SURE-Val: **S**afe **U**rban **R**elevance **E**xtension and **Val**idation

Kai Storms*†,Ken Mori*†und Steven Peters†

Abstract: To evaluate perception components of an automated driving system, it is necessary to define the relevant objects. While the urban domain is popular among perception datasets, relevance is insufficiently specified for this domain. Therefore, this work adopts an existing method to define relevance in the highway domain and expands it to the urban domain. While different conceptualizations and definitions of relevance are present in literature, there is a lack of methods to validate these definitions. Therefore, this work presents a novel relevance validation method leveraging a motion prediction component. The validation leverages the idea that removing irrelevant objects should not influence a prediction component which reflects human driving behavior. The influence on the prediction is quantified by considering the statistical distribution of prediction performance across a large-scale dataset. The validation procedure is verified using criteria specifically designed to exclude relevant objects. The validation method is successfully applied to the relevance criteria from this work, thus supporting their validity.

Keywords: Automated Driving, Perception, Relevance, Safety

1 Introduction

Automated driving (AD) is currently viewed as a key emerging technology. Among the benefits which are attributed to AD are increased comfort, availability of mobility and foremost an increase in safety for the traffic environment [11]. However, safety assurance of AD remains a challenge. Whilst first SAE level 3 AD system (AD) are already available [2] for parts of the highway domain, a current focus in research is the urban domain [13].

One method supporting safety assurance by offering potential benefits regarding the testing effort is modular decomposition [4]. Under this method, all modules of a classic Sense-Plan-Act architecture and its derivatives, such as the perception module, need to be evaluated individually [38]. For this individual evaluation, it is imperative to have knowledge about what is relevant to the subject module, both for completeness and efficiency.

Therefore, in this paper we will consider relevance for the perception module and how it can be evaluated. Current approaches face a variety of problems. Most methods lack generality because they are specific to either one module or a limited set of scenarios. These methods cannot be applied to other modules or scenarios without requiring major redesign efforts.

*contributed equally
†Institute of Automotive Engineering (FZD) at Technical University of Darmstadt, 64287 Darmstadt (e-mail: firstname.lastname@tu-darmstadt.de)

Furthermore, a lack of consistency between relevance results must be noted among different methods. One reason is the lack of validation in previous methods. This situation is exacerbated by the fact that the current state of the art does not provide any methodology to validate relevance criteria.

This paper will expand on the previous work of Mori & Storms [30], henceforth denoted as Stuctured Analysis for Conservative Relevance Estimation in Driving context (SACRED). Firstly, the Safe Urban Relevance Extension (SURE) is introduced. As main contribution, we present a novel methodology that enables the evaluation of validity for a given relevance selection method. This methodology is applied and verified using the expanded method for the urban domain.

2 Related Work

This section covers related works with respect to concepts of relevance as well as their respective validation.

2.1 Relevance

Previous conceptualizations of relevance can be broadly categorized in heuristic approaches, formal approaches and concepts based on a downstream task.

Heuristic approaches typically implicitly define relevance by excluding certain objects based on simple criteria. One application is the datasets used to develop and test perception functions. Here, thresholds on distance [42] or number of points from lidar and radar [9] are applied during annotation. Heuristics are also applicable when defining perception metrics [21]. Criteria such as distance [9], height in camera image plane and occlusion [18] are used by dataset metrics. Other metric proposals in literature follow similar approaches such as leveraging the distance to the ego vehicle [27] or the ego trajectory [8]. Similar ideas are found in neural path planners where relevance is implicitly defined by the network input. Commonly, inputs are restricted to a certain geometric region [6, 19, 34, 40] or a limited number of objects [1, 12, 17, 23, 44].

Formal approaches provide an explicit consideration of relevance based on requirements. Typically, relevance is related to safe behavior of the vehicles by considering reachability or formal planners [21]. Reachability leverages kinematic constraints to define potential collision objects as relevant [3, 43]. Other work leverages formal planners either from preexisting work [45] or by directly specifying context-dependent behavioral requirements [30, 37, 41]

In order to avoid manual specification of relevance, the planning Kullback-Leibler divergence (PKL) [35] and following work [20, 37] propose to leverage neural planners. Here, relevance is conceptualized and quantified as magnitude of the effect of an object on a downstream planner implementation [35]. However, the validity of this approach is limited to the specific implementation of the planning algorithm [36].

Overall, various approaches for the definition of relevance are proposed and reach different conclusions. Notably, there is currently no reconciliation of formal and downstream implementation based approaches.

2.2 Relevance Validation

While different approaches have been suggested to conceptualize object relevance, the validation of these concepts is currently underexplored. Perception datasets such as [18, 42, 47] do not consider the validity of the perception metrics with respect to safety. Other datasets discuss metrics with respect to the weighting of different attributes [28] or validate the ability of the metric to produce a ranking between different types of detectors [9]

Some analytic approaches rely exclusively on plausibilization by visualizing selected scenarios without further validation [37, 43, 45]. In other cases, no further attempts to verify or validate the results are made [14].

