in deep learning for object detection,” *Neurocomputing*, vol. 396, pp. 39–64, 2020.

[106] Y. Wu, S. Ying, and L. Zheng, “Size-to-depth: A new perspective for single image depth estimation,” 2018, *arXiv:1801.04461*.

[107] C. Xu, J. Wang, W. Yang, and L. Yu, “Dot distance for tiny object detection in aerial images,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2021, pp. 1192–1201.

[108] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, “Who are you with and where are you going?,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1345–1352.

[109] J.-J. Yan, B. Lorv, H. Li, and H.-J. Sun, “Visual processing of the impending collision of a looming object: Time to collision revisited,” *J. Vis.*, vol. 11, no. 12, pp. 1–25, 2011.

[110] B. Yang, M. Bai, M. Liang, W. Zeng, and R. Urtasun, “Auto4D: Learning to label 4D objects from sequential point clouds,” 2021, *arXiv:2101.06586*.

[111] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, “IPOD: Intensive point-based object detector for point cloud,” 2018, *arXiv:1812.05276*.

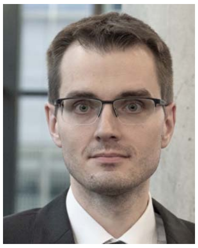
[112] T. Yin, X. Zhou, and P. Krähenbühl, “Center-based 3D object detection and tracking,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11779–11788.

[113] H. Zhao, S. K. S. Hari, T. Tsai, M. B. Sullivan, S. W. Keckler, and J. Zhao, “Suraksha: A quantitative AV safety evaluation framework to analyze safety implications of perception design choices,” in *Proc. IEEE/IFIP 51st Annu. Int. Conf. Dependable Syst. Netw.*, 2021, pp. 35–38.

[114] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, “Distance-IoU loss: Faster and better learning for bounding box regression,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12993–13000.

[115] X. Zhou, V. Koltun, and P. Krähenbühl, “Probabilistic two-stage detection,” 2021, *arXiv:2103.07461v1*.

Ken T. Mori received the B.S. and M.S. degrees in mechanical and process engineering from the Technical University of Darmstadt, Darmstadt, Germany in 2020. He has been a Research Assistant with the Institute of Automotive Engineering, Technical University of Darmstadt, since 2020.



Steven Peters was born in 1987. He received the Ph.D. (Dr.-Ing.) degree from the Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2013. From 2016 to 2022, he was a Manager of artificial intelligence research with Mercedes-Benz AG in Germany. He is cuurently a Full Professor with the Technical University of Darmstadt, Darmstadt, Germany and has been the Head of the Institute of Automotive Engineering, Department of Mechanical Engineering, since 2022.