

Sensitivity Analysis of Deep-Learning Perception Models to Sensing Parameters

Anne Sielemann*, Niklas Bitzer*, Masoud Roschani*, Jens Ziehn*, Juergen Beyerer*[†]

*Fraunhofer IOSB, Fraunhoferstr. 1, 76131 Karlsruhe, Germany

Email: {anne.sielemann, niklas.bitzer, masoud.roschani, jens.ziehn}@iosb.fraunhofer.de

[†]Karlsruhe Institute of Technology (KIT), Vision and Fusion Laboratory (IES), 76131 Karlsruhe, Germany

Abstract—In the study of deep learning-based models and deep neural networks (DNN), measuring their sensitivity or robustness against perturbations has been a key focus. This is especially relevant for safety-critical applications, such as fully-automated vehicles on public roads, which need to demonstrate—besides robustness against intentional adversarial manipulations—the robustness under all conditions of their operational design domain (ODD). In this context, the specific understanding of how deep learning models are affected by ODD aspects is crucial to assure safe and reliable operation.

We present an approach to systematically evaluate the sensitivity of DNNs to variations in environmental conditions through the use of physically-based synthetic data, demonstrated on the task of traffic sign recognition. For this task, a data synthesis pipeline is leveraged, for which the comparability to real-world data was previously demonstrated. This pipeline is used to create a large-scale systematic dataset of perturbations with physical, exposure-related meaning, including 2196 subdatasets, accumulating to 9442800 test images. This dataset is provided as open data to serve as a basis for further in-depth investigations into the performance of DNNs under challenging conditions.

The application of our testing approach indicates that the performance of DNNs can be related to the physical parameters of input image data and that the sensitivity to these parameters can be measured to a significant and reproducible degree; however, the results show high variation and complexity in the behavior of the evaluated DNNs that extend beyond simple data quality models.

Index Terms—computer vision, deep learning, sensitivity analysis, robustness, automated driving, traffic sign recognition, simulation, synthetic data

I. INTRODUCTION

WITHIN the development of artificial intelligence (AI) and machine learning (ML) systems for critical applications (such as automated driving, robotics, healthcare, and security), the prediction of their performance during operation, and the early identification of system limits are pivotal challenges. Due to the complexity of advanced AI-based systems and their operational environments, their robustness against variations in the environment and their sen-

Download Synset Signset Sensitivity Analysis Datasets:
synset.de/datasets/synset-signset-ger/sensitivity-analysis/

This work was funded in part by the German Federal Ministry for Economic Affairs and Energy (BMWE) within the program “New Vehicle and System Technologies” as part of the AVEAS research project (www.aveas.org), and conducted within the scope of the KAMO: Karlsruhe Mobility High Performance Center (www.kamo.one).

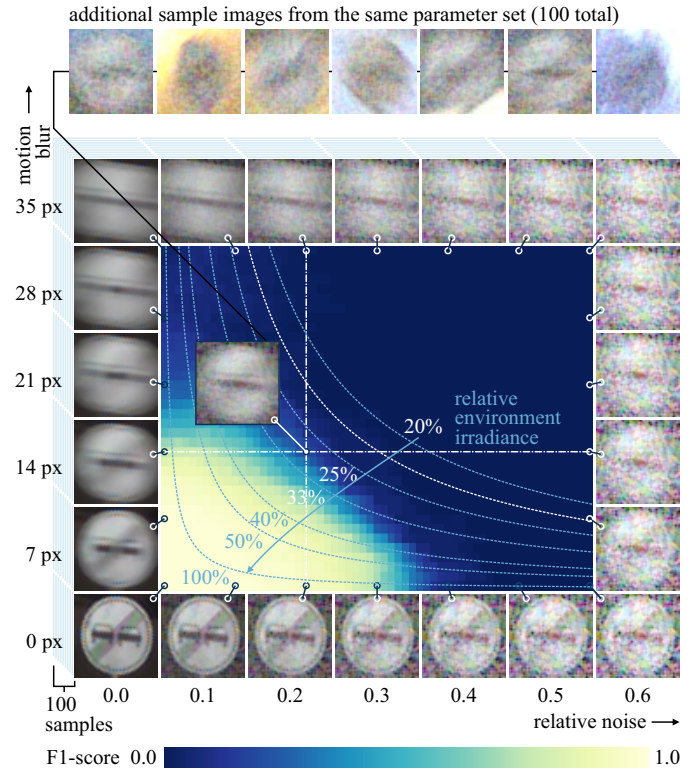


Fig. 1: Example of the applied approach, based on the F1-score of a ConvNext V2 Atto model [1] (trained on GTSRB [2]) on a dataset of physically-based perturbations of a specific traffic sign class. Each color-mapped cell represents the F1-score over 100 sample images. The F1-score generally decreases as the amount of light decreases. But compensation via longer exposure times (which scales motion blur) or via gain or ISO (which scales noise) along an isoline of constant environment brightness (irradiance) has no uniform effect on the F1-score. A systematic evaluation indicates a considerable sensitivity to specific combinations of these camera parameters, to the traffic sign shape, and to the model architecture. The indicated sample in the center is (along with the additional samples at the top) an example of a parameter combination at which the F1-score drops below 0.3. An arbitrary reference irradiance was defined as 100% for the identification of isolines.