The usage of planners in a closed loop simulation has been proposed [43] and also executed for safety aware prediction metrics [24]. However, incorporating a planner into the pipeline incurs the potential of errors in the planner which questions the validity [29]. The results of the PKL metric are verified and validated in two different ways. A first verification step argues plausibility by showing that the metric is sensitive to distance and velocity as intuitively salient features. Furthermore, a human evaluation is conducted by Amazon Mechanical Turk workers, showing an 80% preference for the PKL metric over the nuScenes detection score (NDS).

Overall, current validation approaches for validating relevance criteria are lacking. Furthermore, the suggested approaches generally focus on relative ranking and therefore do not provide acceptance criteria.

3 Methodology

We suggest the following approach for defining and validating relevance in this work. First, a formal analytic method of relevance is selected to facilitate interpretability. Next, a neural trajectory prediction model is utilized to validate the results of the formal model. It will later be shown that neural networks are suitable for validation if large-scale datasets are leveraged to consider uncertainties in the prediction outcome.

As formal relevance method, the prior SACRED method [30] is selected. This method provides a conservative estimation aiming to provide a complete set of relevant objects. The property of completeness will also be focus of the later validation procedure. In addition, the relevance method offers the benefit of few requirements regarding environment knowledge.

4 Application of Relevance Selection Method

This section follows the SACRED relevance method defined in [30]. First, a minimum specification is provided. Next, interactions are decomposed into multiple scenarios for each of which relevance criteria are formulated.

4.1 Specification

In order to apply the relevance methodology, it is first necessary to specify the system and the use case.

The system specification follows SACRED [30]. A Sense-Plan-Act architecture is assumed with an object list as interface between perception and planning. The output trajectories of the planner must generally adhere to physical and legal requirements. This work focuses on the task of collision safety. In order to fulfill these requirements, the system provides guarantees regarding its capabilities. After a specified reaction time, the system is required to act in accordance with the requirements. In order to do so, the system is capable of providing a minimum guaranteed braking deceleration as well as a minimum guaranteed acceleration.

Currently, testing of perception is mainly performed on datasets such as [9] which mainly consider urban environments. Therefore, the use case is selected to be the application in the urban domain. This use case serves to expand the ideas from the highway domain to more complex interactions.

4.2 Decomposition

The next step in determining relevance is the decomposition of the use case into functional scenarios. As in SACRED [30], the distinction is made using polar coordinates for the relative velocities, distinguishing radial and tangential scenarios (first letter of abbreviation). Depending upon whether the ego is moving towards or away from the object of interest (OOI)(second letter of abbreviation) and whether the OOI is moving towards or away from the ego (third letter of abbreviation), the four radial scenarios R.TA, R.AT, R.TT and R.AA are distinguished. The tangential scenarios are distinguished depending on whether the OOI is moving towards or away from the ego vehicle. This is described by the two scenarios T.XT and T.XA.

The results for the R.TA, R.AT, R.AA and T.XA scenario are identical to SACRED [30] and are thus not repeated here. However, the R.TT and the T.XT scenario require additional considerations to consider the urban environment. In order to operate in an urban environment a vehicle needs to be able to pass static obstacles by entering the opposite lane and to traverse intersections. Overtaking dynamic objects is excluded to limit complexity, since it is not required to fulfill a driving task. These are considered by modifying the R.TT and the T.XT scenario respectively. Each of the scenarios requiring modification to the relevance estimation is discussed in the following subsections.

4.3 R.TT'

First, the R.TT scenario is applied as in SACRED [30] without modification. For all objects not considered relevant by the R.TT scenario, a modified R.TT scenario is used. This scenario, shown in Fig. 1, hereafter called R.TT', takes into account the case of passing a static object.

Variables follow the notation from SACRED [30] with three optional indices. The first index identifies the object, with 1, 2 and 3 denoting the ego vehicle, opposing vehicle and static object respectively. The second index gives the frame of reference, either r for radial or t for tangential. The third index denotes the state.

