

# A Concept for Requirements-Driven Identification and Mitigation of Dataset Gaps for Perception Tasks in Automated Driving

Mohamed Sabry Moustafa<sup>1</sup><sup>a</sup>, Maarten Bieshaar<sup>1</sup><sup>b</sup>, Andreas Albrecht<sup>1</sup> and Bernhard Sick<sup>2</sup><sup>c</sup>

<sup>1</sup>Robert Bosch GmbH, Germany

<sup>2</sup>University of Kassel, Chair of Intelligent Embedded Systems, Germany

**Keywords:** Deep Learning, Dataset Coverage Gaps, Safety-Critical Applications, Requirements-Driven Engineering, Data-Driven Engineering, Datasets Augmentation, Synthetic Datasets.


**Abstract:** The development of reliable perception machine learning (ML) models is critical for the safe operation of automated vehicles. However, acquiring sufficient real-world data for testing and training these models is not only time-consuming and dependent on chance, but also presents significant risks in safety-critical situations. To address these challenges, we propose a novel requirements-driven, data-driven methodology leveraging state-of-the-art synthetic data generation techniques in combination with tailoring real-world datasets towards task-specific needs. Our approach involves creating synthetic scenarios that are challenging or impossible to capture in real-world environments. These synthetic datasets are designed to enhance existing real-world datasets by addressing coverage gaps and improving model performance in cases represented by such gaps in real world. Through a rigorous analysis based on predefined safety requirements, we systematically differentiate between gaps arising from insufficient knowledge about the system operational design domain (e.g., underrepresented scenarios) and those inherent to data. This iterative process enables identifying and mitigating underrepresented scenarios, particularly in safety-critical and underrepresented scenarios, leading to local improvement in model performance. By incorporating synthetic data into the training process, our approach effectively mitigates model limitations and contributes to increased system reliability, in alignment with safety standards such as ISO-21448 (SOTIF).


## 1 INTRODUCTION


Deep learning (DL) has significantly advanced computer vision, particularly in tasks like classification, object detection, and segmentation. These advancements enable the integration of DL models into complex systems such as highly-automated vehicles, which demand reliable performance in safety-critical scenarios. The performance of such perception models is evaluated against pre-defined data requirements within the Operational Design Domain (ODD), defining conditions for safe operation. Developing such perception models to perform reliably in safety-critical applications requires training datasets that align with system-level requirements for the specific ODD (Metzen et al., 2023). However, constructing a dataset that fully meets these requirements is challenging,

and this work addresses this critical step within the DL development cycle (Gauerhof et al., 2020). Furthermore, SOTIF (Safety of the Intended Functionality) (Expósito Jiménez et al., 2024) is a safety standard (ISO 21448) that addresses hazards arising from the correct functioning of a system but in unsafe scenarios, particularly relevant in automated and autonomous systems. It focuses on situations where the system performs as designed, yet due to limitations in the design, environmental factors, or the system's interpretation of complex scenarios, safety can be compromised. Therefore, to mitigate risks, a comprehensive ODD analysis is essential, ensuring that both safety-critical and non-critical cases are represented in the data. However, not all scenarios can be covered without investigating specific hazards, as such situations may lead to models making uncertain predictions.

To assess model performance, recorded data - sometimes requiring additional labeling - must be evaluated against the target values specified in the system requirements. The performance of a machine learning

<sup>a</sup> <https://orcid.org/0009-0000-2260-3978>

<sup>b</sup> <https://orcid.org/0000-0002-6471-6062>

<sup>c</sup> <https://orcid.org/0000-0001-9467-656X>

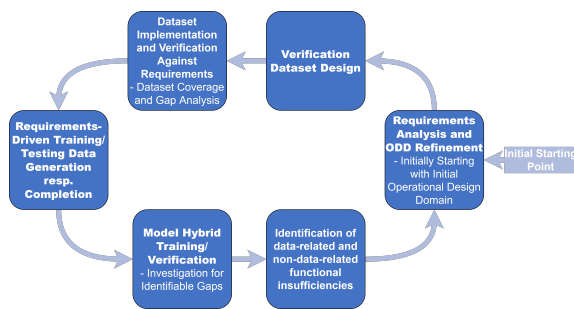


Figure 1: Proposed workflow cycle to iteratively identify and mitigate data-related deficiencies and adapt requirements. Our approach targets building an iterative cycle of model optimization and data refinement towards better dataset designs, including adaptation of ML (data) requirements.

model is influenced by both systematic effects, such as the model’s inability to handle out-of-distribution sample, corner cases, or domain shifts, and by stochastic effects, such as ambiguities or inherent uncertainties in the data (Hüllermeier and Waegeman, 2021). In this context, we rely on the definition of uncertainty in ML that can either stem from systematic gaps in the model’s knowledge and can be mitigated by supplying additional data which is defined as epistemic uncertainty. On the other hand, uncertainty arising from noise or stochastic ambiguities inherent in the data itself and not addressable through augmenting more data samples is defined as aleatoric uncertainty.