sitivity to potential variables are no longer easily described or derived analytically from the system design. Accordingly, the performance evaluation of such data-driven models also depends on using data to evaluate system limits, as well as on the sensitivity of the system to approach said limits due to variations in operational conditions. A particularly important regulation on requirements for the operation of AI/ML-based

systems is the European AI Act, proposed in 2021, which has been current law in the European Union since August 2024, and which states, with regard to the robustness of AI in high-risk applications, i.e., that “Technical robustness is a key requirement for high-risk AI systems” [3, Rec. 75] and that “High-risk AI systems shall be as resilient as possible regarding errors, faults, or inconsistencies that may occur within the system or the environment in which the system operates [...]. Technical and organizational measures shall be taken in this regard.” [3, Art. 15.4] In this context, “high-risk AI systems” refers to, amongst others, “AI systems intended to be used as safety components in the management and operation of critical digital infrastructure [and] road traffic” [3, Anx. III]. It also introduces requirements for technical documentation, which is to include “the validation and testing procedures used, including information about the validation and testing data used and their main characteristics; metrics used to measure accuracy, robustness, and compliance with other relevant requirements” [3, Anx. IV]. Accordingly, the demand increases to map aspects of the expected operational conditions of AI systems (their *operational design domain* (ODD) [4]) to quantitative performance metrics.

A consistent challenge in implementing such a predictive robustness statement is the high degree of variation that can arise in the future ODD and the lack of conclusive data that enables the quantitative prediction of AI module / subsystem performance (for example, accuracy, recall, mIoU, etc.), and from this, the prediction of AI system performance (for example, safety, efficiency, availability, etc.). Various approaches have been proposed to improve the understanding of system performance (see Sec. II for examples), prominently featuring methods to augment datasets of samples under benign conditions by introducing artificial degradations. This has the advantage of leveraging high-quality annotations under robust conditions (both manually-generated and automated) to create samples of challenging conditions. One limitation is that these degradations cannot directly be linked to properties of the target ODD; another is that the effect of ODD conditions on design parameters such as the choice of specific DNN models, the training dataset, and technical system parameters is still not well understood. It is clear that inferior sensor quality, challenging environmental conditions, and generally challenging AI tasks degrade AI performance; yet how this degradation manifests and how knowledge about this can be used directly to adjust system parameters, hyperparameters, or ODD definitions is not yet conclusively established.

We introduce an approach to systematically scan the performance of DNN-based perception models through physically-based simulation in the OCTAS[®] [5] framework that directly links performance to defined combinations of system parameters and environmental conditions. Instead of augmenting “clean” datasets by introducing artificial degenerations, our approach introduces physical phenomena in an analytic, model based process that enables the consistent tracking of physical parameters of the degradations. We demonstrate the approach on the example of traffic sign recognition, as this represents

a well-understood and established task that provides a wide range of publicly available datasets and applicable models. It also represents a task for which a data synthesis pipeline exists [6], for which the adequacy of replacing real data with physically-based synthetic data was demonstrated through the conduction of cross-dataset evaluation. We utilize this pipeline to generate a synthetic dataset of traffic signs with highly-resolved degradations (61×36 subdatasets) in camera parameters (as shown in Fig. 1), comprising 9 442 800 images (plus additional metadata) across 43 traffic sign classes from the *German Traffic Sign Recognition Benchmark* (GTSRB) dataset [2]. Each combination of class and camera parameters was covered with 100 samples of the same traffic sign type under different viewing angles, environmental conditions, and surface variations, enabling to trace the degradation of recognition performance.

In Sec. II we introduce the existing state of the art, based on which we establish the generation of the dataset for the sensitivity analysis in Sec. III. The generated dataset is applied in Sec. IV to yield sensitivity curves across different DNN architectures and training datasets. The results show that DNN performance under challenging perception conditions varies based on the DNN model architecture, as well as primarily on the training dataset used, and that these parameters are strongly correlated in a way that is occasionally counter-intuitive. They show that physically-based degradations do not have a homogeneous or monotonic effect on DNN performance, which is why they cannot simply be inferred. Furthermore, the performance curves of DNN models across degradations exhibit highly different and complex structures that—when known—can be used to specify requirements on the system and the ODD. In Sec. V, we draw conclusions from the study and provide an outlook on how the principle may be leveraged to establish a closer link, across different domains and tasks, between the parameters of the ODD and the performance of AI at the model or subsystem level.