We extend the SACRED method [30] which only considers pairwise interactions. First, the static object is considered as OOI. Since the OOI is static, the R.TT and R.TA are equivalent. In order to distinguish the R.TT' scenario, the static object has to be relevant

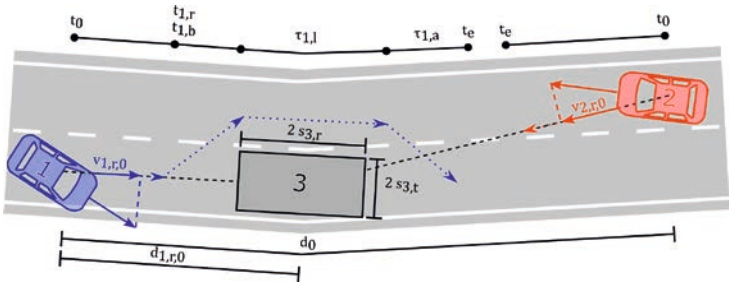

Figure 1: Sequence and variables for the R.TT' scenario.

according to either scenario. Arbitrarily selecting R.TA requires the distance to the static object $d_{1,\mathrm{r},0}$ to be less than the sum of reaction and breaking distance:

$$0 < d_{1,\mathrm{r,min}} = d_{1,\mathrm{r},0} - s_1 - s_{3,\mathrm{r}} - v_{1,\mathrm{r},0}\ t_{1,\mathrm{r}} + \frac{1}{2} a_{\max} t_{1,\mathrm{r}}^2 - \frac{(v_{1,\mathrm{r},0} + t_{1,\mathrm{r}} a_{\max})^2}{2 a_{1,\mathrm{r,b}}} \tag{1}$$

The R.TT' scenario is applied if all following conditions for the initial positions $r_{\mathrm{i},0}$ and velocities $v_{\mathrm{i},0}$ are fulfilled:

- Ego moving towards static object: $(\vec{r}_{3,0} - \vec{r}_{1,0}) \cdot \vec{v}_{1,0} > 0$
- OOI moving towards static object: $(\vec{r}_{3,0} - \vec{r}_{2,0}) \cdot \vec{v}_{2,0} > 0$
- OOI located behind static object: $(\vec{r}_{3,0} - \vec{r}_{1,0}) \cdot (\vec{r}_{3,0} - \vec{r}_{2,0}) < 0$

The scenario is described by the following considerations, according to worst case assumptions. Within the reaction time, the ego vehicle brakes. This results in a distance covered and velocity as defined in (2) and (3). The end time of the braking maneuver $t_{1,\mathrm{b}}$ in (4) is equal to the reaction time or the time to standstill if less.

$$d_{1,\mathrm{b}} = v_{1,\mathrm{r},0} t_{1,\mathrm{b}} - \frac{1}{2} a_{\max} t_{1,\mathrm{b}}^2 \tag{2}$$

$$v_{1,\mathrm{r,b}} = v_{1,\mathrm{r},0} - a_{\max} t_{1,\mathrm{b}} \tag{3}$$

$$t_{1,\mathrm{b}} = min\left\{ t_{1,\mathrm{r}}\ ,\ \frac{v_{1,\mathrm{r},0}}{a_{\max}} \right\} \tag{4}$$

After the reaction time, the passing maneuver is initiated with a lateral movement to the opposite lane, followed by accelerating with the guaranteed acceleration. After passing the static object, the ego vehicle performs a lateral movement back to its initial lane, thus concluding its maneuver. The durations of the lateral lane change $\tau_{1,\mathrm{l}}$ and acceleration $\tau_{1,\mathrm{a}}$ are defined in (5) and (7). The final velocity after accelerating $v_{1,\mathrm{a}}$ is given by (6).

$$\tau_{1,\mathrm{l}} = 2\sqrt{\frac{s_{3,\mathrm{t}} + s_1}{a_{1,\mathrm{g}}}} \tag{5}$$

$$v_{1,\mathrm{r,a}} = v_{1,\mathrm{r,b}} + a_{1,\mathrm{g}}\tau_{1,\mathrm{a}} \tag{6}$$

$$\tau_{1,\mathrm{a}} = \frac{-v_{1,\mathrm{r,b}} + \sqrt{2a_{1,\mathrm{g}}(2s_1 + 2s_{3,\mathrm{r}}) + v_{1,\mathrm{r,b}}^2}}{a_{1,\mathrm{g}}} \tag{7}$$

The final distance the ego vehicle traverses during the scenario is defined in (8) as sum of the individual maneuvers.

$$d_{1,\mathrm{e}} = d_{1,\mathrm{b}} + v_{1,\mathrm{r,b}}\tau_{1,\mathrm{l}} + d_{1,\mathrm{a}} + v_{1,\mathrm{r,a}}\tau_{1,\mathrm{l}} \tag{8}$$

Throughout the scenario, the opposing vehicle performs the maximum acceleration towards the ego, yielding (9) and (10) with (11).

$$d_{2,\mathrm{e}} = v_{2,\mathrm{r},0}t_{\mathrm{e}} + \frac{1}{2}a_{\max}t_{\mathrm{e}}^2 \tag{9}$$

$$v_{2,\mathrm{r,e}} = v_{2,\mathrm{r},0} + a_{\max}t_{\mathrm{e}} \tag{10}$$

$$t_{\mathrm{e}} = t_{1,\mathrm{r}} + 2\tau_{1,\mathrm{l}} + \tau_{1,\mathrm{a}} \tag{11}$$