This article presents an approach for evaluating perception models against safety-related requirements by using synthetic and hybrid datasets to systematically identify and mitigate performance gaps due to functional deficiencies. This iterative evaluation process helps pinpoint two types of performance gaps (Hüllermeier and Waegeman, 2021; Zhang et al., 2021).

The first type of gaps is model-related gaps where systematic model-related gaps can arise from a mismatch of the model’s learning capacity and the functional input-output behavior to be learned given by the training data due to sub-optimal architectural designs. Such mismatch is labeled as model uncertainty (Hüllermeier and Waegeman, 2021). This uncertainty may lead to overfitting, underfitting or inductive biases, which lead to performance limitations within the ODD. As such gaps are systematic in nature they can be categorized under epistemic uncertainty.

The second type is data-relevant gaps that stem from discrepancies between the training data and the real-world environment. Data gaps can arise from (1) data capture issues, introducing aleatoric uncertainty through noise or ambiguity, and (2) limitations in data scope or representativeness, completeness or diversity, for example, creating epistemic uncertainty.

Both types of data-related gaps can lead to underperformance regarding potentially safety-critical scenarios.

In this article, we focus specifically on gaps that models experience epistemic uncertainty with respect to that we label as datasets coverage gaps. To address those gaps, we propose a hybrid training method that generates synthetic samples to target and close identified gaps, iteratively augmenting them to real data. This systematic approach aims to enhance perception model performance in these local data space regions while improving overall performance and generalization capabilities, ensuring compliance with safety requirements for real-world applications.

## 1.1 Possible Contributions

Systematically analyzing deep perception models and their training and test datasets to identify functional deficiencies and coverage gaps and comparing them with the target ODD using a structured, requirements-driven approach with synthetic data is an emerging research area with limited contributions (Metzen et al., 2023; Boreiko et al., 2023; Boreiko et al., 2024; Zhang et al., 2021).

Our approach aims to establish a baseline for an iterative data-driven engineering loop to systematically test for model performance against different elements of the ODD through constructing test datasets that can identify coverage gaps in training datasets that cause poor model performance. This can be achieved through state of the art data generation techniques that help generate data on demand based on defined sets of requirements.

Once datasets coverage gaps are identified, the next step is to address the coverage issues by introducing synthetically generated data samples with missing contents or properties into the training process by augmentation to close the identified coverage gaps, re-train the models, and evaluate their performance in an iterative and combined top-down and bottom-up approach starting from requirements. This way, we can combine top-down aspects from safety with bottom-up aspects from model and dataset analysis. Our top-down approach is adapted from (Zhang et al., 2021) establishing a systematic data-driven engineering loop for automated driving systems and can be seen in 1.

## 2 ALL ABOUT DATASET COVERAGE GAPS

Dataset coverage gaps occur in local data regions lacking sufficient representation of certain characteristics, such as scenario or feature classes. These gaps can

lead to missing data points, imbalanced class distributions, or limited variations, preventing models from learning the full spectrum of patterns needed for reliable predictions. Consequently, models trained with such gaps may perform well in training but struggle in real-world scenarios containing unseen or under-represented patterns. Closing these gaps is essential for building models that generalize well across diverse conditions.

## 2.1 Defining Dataset Coverage Gaps

Dataset gaps can lead to model biases, hindering generalization and causing high error rates on new data, especially in safety-critical applications. Strategies like data augmentation, synthetic data generation, and goal-oriented data collection can mitigate these issues. Understanding coverage gaps requires knowledge of ML pipelines, as shown in Fig. 2. Key stages include data preparation, model training, and deployment. Data collection and exploration are crucial for identifying gaps and inconsistencies, such as rare events, long-tail issues, sampling biases, and domain shifts. Datasets must adequately represent the ODD to ensure model effectiveness. Discrepancies between training data and the target ODD can lead to systematic errors.

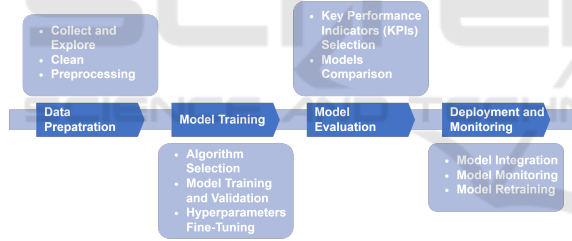


Figure 2: A typical ML pipeline from data preparation to deployment.