II. STATE OF THE ART

The robustness—or conversely the sensitivity—of analytic deep learning models w.r.t. objective parameters of the subject of analysis, both deliberate adversarial modifications and incidental (non-adversarial) perturbations, has been a long-standing research topic. In [7], a systematic sensitivity evaluation is performed by defining a perturbation manifold from which variations of datasets and network parameters are generated in various classification applications. An overview of different robustness evaluation approaches is provided in [8], which distinguishes between environmental effects (i.e., lighting and weather), rendering (i.e., filters, compression, drawing style), and sensor (i.e., motion blur, noise), stating, “Of the studies which examined multiple corruption categories (31), noise (20) and blur (5) sensor (S) corruptions were found to have the strongest impact on performance.” With DIN SPEC 91527:2025-12 [9], a PAS standard was published that provides a framework for runtime-monitoring of AI systems according to different metrics and reacting to a detected lack

of robustness during operation. In [10], the robustness against different types of anonymization is evaluated, representing a third type of perturbation besides adversarial and incidental. In [11], the sensitivity of models against variations in hyperparameters is evaluated, while [12] evaluates the sensitivity of learning quality with respect to continual learning paradigms, and [13] uses parameter variation to derive model sensitivities in classical ML behavior modeling. The effect of artifacts introduced by image compression and virtualization is assessed in [14]. In [15], the impact of motion blur on camera data is evaluated through post-processing images, and mitigation methods are derived; similarly, [16] evaluates the effect of motion blur in robotic inspection. In [17], a systematic analysis of motion blur for image analysis is conducted by post-processing images through adversarial kernels, both on simulated and real imagery. In [18], simulated data is used in conjunction with post-processing over realistic and non-realistic perturbations on image and depth data to benchmark the robustness of SLAM methods. Additional applications of synthetic data for computer vision and perception tasks can be found in [19], [20], [21]. Corresponding datasets are based directly on computer games [22], [23], [24] or computer game engines (such as the Unity [25] or the Unreal engine) [26], [27], [28], [29], [30], [31], [32], including data simulated in the Carla Simulator [33].

The specific use case of traffic sign recognition has repeatedly been considered a benchmark for ML applications, partly due to the clear scope and wide availability of datasets and reference models, with [34] using this task (among others) as a demonstration of deep learning-based methods outperforming humans in a vision task. Specific approaches towards this task through synthetic data include real data modified through augmentation and recombination (e.g., [35], [36], [37], [38]), or the use of generative AI to create sensor data, such as [39], [40] for the specific task of traffic sign recognition. In [41], a framework was presented for the systematic synthesis of sensor data based on physically-based rendering, which was applied in [6] to generate a traffic sign recognition dataset that achieved training results outperforming the use of real data. In [42], this approach was used to evaluate the interpretation of explainable AI (XAI) approaches for deep learning attention on the same task.

III. SYNTHETIC TEST DATASETS

As outlined in the introduction, the focus of this paper is on the pipeline to generate arrays of datasets with finely-resolved, physically-based imaging artifacts to measure their effect onto the performance of different model architectures, training parameters, and subtasks. The purpose of the synthetic test datasets is to systematically cover the task of traffic sign recognition over incremental degradations that preserve a physical meaning w.r.t. properties of the ODD. As a basis, we choose the synthesis pipeline presented in [6], where the *Synset Signset Germany* (SSG) dataset, which represents the “virtual twin” of the GTSRB dataset [2], [43], was created, and where a training / testing comparison showed that the synthetic

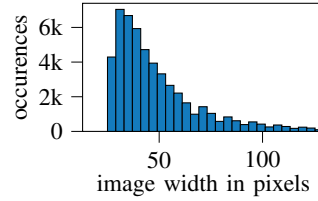


Fig. 2: Distribution of image widths in GTSRB [2] (cropped to max 130px).



Fig. 3: Examples of provided mask and segmentation images.

data can adequately replace real-world data within the given scope. The synthesis pipeline uses physically-based rendering and procedural, stochastically-based scenario sampling in the OCTAS[®] framework to generate pictures of traffic signs with realistic variations in viewing angles and environments, as well as generative wear and tear on the traffic sign surface based on real-world data. The pipeline was already used in [42] to explore XAI approaches w.r.t. learning performance of DNNs.

In the physically-based framework, degradations must be modeled based on environmental and sensor properties rather than through post-processing effects; accordingly, a model is required that represents variations in the ODD and in sensor properties. To demonstrate the approach, we chose to relate exposure parameters to DNN performance, namely two parameters to compensate for low-light situations: Longer exposure times (introducing motion blur) and stronger camera gain or higher ISO, increasing sensitivity to light at the expense of increased noise levels.