Given the resulting distances and velocities of the defined events, the relevance of the opposing vehicle can then be determined by using them as the initial values for R.TT from SACRED [30] as defined in equation (12).

$$\begin{aligned} 0 < d_{\min} = (d_0 - d_{1,\mathrm{e}} - d_{2,\mathrm{e}}) - s_1 - s_2 - v_{1,\mathrm{r,a}}\, t_{1,\mathrm{r}} \\ - \frac{1}{2}a_{\max}t_{1,\mathrm{r}}^2 - \frac{(v_{1,\mathrm{r,a}} + t_{1,\mathrm{r}}a_{\max})^2}{2a_{1,\mathrm{r,b}}} - v_{2,\mathrm{r,e}}t_{1,\mathrm{b}} - \frac{1}{2}a_{\max}t_{1,\mathrm{b}} \end{aligned} \tag{12}$$

Exemplary scenarios yield distances in the hundreds of meters, well beyond typical dataset annotation ranges. In addition, passing only needs to be considered given an ego intention. Since this information is not available on datasets, this scenario is neglected for the practical implementation.

4.4 T.XT'

This scenario covers tangential movements such as merging and intersections. The merging covered in SACRED [30] is extended to intersections encountered in the urban domain.

Within an intersection, there are two distinct maneuvers available to the ego vehicle as shown in Fig. 2. It can either merely pass through (A) or turn onto the path of another vehicle travelling (B). The requirement in either case is not to impede another vehicle. The latter case represents the worst case since ego vehicle must not only leave the intersection, but additionally accelerate to an adequate speed.

Thus, the worst case in an intersection is adequately described by the T.XT scenario from SACRED [30]. The main difference is that the lateral distance to the travelling direction of the other vehicle can be arbitrarily large unlike for the case of the highway domain. Therefore, the ego vehicle has the opportunity to avoid entering the intersection by braking in time. The corresponding braking distance is:

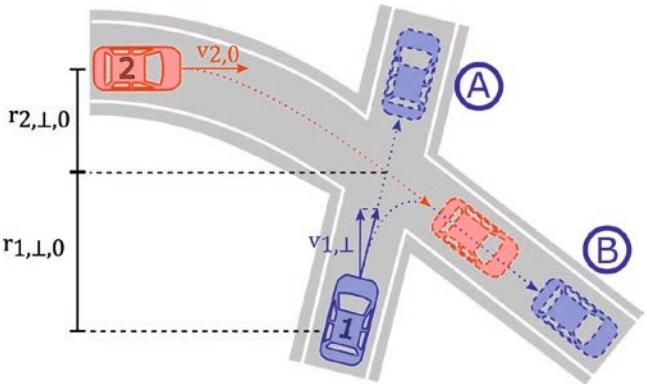

Figure 2: Worst-case intersection and variables in the T.XT' scenario.

$$s_{1,\perp,\mathrm{b}} = v_{1,\perp,0} t_{1,\mathrm{r}} + \frac{v_{1,\perp,0}^2}{2a_{1,\perp,\mathrm{b}}} \tag{13}$$

Additionally, we assume the location of the intersection to be unknown. Therefore, the intersection may be closer to the ego than the heading of the other vehicle. In the worst case, the other vehicle would perform a maximum lateral acceleration towards the ego vehicle to reach the intersection. The corresponding lateral distance traveled with the corresponding braking time of the ego vehicle is:

$$d_{2,\perp} = \frac{1}{2} a_{\max} t_{1,\mathrm{b}}^2 \tag{14}$$

$$t_{1,\mathrm{b}} = t_{1,\mathrm{r}} + \frac{v_{1,\perp,0} + t_{1,\mathrm{r}} a_{\max}}{a_{1,\mathrm{b}}} \tag{15}$$

Considering these two components, the maximum lateral distance for which the merging scenario from SACRED [30] is considered is limited to:

$$r_{1,\perp,0} < v_{1,\perp,0} t_{\mathrm{r}} + \sqrt{\frac{v_{1,\perp,0}^2}{2a_{1,\perp,\mathrm{b}}}} + \frac{1}{2} a_{\max} \left[t_{1,\mathrm{r}} + \frac{v_{1,\perp,0} + t_{1,\mathrm{r}} a_{\max}}{a_{1,\mathrm{b}}} \right]^2 \tag{16}$$

If this equation is fulfilled, the merging scenario from SACRED [30] is applied. Otherwise, the respective object is considered irrelevant according to this scenario.

5 Validation

This section discusses the validation of the results presented thus far. First, some preliminary discussion relating to prior work is shown. Next, the approach of this work is presented. Finally, the approach is applied and verified.

5.1 Preliminaries

As noted in previous sections, the validation of relevance criteria has received limited attention with no generally accepted methodology available. Previous relevance concepts either take an argumentative approach or include a concrete downstream planning task. In addition, the notion of a human baseline is present in one work [35]. We believe that reconciliation of these complimentary approaches is required in order to argue validity.