## 2.2 Safety of the Intended Functionality and Triggering Conditions

One approach to identify Datasets gaps is by examining the divergence between training and target distributions within the intended ODD, particularly for image classifiers. Let the classifier  $f_{\Theta} : \mathbf{X} \mapsto \mathbf{Y}$  predict the probability  $f_{\Theta}(\mathbf{y}|\mathbf{x})$  for class  $\mathbf{y} \in \mathbf{Y} = \{1, \dots, \mathbf{C}\}$  given an image  $\mathbf{x} \in \mathbf{X}$ , where  $\mathbf{x}$  and  $\mathbf{y}$  follow a distribution  $p(\mathbf{x}, \mathbf{y})$  and  $\Theta$  denotes model parameters. Here, we focus on the model's behavior across semantically coherent data subgroups (e.g., "red cars in an urban setting"), represented by conditioning on a latent variable  $\mathbf{o}$  with  $p(\mathbf{x}, \mathbf{y}|\mathbf{o})$ .

The ODD is compositional,  $O = O_0 \times \dots \times O_{n_O-1}$ , where each  $O_i$  is a semantic dimension. Each  $\mathbf{o} \in O$

is a tuple with  $n_o$  values, as exemplified in Table 1 for pedestrian detection in automated driving.

Table 1: Example ODD variables for pedestrian detection in automated driving, with valid ranges.

ODD Variable	Valid Range or Set
Time of Day ( $O_0$ )	$[T_{min}, T_{max}]$ (e.g., 6 AM - 6 PM)
Weather ( $O_1$ )	{Sunny, Cloudy, Rainy}
$\vdots$	$\vdots$
Distance to pedestrian ( $O_N$ )	$[d_{min}, d_{max}]$

Instead of a deterministic mapping between inputs and outputs, we use a probabilistic approach to model  $p(\mathbf{y}|\mathbf{x}, \Theta)$  (Bishop and Nasrabadi, 2006; Hüllermeier and Waegeman, 2021). This probabilistic framework is essential in capturing prediction uncertainties, particularly for safety-critical applications like automated driving. Integrating the ODD within this framework, we acknowledge that both  $\mathbf{x}$  and  $\mathbf{y}$  are influenced by ODD variables  $\mathbf{o} \in O$ . These variables represent conditions under which the system operates, such as time of day or weather. Figure 4 illustrates this probabilistic approach, with a focus on potential data gaps that impact performance within the ODD.

Dataset gaps are quantified by comparing semantically coherent subgroup distributions in the training data and target ODD using the marginal probability  $p(\mathbf{o})$ , which indicates subgroup  $\mathbf{o}$ 's representation. Significant differences reveal underrepresented or missing subgroups. In this regard, SOTIF (Safety of the Intended Functionality) is a safety standard (ISO 21448) that addresses hazards arising from the correct functioning of a system but in unsafe scenarios, particularly relevant in automated and autonomous systems. It focuses on situations where the system performs as designed, yet due to limitations in the design, environmental factors, or the system's interpretation of complex scenarios, safety can be compromised. Formally, a triggering condition exists in an input subspace  $\mathbf{x}_{trig} \in \mathbf{X}_{trig} \subset \mathbf{X}$  where the model  $\Theta$  is susceptible to error if a certain threshold  $\delta_{trig}$  is surpassed:

$$p(\mathbf{y}_{error}|\mathbf{x}_{trig}^{(i)}, \Theta) > \delta_{trig}, \quad (1)$$

where  $\mathbf{x}_{trig}^{(i)} \in \mathbf{X}_{trig}^{(i)}$  is linked to the  $i$ -th triggering condition and  $\mathbf{y}_{error}$  represents erroneous predictions. While some conditions may be outside the ODD, we limit our analysis to  $\mathbf{o}_{trig} \in O$  under the assumption that subspaces  $\mathbf{X}_{trig}^{(i)}$  are well-represented by  $p(\mathbf{x}|\mathbf{o}_{trig})$ , thus forming the subspace  $\mathbf{X}|\mathbf{O}_{trig}^{(i)}$ . This enables systematic testing across ODD scenarios. Triggering conditions outside  $O$  may require supervision and fallback mechanisms (Mekki-Mokhtar et al., 2012). The risk from such conditions can be assessed using joint