On the other hand, for meaningful and transferable results, it is necessary to parameterize these effects in a way that maps readily to various practical applications rather than just a single one. A wide range of related parameters affect camera exposure: Different noise models are introduced from different physical effects; exposure times introduce motion blur not directly, but indirectly through angular or spatial motion of the camera, object distance, focal length, and sensor size; and all of these effects can again be varied by a third exposure parameter, the aperture, which controls the incoming light. The aperture introduces another complex set of parameters, namely depth of field or defocus blur for wide apertures and the introduction of diffraction patterns for narrow apertures, both of which again depend not only on the numerical aperture value but also on the geometric shape of the aperture diaphragm. Additionally, several other effects, such as chromatic aberrations and geometric variations [6], will also occur in practice that have a non-negligible effect on DNN performance, as demonstrated in [42].

A. Dataset Generation

To disentangle this complexity and establish transferable, generic metrics for degeneration effects while maintaining transparent model assumptions and physical meaning, the following assumptions were taken on basis of the data generation procedure in [6].

The most fundamental simplification is the limitation to a *single image resolution* of 48×48 pixels. This limitation

is based on the fact that effects such as noise and motion blur are related to image size (in pixels and in field of view), such that the effect of constant noise or motion blur values within a given pixel neighborhood can vary depending on whether the overall image resolution is high or low. In turn, DNNs that take smaller input resolutions are rather unaffected by moderate noise levels introduced into high-resolution images. Additionally, very high resolutions may be unrealistic in practical applications (typically traffic sign detection followed by recognition), while very low resolutions preclude reasonable DNN performance. Hence, to eliminate these complexities, all images were rendered at a resolution of 48×48 , which represents the median sample width of 43 pixels in the reference dataset GTSRB [2] (cf. Fig. 2), rounded to a multiple of 8 to improve rendering performance. Images are generated such that this size represents raw sensor size, including, for example, Bayer pattern artifacts and noise.

Concerning the representation of *motion blur*, we limit our consideration to the recognition of traffic signs at relatively long distances (above around 50 m or 150 ft.) and corresponding narrow, forward fields of view, where the perspective translation of the camera w.r.t. the traffic sign during the exposure time is negligible. In this case, motion blur can, with acceptable approximation, be represented as a purely linear motion between two angles, modeled as an angular shift due to pitch and yaw changes of the camera with respect to the sign. To establish a parameterization that does not rely on focal length, distances, camera motion models, etc., we measure motion blur in pixels of image size (see above), modeled as linear motion at a uniformly-sampled angle $\beta \sim \mathcal{U}(-\pi, \pi)$.

Regarding the representation of *image noise*, we again solely consider additive noise whose scale is measured w.r.t. the overall image intensity. This normalizes all image intensities (regardless of their scene environment, such as indoor parking, outdoor, day, night, ...) to a common scale, but removes the need to specify camera sensitivities and electrical parameters.

All other effects and artifacts applied for the generation of SSG [6] were either disabled or ensured to be parameterized to a constant, realistic, neutral level: The traffic sign pole is configured to permanently remain in a vertical orientation. Additional traffic signs on the traffic sign pole and the tree geometry for occlusions and shadow casting are deactivated. We chose to apply a moderate variation of camera viewing angles sampled from normal distributions $\mathcal{N}(\mu, \sigma^2)$ of roll $\sim \mathcal{N}(0.0^\circ, 9.0^\circ)$, pitch $\sim \mathcal{N}(0.0^\circ, 64.0^\circ)$, and yaw $\sim \mathcal{N}(0.0^\circ, 256.0^\circ)$. The template image generation, Pix2Pix-based [44] texture and defect synthesis, the traffic sign material, traffic sign pole diameter, and pole material variation remain unchanged. The environment is likewise represented as in [6] using image-based lighting (IBL). For this purpose, 523 uniformly sampled environment maps from Polyhaven [45] were utilized. Concerning the image artifact configuration, lens flares and chromatic aberrations were excluded. Limited digital image sharpening and Bayer pattern artifacts are present. The point spread functions (PSF) are constant and based on the

original parameterization in [6]. Internally, we improved the white balancing and AEC error calculation from [6].

For rendering the images, OCTAS[®] currently provides APIs to the rasterization-based engine *OGRE3D* [46] and the path-tracing engine *Cycles* [47]. However, as the evaluation in [6] demonstrated that the generated data from both methods resulted in comparable outcomes, we opted to use only the less computationally demanding OGRE3D engine for this work.

The result is shown in Fig. 1, where the modeling of motion blur and image noise allows to define isolines of environmental lighting conditions or effective *irradiance*, such that any image sampled along the line would yield the same brightness according to automatic exposure control (AEC); yet each point along the line represents a different permissible tradeoff between exposure time and gain. This perspective can be mapped to any practical application where relative noise scale and motion blur can be measured or estimated from system models.

B. Dataset Properties

The resulting test dataset comprises 2 196 individual sub-datasets, generated by combining 36 levels of motion blur $\times$ 61 levels of additive noise. For motion blur, the dataset covers pixel-wise degradations from 0 to 35 pixels; for the additive image noise, the subdatasets were generated using a scaling factor that increases in increments of 0.01, ranging from 0.0 to 0.6. Each of the subdatasets includes the 43 traffic sign classes of GTSRB [2], with 100 images per class, which sum up to 9 442 800 images overall. Together with the RGB images, we provide mask and segmentation images (cf. Fig. 3), as well as image-wise json metadata files containing the stochastically selected environment and imaging effect parameters.