One basis for the proposed validation approach is the human baseline. This idea has previously been applied for accident rates of human driving performance [25,26,33] as well as for perception performance [39]. Works incorporating downstream planners implicitly also include this concept. However, planning goals may include additional aspects other than imitating human behavior [6] such as infractions, mission goals [16] and comfort [10]. Therefore, these goals are ambiguous and lack consistent evaluation metrics [10].

5.2 Proposed Method

We conceptualize relevance validation by considering the human behavior as baseline. Ideally, the influence of removing objects considered irrelevant might be studied in driving simulators to directly quantify the behavioral changes in human subjects. To avoid the substantial costs such an approach incurs, we propose to approximate human driving behavior with a motion prediction algorithm. This approach follows the approach proposed in [30] which is adapted from [35]. A motion prediction algorithm has several advantages over using a path planning algorithm. The objective of the motion prediction is unambiguous and can be evaluated in an open-loop setting with real-world data without relying on a simulation environment. This work uses the predictor as proxy for human behavior rather than as component to create a full AD pipeline as in [35]. Additionally, this work explicitly quantifies the prediction performance as opposed to prior work which explicitly or implicitly assumes the planner to be valid [24, 35].

The validation procedure consists of running the prediction network with different inputs. Predictions are calculated multiple times for each input to account for potential uncertainties and non-deterministic elements of the prediction. The validation procedure only considers the empirical cumulative distribution function (ECDF) of prediction errors across a large-scale dataset to avoid local performance issues and effects of non-deterministic predictions. A relevance criterion is considered valid if the prediction error distribution remains unchanged between the original input and the filtered input. Whether two distributions are identical is evaluated using the Cramer-von Mises test for two empirical distributions [5]. This test provides a confidence value, based on distribution similarity and sample size, for which an acceptance criterion is required.

To study the utility of the proposed validation method, an implementation on the nuScenes dataset [9] is performed. All experimental results are reported for the standard val split. The prediction network is selected from the nuScenes motion prediction leaderboard [32] among entries with an open implementation. We select the PGP algorithm [31] as implemented by [15] using the default settings. The implementation filters inputs by location in a square region from [-20 m, 80 m] in longitudinal and [-50 m, 50 m] in lateral direction. This filtering is maintained for all experimental conditions with the relevance filtering criteria only being additionally applied. The prediction errors are evaluated using the average distance error (ADE) metric for the top 10 trajectories.

5.3 Results

The prediction network is run ten times each with four different inputs. Firstly, multiple prediction runs with all inputs abbreviated as 'A' are compared to each other to determine the prediction noise. The ECDF of this noise is depicted in Fig. 3, showing that the detector exhibits significant noise averaging at 0.33 m. The error distribution is depicted as ECDF averaging at 0.96 m. Noise also affects the error distributions. Therefore, different runs using identical inputs with all objects (A-A) are compared using statistical tests for equality of distribution. The results are displayed in Fig. 4 as box plot of p-values for different runs. Since the distributions are similar despite the presence of noise, the p-values are high. This provides a reference for the p-values when accounting for noise.

To verify the validation procedure, two artificial verification inputs abbreviated with 'RV' and 'RV2' are constructed and compared with the original input. The first is simply deleting all objects in a scene. The second is to delete all vehicles which are within 2 m from the heading axis of the ego vehicle. Both inputs are constructed to be implausible relevance criteria. The latter filters out 5% of the objects within the heuristic input region of the prediction network, thus filtering out fewer objects than the relevance criteria developed in this work. The error distributions in Fig. 3 exhibit general visual similarity, all being larger than the prediction noise. Nevertheless, the enlarged image indicates that the verification inputs differ from the case using all inputs. Testing for equality of distributions between verification and all inputs (A-RV and A-RV2) in Fig. 4 shows p-values which are orders of magnitude smaller than for A-A. The large spread in p-values is a consequence of prediction noise. Average p-values are below the exemplary threshold value of 0.005 [7]. This indicates a high confidence that the error distributions are different for the verification input and all inputs. Since the prediction is impacted by the verification input, the verification criteria are successfully identified as invalid.

Similar to the verification inputs, the prediction is performed on the input filtered according to the relevance criteria of this work abbreviated as 'R'. The relevance criteria filter out 10% of the objects included within the heuristic input region of the prediction network. Fig. 3 shows that even with an enlarged image, the distributions appear visually similar. The results of testing for equality of all and of relevant inputs (A-R) are shown in Fig. 4 The p-values are similar to A-A, indicating that any dissimilarity in error distributions is within the range caused by noise. Since the prediction performance is unaffected, the relevance criteria of this work are not falsified. Accordingly, the validation results support the relevance criteria from this work.