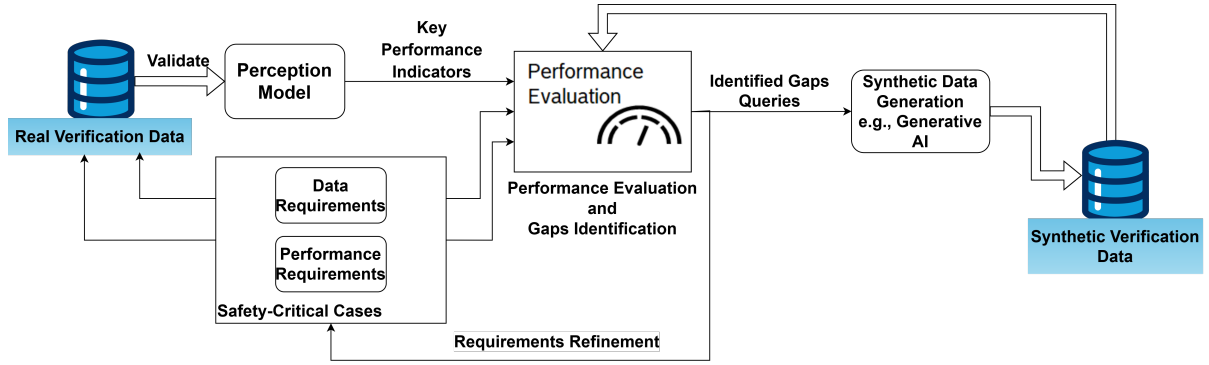


Figure 3: Verification loop using synthetic data to identify coverage gaps, creating test datasets to assess model deficiencies and meet requirements.

distributions of failure probability, criticality, controllability, and severity (Expósito Jiménez et al., 2024). Minimum performance targets for perception in automated driving are common, as noted in (Zhang et al., 2021). For instance, object detection is often evaluated with metrics like log-average miss rate, which reflects the probability of failure. Performance metrics, as seen in Eq. 1, serve to gauge model reliability across the ODD and for specific triggering conditions. Let  $s \in \mathbb{R}$  denote model performance, which is assessed against targets in data requirements. Performance variations stem from systematic issues (e.g., handling out-of-distribution samples (Hendrycks and Gimpel, 2016), corner cases (Heidecker et al., 2024), or domain shifts (Candela et al., 2009)) and stochastic factors due to ambiguity or data uncertainty (Hüllermeier and Waegeman, 2021). To model the performance  $s$  for a given ODD point  $\mathbf{o}$ , we use a probability density  $p_{\Theta}(s|\mathbf{o})$ , allowing for uncertainty assessments per gap. This indicates if a gap stems from epistemic uncertainty, where more data would aid learning, or aleatoric uncertainty, which is irreducible (Hüllermeier and Waegeman, 2021). Model  $f_{\Theta}$  achieves acceptable performance in an ODD region if:

$$P_{\Theta}(s \geq \tau|\mathbf{o}) \geq 1 - \alpha, \quad (2)$$

where  $\alpha \in [0, 1]$  is the confidence level. When Eq. 2 does not hold, this signals potential performance issues in specific ODD conditions. Since obtaining exact values for  $p_{\Theta}(s|\mathbf{o})$  is impractical, we approximate it with a parameterized distribution  $p_{\Theta}(s|\mathbf{o}, \Psi)$ , where  $\Psi$  represents the parameters of the distribution. Therefore, we estimate performance for model  $\Theta$  at input  $\mathbf{o}$ . Using Bayesian parameter estimation (Bishop and Nasrabadi, 2006), the true probability  $p_{\Theta}(s|\mathbf{o})$  can be inferred via:

$$p_{\Theta}(s|\mathbf{o}) = \int p_{\Theta}(s|\mathbf{o}, \Psi) p(\Psi) d\Psi. \quad (3)$$

This probabilistic approach facilitates categorizing uncertainties as: (1) Data Gaps, due to insufficient data

for reliable performance estimates, and (2) Confirmed Performance Gaps, where model deficiencies are clear. The total uncertainty  $S(s)$  of the performance  $s$  can be expressed as:

$$S(s) = I(s, \Psi) + S(s|\Psi), \quad (4)$$

where  $S(s)$  represents the total uncertainty in the performance of the model.  $I(s, \Psi)$  represents the epistemic uncertainty. It is the mutual information between the model performance  $s$  and the model parameters  $\Psi$ , indicating how much uncertainty about the performance can be attributed to uncertainty in the model parameters. This component can be reduced by gathering more data or improving model training. Finally,  $S(s|\Psi)$  represents the aleatoric uncertainty. It quantifies the residual uncertainty in the model's performance given the model parameters  $\Psi$ . This uncertainty is inherent to the data and cannot be reduced by additional data collection (Hüllermeier and Waegeman, 2021).

### 2.3 How Dataset Coverage Gaps Can Affect Performance

One of the key performance indicators that dataset coverage gaps can influence is the model risk. Hence, we can investigate the probabilistic model introduced in 2.2 by introducing multiple granularity levels of gaps and investigate how they can influence the model risk. The conditional probability  $p(\mathbf{x}|\mathbf{y}, \mathbf{o})$  describes the distribution of images  $\mathbf{x}$  for class  $\mathbf{y}$  and subgroup  $\mathbf{o}$ , identifying visual gaps (e.g., "red ( $o_i$ ) cars ( $y$ ) in forests ( $o_j$ )"). Similarly,  $p(\mathbf{y}|\mathbf{o})$ , the likelihood of class  $\mathbf{y}$  within subgroup  $\mathbf{o}$ , highlights class imbalances or biases, signaling potential classification errors.