IV. EVALUATION

A. Evaluated Network Architectures and Training Setup

For the evaluation, we employ two state-of-the-art image classification architectures: ConvNeXt V2 [1], a pure CNN, and MaxViT [48], a CNN-ViT hybrid model. To gain insight into the potential influence of network capacity, we train ConvNeXt V2 at model scales Base (89M parameters), Nano (15.6M), and Atto (3.7M), alongside MaxViT Tiny (31M). Training follows the official ConvNeXt V2 fine-tuning pipeline with minor modifications. Each model is initialized with ImageNet-1k [49] pre-trained weights and fine-tuned separately on GTSRB [2] and the corresponding 43-class subset of SSG [6] for 100 epochs at an input resolution of 224×224 with a batch size of 128. For the ConvNeXt V2 models, the first two stages remain frozen. We use the AdamW optimizer with an initial learning rate of 10^{-3} , cosine annealing to 10^{-6} , a linear warmup over 10 epochs, and a weight decay of 10^{-2} . Regularization includes label smoothing (0.1), drop path (0.1), random erasing, and RandAugment [50]. Since traffic signs are sensitive to orientation, geometric augmentations (e.g., horizontal/vertical flipping) are disabled. To determine the influence of the random seed, we conduct eight additional training runs for ConvNeXt V2 Atto with differing seeds.

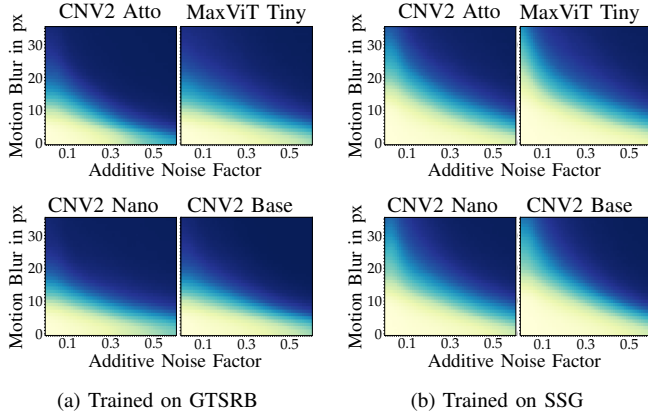


Fig. 4: Comparison of the overall F1-scores achieved per network architecture (upper row) and for additional ConvNeXt V2 (CNV2) [1] sizes (lower row) evaluated on the presented synthetic test dataset.

B. Evaluation Results

All result visualizations presented in this section follow the representation and coloration in Fig. 1. The first experiment evaluates the overall F1-scores measured on the presented synthetic test dataset achieved by the network architectures under consideration: ConvNeXt V2 (Atto) [1] and MaxViT (Tiny) [48], which were respectively trained on GTSRB [2] and SSG [6]. The findings shown in Fig. 4 (upper row) indicate that, overall, the networks trained on GTSRB exhibit greater sensitivity to additive noise, and particularly to motion blur. This is presumably because, on the one hand, the entire GTSRB dataset was recorded using a single camera setup (mounted on a mobile platform) [43], which likely results in less variety in motion blur and additive noise compared to SSG, where different exposure and noise parameters were explicitly simulated [6]. On the other hand, it should be considered that the domain gap between SSG and the generated test dataset is probably smaller than that to GTSRB since both originate from the same simulation source, albeit with differing parameterizations.

Furthermore, comparing the network architectures reveals that the variations between architectures are smaller than the variations observed when training on different datasets. The areas with rather high F1-scores (yellow) have nearly similar dimensions and shapes for both architectures. This observation is also supported by the evaluations of different ConvNeXt V2 sizes; the results for Nano and Base are complemented in the lower row of Fig. 4. The slight variation across model sizes suggests that the potential of the larger models additional parameters is not fully utilized. This is supported by the nearly identical validation accuracies achieved across the tested sizes (e.g., F1 on SSG for ConvNeXt V2 Atto: 0.994; Nano: 0.991; Base: 0.995) and hints at a mismatch between model capacity and task complexity for models of these sizes. However, the comparison of high F1-score area shapes between the training datasets reveals differences: when training on GTSRB, these areas are shaped triangularly across the architectures, while they possess a slightly concave shape due to further expansion

in the near-axis regions when training on SSG. This shows not only that training on both datasets leads to different sensitivities to sensing parameters despite very similar performance results on common test splits (cf. [6]), but also that the sensitivity to the combination of parameters cannot accurately be inferred from the analysis of the sensitivity to the individual sensing parameters.