6 Discussion

This section discusses the results of both the validation method proposed as well as the validation of the extended urban relevance model.

6.1 Validation Method

The proposed validation method relies on a motion prediction network. Compared to PKL [35], fewer requirements are imposed upon the motion prediction. Discrete output trajectories and non-deterministic implementations are applicable without modification.

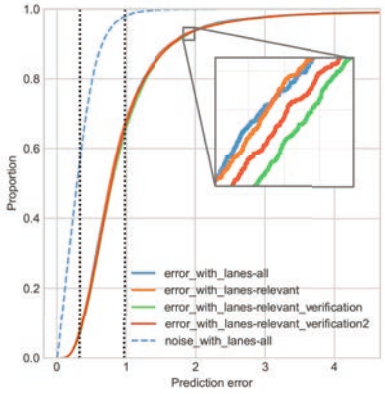

Figure 3: ECDF prediction error as ADE for top 10 trajectories for different inputs and noise for multiple runs with same inputs.

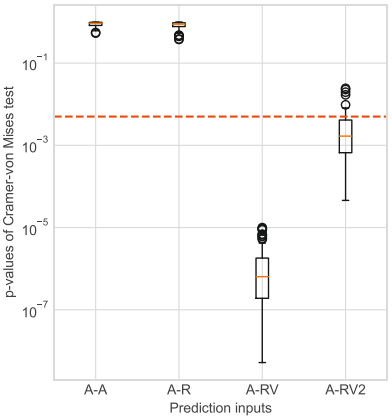

Figure 4: Boxplot of p-values testing for equality of distribution for pairwise combinations. Exemplary threshold indicated by red dashed line.

The method explicitly considers the global performance of the prediction, showing the error to be in the order of 1 m. Notably, the global performance is robust to local noise which may appear in the prediction for a single scenario even with identical inputs. We suspect the inherent uncertainty of future trajectories causes stochastic behavior in the predictions. Since the underlying trajectory distribution is unknown, the prediction performance can only be assessed over a large number of scenarios. Conversely, this means that the local prediction performance in a single scenario is not meaningful since it cannot be disentangled from the stochastic component. Therefore, the local stability of the prediction as used by PKL [35] is not suited as relevance measure for some prediction networks. This is visible in the noise results as exhibited by a non-deterministic prediction implementation as used in this work.

The validation method of this work is successful in falsifying the verification input. The global performance of the prediction deteriorates, which agrees with the intuition that the vehicle directly in front is relevant to the vehicle. This verification succeeds despite the fact that only 5% of the objects are filtered out, as opposed to the 10% for the relevance criteria. However, further study is required to understand the impact of thresholds for p-values. In addition, further research is required to determine to what degree invalid samples can be resolved in large datasets. However, higher sensitivity of the validation may be possible if subsets of specific scenarios are considered individually.

6.2 Extended Urban Relevance Model

Generally, we first develop analytic relevance criteria which are then reconciled with a prediction component at validation. We consider this approach to have the following ad-

vantages. Firstly, no complex neural networks are required when applying the relevance criteria. This is beneficial with regards to implementation effort [48] and computational resources. Additionally, neural networks are prone to failures due to their lack of robustness [22, 46] which may impact directly determining relevance. However, the proposed approach is robust to these failures since the distribution across many samples is considered. Additionally, the analytic relevance criteria are fully interpretable.

These criteria are developed by applying an existing method to derive relevance. The method is successfully applied to extend the results to the urban domain. Applying the proposed validation method on a large-scale dataset shows that the prediction performance is unaffected by the relevance criteria proposed in this work. The validation is currently restricted to a single prediction network and one dataset. For this case, the validity of the relevance criteria is supported. However, only 10% of objects can be excluded from the data based on relevance. It is currently not known if the specificity of the relevance model can be improved while maintaining validity.

7 Conclusion and Outlook

This work presents SURE-Val, the extension of a recent method to derive relevance to the urban domain. For this purpose, the methodology is applied to the intersection and the passing scenario. Additionally, a novel method for validating analytic relevance criteria using a motion prediction is introduced. The relevance criteria and the validation method are applied on a public dataset using an exemplary motion prediction component. The verification of the validation procedure shows that the validation method itself is feasible. At the same time, the validation results support the relevance criteria obtained in this work. Additionally, the analytic relevance criteria are computationally efficient, interpretable and agnostic with regards to the implementation of downstream components. We hope that both the relevance criteria as well as the validation method can serve as reference to consider these aspects more explicitly in the future.

For future research, it is desirable to further explore the limitations of the relevance criteria and their validation. One aspect is the impact of the prediction architecture on relevance. Another aspect is the consideration of datasets including more varied scenarios. Further validation on such datasets which include unusual and dangerous driving situations is required to gain confidence in relevance criteria in general.