Decomposing the joint distribution  $p(\mathbf{x}, \mathbf{y}, \mathbf{o})$  into these components (Figure 4) helps identify subgroup-level gaps where  $p(\mathbf{o})$  reveals missing or underrepresented subgroups, class-level gaps within subgroups



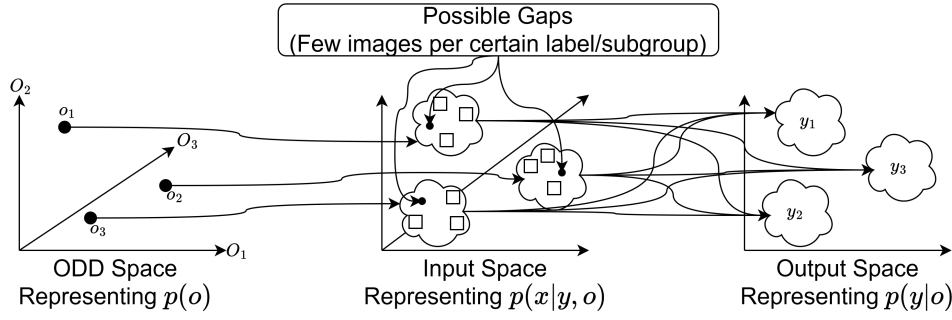


Figure 4: Schematic illustration of data-related gaps arising during model training, focusing on gaps in semantically coherent subgroups of image data and class distributions.

such that  $p(\mathbf{y}|\mathbf{o})$  identifies class imbalances in specific subgroups, causing biased predictions, in addition to image-level gaps can be identified since  $p(\mathbf{x}|\mathbf{y}, \mathbf{o})$  highlights limited visual diversity in class-subgroup pairs. While  $p(\mathbf{o})$  directly captures subgroup coverage,  $p(\mathbf{x}|\mathbf{y}, \mathbf{o})$  and  $p(\mathbf{y}|\mathbf{o})$  provide insights into image and class distributions. In contrast, other quantities like  $p(\mathbf{o}|\mathbf{y})$  or  $p(\mathbf{o}|\mathbf{x}, \mathbf{y})$  are less intuitive for detecting gaps, as they obscure subgroup representation.

The influence of coverage gaps on model performance can be analyzed by associating coverage gaps with an estimated risk posed by the model on each subgroup and overall. The risk of a classifier to fail on a specific subgroup represents the expected loss of the classifier over the distribution of data points within that subgroup (Metzen et al., 2023). The risk of a classifier  $f_{\Theta}$  on a subgroup  $\mathbf{o}$  is given by:

$$R_{f_{\Theta}}(\mathbf{o}) = \mathbf{E}_{(\mathbf{x}, \mathbf{y}) \sim p(\mathbf{x}, \mathbf{y}|\mathbf{o})} [L(f_{\Theta}(\cdot|\mathbf{x}), \mathbf{y})]. \quad (5)$$

Here, the expectation is taken over the conditional distribution of images  $\mathbf{x}$  and labels  $\mathbf{y}$  given subgroup  $\mathbf{o}$ , and  $L(f(\cdot|\mathbf{x}), \mathbf{y})$  is the loss function that measures the discrepancy between the classifier's prediction  $f(\cdot|\mathbf{x})$  and the true label  $\mathbf{y}$ . The loss function  $L: [0, 1]^C \times \mathcal{Y} \mapsto \mathbb{R}$  applies to  $C$  classes, with  $\mathcal{Y}$  being the label space. The risk  $R_{f_{\Theta}}(\mathbf{o})$  reflects the average loss over all data points in subgroup  $\mathbf{o}$ . Using probabilistic quantities, it can be rewritten as:

$$R_{f_{\Theta}}(\mathbf{o}) = \int_{\mathbf{x}} \sum_{\mathbf{y}} L(f_{\Theta}(\cdot|\mathbf{x}), \mathbf{y}) p(\mathbf{y}|\mathbf{o}) p(\mathbf{x}|\mathbf{y}, \mathbf{o}) d\mathbf{x}. \quad (6)$$

This formulation shows that risk depends on both the probability of observing an image  $\mathbf{x}$  given a class  $\mathbf{y}$  and subgroup  $\mathbf{o}$ , i.e.,  $p(\mathbf{x}|\mathbf{y}, \mathbf{o})$ , and the class distribution within the subgroup  $p(\mathbf{y}|\mathbf{o})$ . It highlights how risk is influenced by both image distribution and class distribution within subgroups. The total expected risk over all subgroups is the weighted sum of the risks for each subgroup, with weights determined by the

subgroup probabilities  $p(\mathbf{o})$ :

$$R_{f_{\Theta}} = \sum_{\mathbf{o}} p(\mathbf{o}) R_{f_{\Theta}}(\mathbf{o}) = \sum_{\mathbf{o}} p(\mathbf{o}) \left( \int_{\mathbf{x}} \sum_{\mathbf{y}} L(f_{\Theta}(\cdot|\mathbf{x}), \mathbf{y}) p(\mathbf{x}|\mathbf{y}, \mathbf{o}) p(\mathbf{y}|\mathbf{o}) d\mathbf{x} \right). \quad (7)$$