In addition to the overall results, we tracked the class-wise F1-scores, which are reported exemplarily for ConvNeXt V2 Atto in Fig. 5. On the instance level, the results reveal that DNNs react with varying sensitivity to the degradation of the sensing conditions of individual classes. This is particularly evident when comparing different logical traffic sign groups with one another (cf. Fig. 6): the investigated DNNs are comparably robust in the classification of circular blue regulatory signs under noisy and (motion) blurry conditions. The robustness of speed limit recognition achieves medium robustness in both aspects, while the DNNs are most sensitive to the classification of danger signs under degraded sensing conditions. The comparison of the per-class results of the two network instances shows that, as assumed, the areas with relatively high F1-scores (yellow) for the individual classes tend to exhibit the dataset-typical shape. However, the robustness distribution across the classes is comparatively similar (e.g., for the danger sign of bicycles, both networks react most sensitively), which indicates that the DNNs are capable of learning similar features from both datasets and underscores the practical usefulness of SSG.

To be capable of estimating the result variability due to the random seed used during training, we evaluated eight additional training runs for ConvNeXt V2 Atto [1] on GTSRB [2] and SSG [6]. The performance results measured on the corresponding test splits (GTSRB: $F1_{\text{weighted}} = 0.998 \pm 0.0004$, SSG: $F1 = 0.993 \pm 0.0016$) exhibit a small standard deviation, and the resulting F1-score matrices do not significantly deviate from the matrices reported in Fig. 4, as demonstrated by the standard deviation plot over all seeds in Fig. 7. Nevertheless, we observed, on the one hand, that single seeds could result in notably less robustness (cf. Fig. 7), and on the other hand, a high variation in class collapsing behavior between different architecture types, even when training the same architecture with different seeds (cf. Fig. 8). This finding highlights the need to accurately define and test the ODD for safety-critical applications; otherwise, random effects might occur.

Overall, the proposed testing approach on the presented dataset provides valuable additional insights into the sensitivity of DNNs to sensing parameters. However, when using the test dataset for the sensitivity analysis of custom DNNs, it should be considered that metrics such as precision or recall should NOT be used in isolation. There are single classes where the precision and recall results both serve as good performance indicators (cf. Fig. 9 A). But the sensitivity analysis aims to systematically test system limits, where both metrics alone could lead to incorrect interpretations. The precision metric, defined as $\text{precision} = \frac{TP}{TP+FP}$, can produce a step-like behavior (cf. Fig. 9 B) when the respective class is rarely predicted