8 Acknowledgement

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Climate Action within the project VVM - Verification Validation Methods under grant number 19A19002S, as well as by the German Federal Ministry of Education and Research within the project VIVID with the grant number 16ME0173 based on a decision of the Deutscher Bundestag. The authors would like to thank the consortia for the successful cooperation.

References

[1] Zero-shot deep reinforcement learning driving policy transfer for autonomous vehicles based on robust control. *ArXiv*, 1812.03, 2018.

[2] Mercedes-Benz AG. Mercedes-benz drive pilot. https://www.mercedes-benz.de/passengercars/technology/drive-pilot.html, last accessed 17.07.2023.

[3] Matthias Althoff. *Reachability Analysis and its Application to the Safety Assessment of Autonomous Cars*. Dissertation, Technische Universität München, München, 2010.

[4] Christian Amersbach. *Functional Decomposition Approach: Reducing the Safety Validation Effort for Highly Automated Driving*. Dissertation, Technische Universität Darmstadt, Darmstadt, 26.09.2019.

[5] T. W. Anderson. On the distribution of the two-sample cramer-von mises criterion. *The Annals of Mathematical Statistics*, 33(3):1148–1159, 1962.

[6] Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *ArXiv*, 2018.

[7] Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E-J Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul de Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony G. Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don A. Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E. Johnson. Redefine statistical significance. *Nature human behaviour*, 2(1):6–10, 2018.

[8] Mario Berk, Olaf Schubert, Hans-Martin Kroll, Boris Buschardt, and Daniel Straub. Assessing the safety of environment perception in automated driving vehicles. *SAE International Journal of Transportation Safety*, 8(1):49–74, 2020.

[9] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[10] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric M. Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *ArXiv*, 2021.

[11] Bill Canis. Autonomous vehicles: Emerging policy issues. *Congressional Research Service*, 23.05.2017.

[12] K. Cho, T. Ha, G. Lee, and S. Oh. Deep predictive autonomous driving using multi-agent joint trajectory prediction and traffic rules. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2076–2081, 2019.

[13] VVM Consortium. Verification & validation methods project. https://www.vvm-projekt.de/, last accessed 17.07.2023.

[14] Boyang Deng, C. Qi, Mahyar Najibi, Thomas A. Funkhouser, Yin Zhou, and Drago Anguelov. Revisiting 3d object detection from an egocentric perspective. In *NeurIPS*, 2021.

[15] Nachiket Deo, Eric Wolff, and Oscar Beijbom. Multimodal trajectory prediction conditioned on lane-graph traversals. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 203–212. PMLR, 2022.

[16] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. *ArXiv*, 2017.

[17] Scott M. Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Benjamin Sapp, C. Qi, Yin Zhou, Zoey Yang, Aurelien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Drago Anguelov. Large scale interactive motion forecasting for autonomous driving : The waymo open motion dataset. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9690–9699, 2021.

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.

[19] Sven Hallerbach, Yiqun Xia, Ulrich Eberle, and Frank Koester. Simulation-based identification of critical scenarios for cooperative and automated vehicles. *SAE International Journal of Connected and Automated Vehicles*, 1(2):93–106, 2018.

[20] Franziska Henze, Dennis Fabender, and Christoph Stiller. Identifying admissible uncertainty bounds for the input of planning algorithms. *IEEE Transactions on Intelligent Vehicles*, page 1, 2021.

[21] Michael Hoss, Maike Scholtes, and Lutz Eckstein. A review of testing object-based environment perception for safe automated driving. *Automotive Innovation*, 5(3):223–250, 2022.

[22] Sebastian Houben, Stephanie Abrecht, Maram Akila, Andreas Bär, Felix Brockherde, Patrick Feifel, Tim Fingscheidt, Sujan Sai Gannamaneni, Seyed Eghbal Ghobadi, Ahmed Hammam, Anselm Haselhoff, Felix Hauser, Christian Heinzemann, Marco Hoffmann, Nikhil Kapoor, Falk Kappel, Marvin Klingner, Jan Kronenberger, Fabian

Küppers, Jonas Löhdefink, Michael Mlynarski, Michael Mock, Firas Mualla, Svetlana Pavlitskaya, Maximilian Poretschkin, Alexander Pohl, Varun Ravi-Kumar, Julia Rosenzweig, Matthias Rottmann, Stefan Rüping, Timo Sämann, Jan David Schneider, Elena Schulz, Gesina Schwalbe, Joachim Sicking, Toshika Srivastava, Serin Varghese, Michael Weber, Sebastian Wirkert, Tim Wirtz, and Matthias Woehrle. Inspect, understand, overcome: A survey of practical methods for ai safety. In Tim Fingscheidt, Hanno Gottschalk, and Sebastian Houben, editors, *Deep Neural Networks and Data for Automated Driving*, pages 3–78. Springer Cham, 2022.

[23] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. *ArXiv*, 2020.