This equation shows that the total risk  $R_{f_{\Theta}}$  is shaped by the individual risks  $R_{f_{\Theta}}(\mathbf{o})$  for each subgroup, weighted by the probability of each subgroup  $p(\mathbf{o})$ . Low representation or coverage of a subgroup in the training data can lead to elevated risk for the following reasons:

- **Inadequate Representation:** When  $p_{\text{train}}(\mathbf{o})$  is small, insufficient samples from subgroup  $\mathbf{o}$  impair the model's generalization capability, leading to  $R_{f_{\Theta}}(\mathbf{o})$  increasing.
- **Loss Sensitivity:** With few data points,  $p(\mathbf{x}, \mathbf{y}|\mathbf{o})$  amplifies the effect of loss  $L(f(\cdot|\mathbf{x}), \mathbf{y})$  on  $R_{f_{\Theta}}(\mathbf{o})$ , especially if the model struggles with rare samples.
- **Inverse Relation of Probability and Risk:** As expressed in:

$$R_{f_{\Theta}}(\mathbf{o}) \approx \frac{1}{p(\mathbf{o})} \mathbf{E}_{(\mathbf{x}, \mathbf{y}) \sim p(\mathbf{x}, \mathbf{y}|\mathbf{o})} [L(f_{\Theta}(\cdot|\mathbf{x}), \mathbf{y}) p(\mathbf{x}, \mathbf{y}|\mathbf{o})]. \quad (8)$$

Low  $p(\mathbf{o})$  typically leads to higher  $R_{f_{\Theta}}(\mathbf{o})$ , implying that even small errors in low-probability subgroups can significantly raise the overall risk.

### 3 TRAINING DEEP PERCEPTION MODELS WITH MIXED REAL AND SYNTHETIC DATA

To address challenges in data scarcity, cost, and annotation time, synthetic data generation has become essential in training deep perception models (Liu and

Mildner, 2020). This method offers diverse and abundant examples, enhancing model reliability and generalization, especially in safety-critical applications like automated vehicles and surveillance. Synthetic datasets can also replicate real-world scenarios, accelerating model training for novel environments (Song et al., 2023). Synthetic data generation encompasses various methods: **(1) 3D Engines** like Unreal Engine and Unity create photo-realistic datasets such as SYNTHIA (Ros et al., 2016), VirtualKitti (Gaidon et al., 2016), and VirtualKitti V2 (Cabon et al., 2020); **(2) Video Game Capture** leverages game imagery, producing datasets like DITM (Johnson-Roberson et al., 2016) and VIPER (Richter et al., 2017); **(3) Generative AI** employs techniques like GANs (Wang et al., 2017), diffusion models (Zhang et al., 2023), and unsupervised learning (Hu et al., 2023) to synthesize high-quality images.

However, synthetic data alone introduces a domain gap that may limit real-world generalizability (Tremblay et al., 2018). We propose a mixed training strategy that integrates real and synthetic datasets, combining their strengths to better handle real-world scenarios. Techniques like domain randomization (Zhu et al., 2023; Yue et al., 2019; Tremblay et al., 2018) and curriculum learning (Wang et al., 2021; Soviany et al., 2022) aid in bridging this gap, transferring reliable features learned from synthetic data to real-world environments. This mixed approach improves model performance, reliability, and safety in complex applications (Keser et al., 2021; Schneider and Stemmer, 2023).

## 4 REQUIREMENTS-DRIVEN DATASET COVERAGE GAPS IDENTIFICATION AND MITIGATION

Developing reliable perception models for safety-critical uses requires datasets capturing safety features across task- and scenario-specific aspects of the ODD. Gaps in datasets coverage often lead to functional deficiencies in ML models, especially in critical situations. Therefore, identifying and addressing these gaps systematically is essential (Zhang et al., 2021). Missing data samples can be generated (Boreiko et al., 2024; Boreiko et al., 2023) or sampled (Settles, 2009) to improve model performance within those specific scenarios. This section outlines the two main stages of our concept, illustrated in Fig. 3 and 5. Given a pre-trained perception model and initial training dataset, we first curate test datasets aligning with the ODD to evaluate

model performance against specific criteria, identifying coverage gaps using state-of-the-art synthetic data. In the second stage, these gaps are iteratively closed by generating new training samples through the same pipeline, refining data and performance requirements based on emerging insights. This approach sets a new SOTA in controlled, requirements-driven data generation and training for real and synthetic datasets.