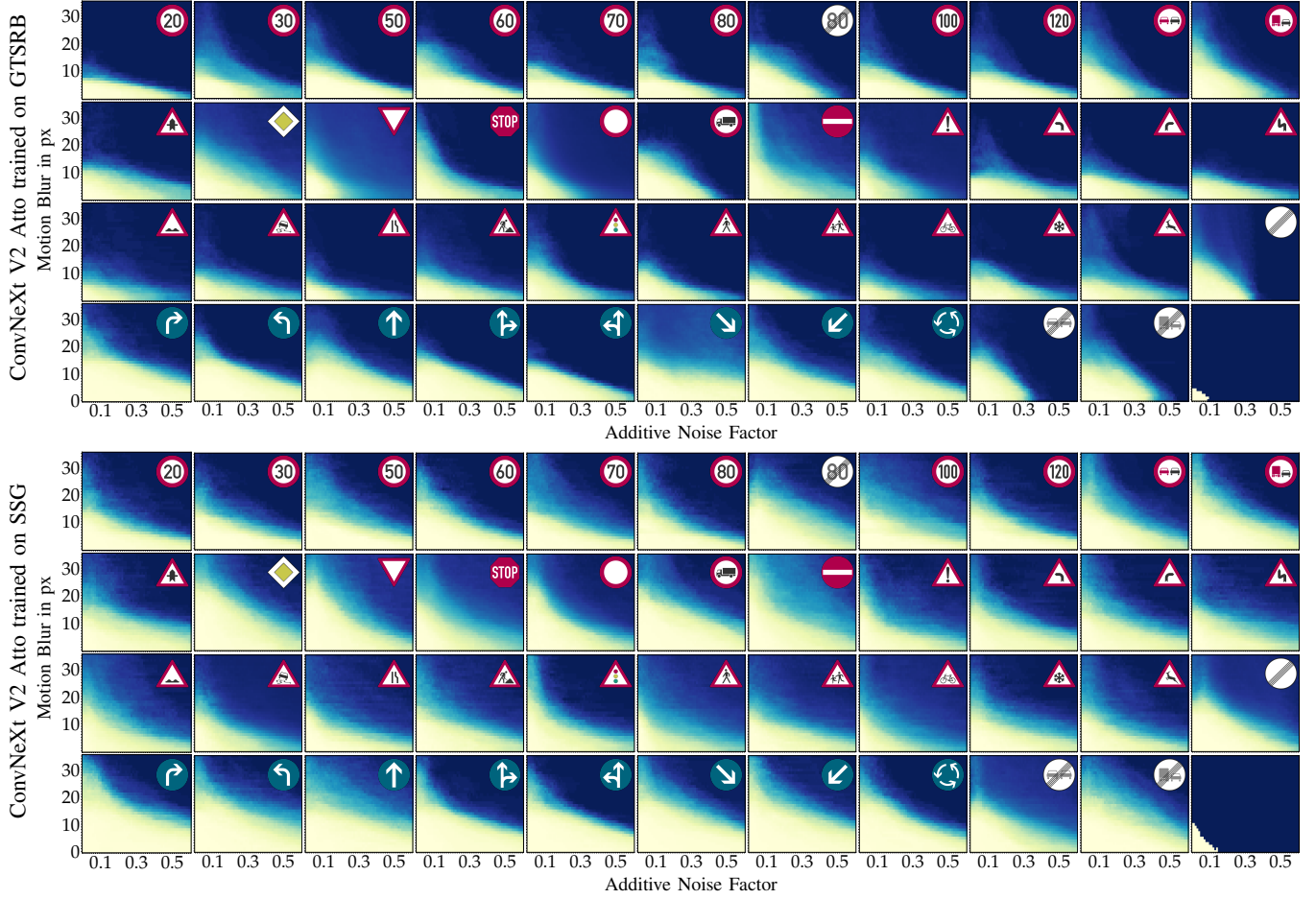


Fig. 5: Class-wise evaluation of the F1-scores obtained by the ConvNeXt V2 Atto [1] instance trained on the GTSRB dataset [2] (top) and Synset Signset Germany [6] (bottom). The threshold matrix in the respectively lower right corner reports the intersection of all regions in which the individual classes achieved an F1-score of over 0.9 (pale yellow, dark blue otherwise).

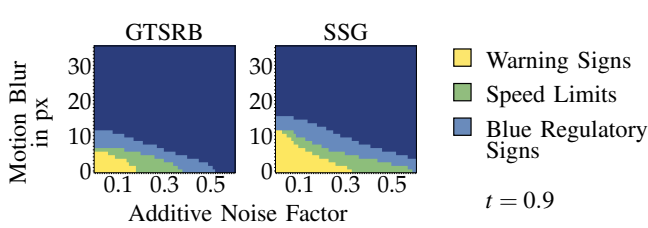


Fig. 6: Visualization of robustness differences regarding to logical traffic sign groups. Reported are the intersections of the regions where all of the traffic signs belonging to the respective group achieved a F1-score of over 0.9. Note that the threshold areas are plotted on top of each other.

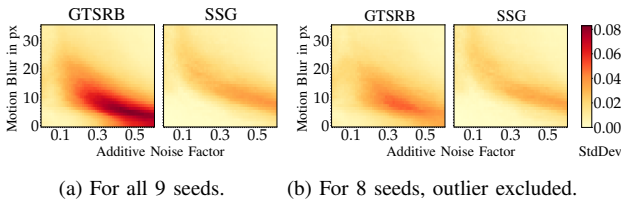


Fig. 7: Standard deviations of the overall F1-scores achieved with different training seeds for ConvNeXt V2 Atto [1]. Since the resulting network of training on GTSRB with one specific random seed (67) is considerably less robust to noise, we reported the results (a) with and (b) without this outlier.

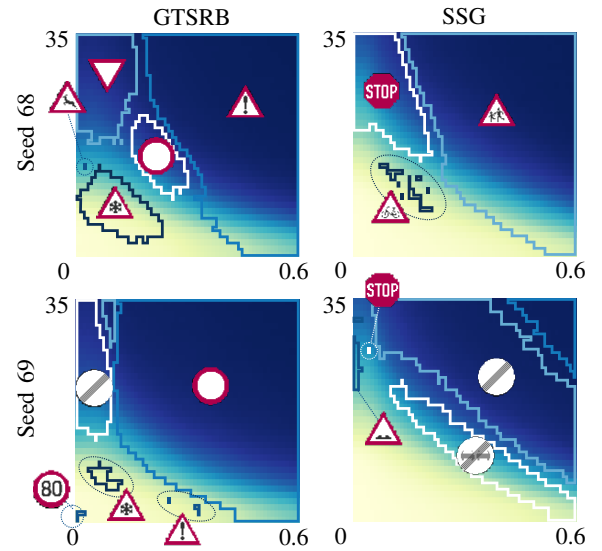


Fig. 8: Overview of the variation of class collapsing behavior when training ConvNeXt V2 Atto [1] with different random seeds. Thereby, we define a matrix cell as collapsed to a particular class if that class is predicted as a false positive (FP) at least 50 times more often than the class with the second-highest number of FP predictions. Where no region is explicitly marked, no such distinct class exists.