[24] Saurabh Jha, Shengkun Cui, Zbigniew Kalbarczyk, and Ravishankar K. Iyer. Watch out for the safety-threatening actors: Proactively mitigating safety hazards. *ArXiv*, 2022.

[25] Philipp Junietz, Udo Steininger, and Hermann Winner. Macroscopic safety requirements for highly automated driving. *Transportation Research Record*, 2673(3):1–10, 2019.

[26] Peng Liu, Run Yang, and Zhigang Xu. How safe is safe enough for self-driving vehicles? *Risk analysis : an official publication of the Society for Risk Analysis*, 39(2):315–325, 2019.

[27] Maria Lyssenko, Christoph Gladisch, Christian Heinzemann, M. Woehrle, Rudolph Triebel, and R. Bosch. From evaluation to verification: Towards task-oriented relevance metrics for pedestrian detection in safety-critical domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 38–45, 2021.

[28] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, Hang Xu, and Chunjing Xu. One million scenes for autonomous driving: Once dataset. *ArXiv*, 2021.

[29] Jiageng Mao, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 3d object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 2023.

[30] Ken Mori, Kai Storms, and Steven Peters. Conservative estimation of perception relevance of dynamic objects for safe trajectories in automotive scenarios. *ArXiv*, 2023.

[31] nachiket92. Pgp: Multimodal trajectory prediction conditioned on lane-graph traversals, 2022. https://github.com/nachiket92/PGP, last accessed 23.05.2023.

[32] nuScenes. nuscenes prediction task: Leaderboard, 2020. https://www.nuscenes.org/prediction, last accessed 23.05.2023.

[33] PEGASUS Project. Pegasus method: An overview, 2019. https://www.pegasusprojekt.de/files/tmpl/Pegasus-Abschlussveranstaltung/PEGASUS-Gesamtmethode.pdf, last accessed 28.06.2022.

[34] Jonah Philion and S. Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, 2020.

[35] Jonah Philion, Amlan Kar, and S. Fidler. Learning to evaluate perception models using planner-centric metrics. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14052–14061, 2020.

[36] Robin Philipp, Hedan Qian, Lukas Hartjen, Fabian Schuldt, and Falk Howar. Simulation-based elicitation of accuracy requirements for the environmental perception of autonomous vehicles. In Tiziana Margaria and Bernhard Steffen, editors, *Leveraging Applications of Formal Methods, Verification and Validation*, pages 129–145, Cham, 2021. Springer International Publishing.

[37] Robin Philipp, Jana Rehbein, Felix Grun, Lukas Hartjen, Zhijing Zhu, Fabian Schuldt, and Falk Howar. Systematization of relevant road users for the evaluation of autonomous vehicle perception. In *2022 IEEE International Systems Conference (SysCon)*, pages 1–8.

[38] Robin Philipp, Fabian Schuldt, and Falk Howar. Functional decomposition of automated driving systems for the classification and evaluation of perceptual threats. In *13. Uni-DAS eV Workshop Fahrerassistenz und automatisiertes Fahren 2020*. Walting, 2020.

[39] Charles R. Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa Vo, Boyang Deng, and Dragomir Anguelov. Offboard 3d object detection from point cloud sequences. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6130–6140, 2021.

[40] A. Sadat, S. Casas, Mengye Ren, X. Wu, Pranaab Dhawan, and R. Urtasun. Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. In *ECCV*, 2020.

[41] Valerij Schönemann, Mara Duschek, and Hermann Winner. Maneuver-based adaptive safety zone for infrastructure-supported automated valet parking. *2184-495X*, 2019.

[42] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Sheng Zhao, Shuyang Cheng, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. *ArXiv*, 2019.

[43] Sever Topan, Karen Leung, Yuxiao Chen, Pritish Tupekar, Edward Schmerling, Jonas Nilsson, Michael Cox, and Marco Pavone. Interaction-dynamics-aware perception zones for obstacle detection safety evaluation. *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1201–1210, 2022.

[44] José Luis Vázquez, Alexander Liniger, Wilko Schwarting, Daniela Rus, and Luc van Gool. Deep interactive motion prediction and planning: Playing games with motion prediction models. *ArXiv*, 2022.

[45] Georg Volk, Jörg Gamerdinger, Alexander von Betnuth, and O. Bringmann. A comprehensive safety metric to evaluate perception in autonomous systems. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8, 2020.

[46] Oliver Willers, Sebastian Sudholt, Shervin Raafatnia, and Stephanie Abrecht. Safety concerns and mitigation approaches regarding the use of deep learning in safety-critical perception tasks. In António Casimiro, editor, *Computer safety, reliability, and security*, volume 12235 of *Lecture Notes in Computer Science*, pages 336–350. Springer, Cham, 2020.

[47] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, 2021.

[48] Mirja Wolf, Luiz R. Douat, and Michael Erz. Safety-aware metric for people detection. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2759–2765. IEEE, 2021.