### 4.1 Systematic Identification of Performance Deficiencies

Verifying perception models against predefined requirements to identify performance deficiencies is an active area of research (Gauerhof et al., 2020; Hawkins et al., 2021), expected to grow with the formal release of ISO 21448 (SOTIF) (Expósito Jiménez et al., 2024). Prior work (Metzen et al., 2023; Boreiko et al., 2023; Boreiko et al., 2024) has developed methods to verify DL models on rare data subgroups by defining requirements that target specific visual and geometric features. A primary challenge remains in constructing test datasets that align with requirements to expose potential coverage gaps, thereby enhancing understanding of model performance across different data regions. Achieving this requires samples that can reveal performance deficiencies in critical ODD data spaces. Consequently, it is essential to introduce additional data samples in those regions identified as high-priority (Expósito Jiménez et al., 2024).

#### 4.1.1 Model Verification Against Requirements

We assess pre-trained perception models against safety-critical cases in line with data requirements, as described in (Metzen et al., 2023) and can be seen in figure 3. Once gaps are identified, we apply n-wise combinatorial testing to capture a broader subspace of the ODD, reducing test cases while preserving coverage (Metzen et al., 2023). This approach mitigates combinatorial explosion by selectively increasing n for safety-critical combinations only. Further analysis of deficiencies across feature subgroups at varying granularity levels, using sensitivity analysis (Zhang et al., 2021) and uncertainty measures (Hüllermeier and Waegeman, 2021), pinpoints epistemic uncertainty gaps in training data coverage. Off-the-shelf datasets are often inadequate for this task. Hence, (Boreiko et al., 2023) proposes a synthetic sample generation pipeline to discover rare data subgroups for classification models. We adapt this approach to iteratively generate synthetic datasets in a requirements-driven manner, ensuring alignment with the ODD for thorough model testing.

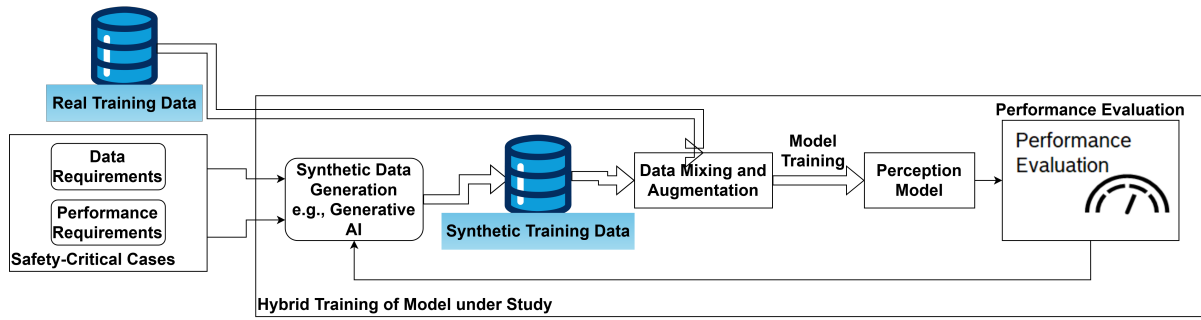


Figure 5: Our requirements-driven hybrid training approach: an iterative process utilizing insights from Fig. 3 to identify gaps, generate synthetic samples, and augment training data to improve model performance.

## 4.2 Targeted Data Generation for Mitigating Gaps

Active learning (Settles, 2009) is an effective method for training detectors with real and synthetic data, as it selects samples based on model uncertainty, focusing near the decision boundary and employing a bottom-up approach from the model’s perspective (Settles, 2009). However, in safety-critical applications, active learning lacks contextual insight into the ODD, system architecture, sensor configuration, and safety measures. We propose enhancing active learning with a requirement-based dataset design to add top-down safety perspectives absent in the model’s view.

### 4.2.1 Enhancing Coverage and Model Performance

To our knowledge, no prior research has been done on integrating controlled synthetic data generation with active learning. Moreover, as incorporating the ODD’s safety-critical aspects within the active learning framework seems feasible, we propose a combination of active learning and a structured data-driven engineering approach. In part 2 as seen in figure 5, we use synthetic samples to systematically close real-world dataset gaps identified in part 1, improving detector performance in critical data regions. We intend to examine the influence of appearance gaps (e.g., color distortion, motion blur) and metadata discrepancies (e.g., environmental conditions, object position) between real and synthetic data, as well as potential label and distribution differences to perform reliable augmentation of synthetic data to real-world training datasets and address the identified gaps.

## 5 CONCLUSION

In this article, we presented a requirements-driven methodology to systematically investigate ML per-

ception models for model- and data-related gaps that could lead to performance deficiencies, particularly in safety-critical contexts. By employing state-of-the-art synthetic data generation, we first create test datasets aligned with predefined safety requirements, allowing us to identify potential performance gaps in these critical scenarios. We then utilize a probabilistic model to quantify these gaps, linking ML model limitations to data coverage issues. When validated, these gaps are addressed by augmenting training datasets with targeted samples to improve the model’s reliability and generalization.

This iterative augmentation incorporates synthetic samples representing identified gaps, creating a hybrid dataset of real-world and synthetic data for retraining the ML perception models. Our top-down approach, guided by explicit ML requirements and compliant with SOTIF (ISO-21448) (Expósito Jiménez et al., 2024) standards, rigorously upholds safety measures throughout the process.

## REFERENCES

- Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 1. springer New York.
- Boreiko, V., Hein, M., and Metzen, J. H. (2023). Identifying systematic errors in object detectors with the scrod pipeline. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4090–4099.
- Boreiko, V., Hein, M., and Metzen, J. H. (2024). Identification of fine-grained systematic errors via controlled scene generation. *arXiv preprint arXiv:2404.07045*.
- Cabon, Y., Murray, N., and Humenberger, M. (2020). Virtual kitti 2. *arXiv preprint arXiv:2001.10773*.
- Candela, J. Q., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). Dataset shift in machine learning. *The MIT Press*, 1:5.
- Expósito Jiménez, V. J., Macher, G., Watzenig, D., and Brenner, E. (2024). Safety of the intended functionality

- validation for automated driving systems by using perception performance insufficiencies injection. *Vehicles*, 6(3):1164–1184.
- Gaidon, A., Wang, Q., Cabon, Y., and Vig, E. (2016). Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349.
- Gauerhof, L., Hawkins, R., Picardi, C., Paterson, C., Hagiwara, Y., and Habli, I. (2020). Assuring the safety of machine learning for pedestrian detection at crossings. In *Computer Safety, Reliability, and Security: 39th International Conference, SAFECOMP 2020, Lisbon, Portugal, September 16–18, 2020, Proceedings 39*, pages 197–212. Springer.
- Hawkins, R., Paterson, C., Picardi, C., Jia, Y., Calinescu, R., and Habli, I. (2021). Guidance on the assurance of machine learning in autonomous systems (amlas). *arXiv preprint arXiv:2102.01564*.
- Heidecker, F., Bieshaar, M., and Sick, B. (2024). Corner cases in machine learning processes. *AI Perspectives & Advances*, 6(1):1.
- Hendrycks, D. and Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.
- Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., and Corrado, G. (2023). Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*.
- Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506.
- Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S. N., Rosaen, K., and Vasudevan, R. (2016). Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*.
- Keser, M., Savkin, A., and Tombari, F. (2021). Content disentanglement for semantically consistent synthetic-to-real domain adaptation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3844–3849. IEEE.
- Liu, T. and Mildner, A. (2020). Training deep neural networks on synthetic data. *LU-CS-EX*.
- Mekki-Mokhtar, A., Blanquart, J.-P., Guiochet, J., Powell, D., and Roy, M. (2012). Safety trigger conditions for critical autonomous systems. In *2012 IEEE 18th Pacific Rim International Symposium on Dependable Computing*, pages 61–69. IEEE.
- Metzen, J. H., Hutmacher, R., Hua, N. G., Boreiko, V., and Zhang, D. (2023). Identification of systematic errors of image classifiers on rare subgroups. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5064–5073.
- Richter, S. R., Hayder, Z., and Koltun, V. (2017). Playing for benchmarks. In *Proceedings of the IEEE international conference on computer vision*, pages 2213–2222.
- Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243.
- Schneider, D. G. and Stemmer, M. R. (2023). Synthetic data generation on dynamic industrial environment for object detection, tracking, and segmentation cnns. In *Doctoral Conference on Computing, Electrical and Industrial Systems*, pages 135–146. Springer.
- Settles, B. (2009). Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison.
- Song, Z., He, Z., Li, X., Ma, Q., Ming, R., Mao, Z., Pei, H., Peng, L., Hu, J., Yao, D., et al. (2023). Synthetic datasets for autonomous driving: A survey. *IEEE Transactions on Intelligent Vehicles*.
- Soviany, P., Ionescu, R. T., Rota, P., and Sebe, N. (2022). Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565.
- Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Bochoon, S., and Birchfield, S. (2018). Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 969–977.
- Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., and Wang, F.-Y. (2017). Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, 4(4):588–598.
- Wang, X., Chen, Y., and Zhu, W. (2021). A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576.
- Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A., Keutzer, K., and Gong, B. (2019). Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2100–2110.
- Zhang, L., Rao, A., and Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.
- Zhang, R., Albrecht, A., Kausch, J., Putzer, H. J., Geipel, T., and Halady, P. (2021). Dde process: A requirements engineering approach for machine learning in automated driving. In *2021 IEEE 29th International Requirements Engineering Conference (RE)*, pages 269–279. IEEE.
- Zhu, X., Bilal, T., Mårtensson, P., Hanson, L., Björkman, M., and Maki, A. (2023). Towards sim-to-real industrial parts classification with synthetic dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4454–4463.