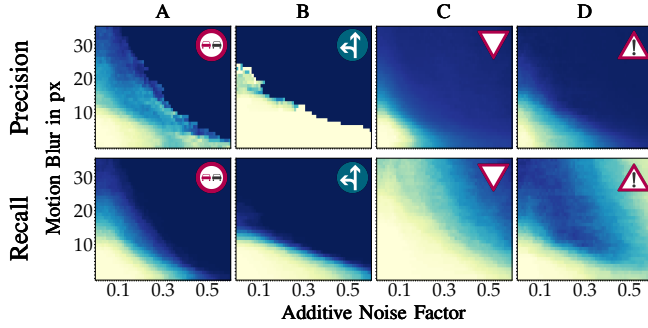


Fig. 9: Single class precision and recall examples accomplished by the ConvNeXt V2 Atto trained on GTSRB [2] and tested on the proposed testing dataset. While for A, both metrics seem to be good standalone performance indicators, for precision the example B and for recall the examples C and D show that both metrics can be misleading when considered isolated.

in combination with nearly no FPs. Consequently, it indicates good performance in the border areas, where performance degrades according to recall. Conversely, the recall metric, defined as $\text{recall} = \text{TP} / (\text{TP} + \text{FN})$, indicates good performance for classes to which the network collapses (cf. Fig. 9 C D). In addition to more comprehensive metrics such as the F1-score, which combines both metrics, considering precision and recall individually can lead to extra information and thus a better understanding.

V. CONCLUSION AND OUTLOOK

In this work, we presented an approach to the sensitivity analysis of DNNs to sensing parameters through synthetic data, performed exemplarily on the task of traffic sign recognition. This includes the following contributions:

(I) The development of a systematic testing approach through the use of physically-based synthetic data, which is applicable to a wide variety of DL perception tasks.

(II) A publicly available, physically-based synthetic testing dataset for the task of traffic sign recognition, following the proposed testing approach and including 2 196 subdatasets with 43 classes each, summing up to 9 442 800 samples, plus additional metadata. The dataset can serve as a basis for further in-depth investigation into the performance of DNNs under challenging conditions.

(III) Exemplary testing results for the DNN architectures ConvNeXt V2 (Atto, Nano, Base) [1] and MaxViT (Tiny) [48], respectively trained on GTSRB [2] and SSG [6]. Through the linkage of the achieved performance to defined combinations of system parameters and environmental conditions, operators of the same architecture and training dataset combinations can use the reported results to either facilitate their estimation of the ODD or to choose and parameterize their camera system.

Our results demonstrate that the properties of the training dataset have a stronger impact on a DNN’s sensitivity to sensing parameters than the specific network architecture. Furthermore, training on different datasets can lead to differing sensitivity patterns with respect to sensing parameters, even when the overall performance metrics measured on common validation datasets appear nearly similar. Moreover, the sensitivity to a combination of sensing parameters cannot

reliably be inferred from the sensitivities to individual parameters, highlighting the need for dedicated analyzes of (sensing) parameter interactions. In addition, our results underline the meaningfulness of analyzing class-wise performances for safety-critical applications: the influence of sensing parameters can vary substantially across classes or groups of classes and is not adequately captured by averaging or aggregating class-wise metrics alone. Finally, our results show that the selected random seed strongly affects which classes the DNN collapses to in edge cases, emphasizing the need to investigate and understand the scope of the ODD to prevent random effects and ensure safe and reliable operation.

Outlook

Overall, the proposed testing approach has demonstrated, in the context of the traffic sign recognition task, that it can yield meaningful insights into the intricate behavior of the investigated DNNs—insights that cannot be achieved by merely evaluating on validation datasets or relying on simple quality measures. This underlines the usefulness of synthetic data for testing deep learning-based applications and renders the approach interesting for further tasks. As pointed out in the study, it is pivotal to base the testing approach on physically-based synthetic data to be capable of linking the results to the ODD. Furthermore, it is urgently necessary to examine the comparability of real data to the synthetic data before using them as test data, which we have done for our data in [6]. The results of the presented analysis showed that metrics such as precision and recall are not good performance indicators in isolation; instead, more comprehensive metrics, such as (for classification tasks) the F1-score, should be used. Additionally, we assume that, for more complex tasks, the choice of network architecture influences a DNN’s sensitivity to sensing parameters more strongly than indicated by our analysis. For new tasks, and especially for more complex tasks, this should be further investigated.

REFERENCES

- [1] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, “Convnext v2: Co-designing and scaling convnets with masked autoencoders,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 16 133–16 142.
- [2] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, “The German traffic sign recognition benchmark: a multi-class classification competition,” in *The 2011 international joint conference on neural networks*. IEEE, 2011, pp. 1453–1460.
- [3] European Commission, “AI Act: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act),” Jul. 2024.
- [4] —, “Commission Implementing Regulation (EU) 2022/1426 of 5 August 2022 laying down rules for the application of Regulation (EU) 2019/2144 of the European Parliament and of the Council as regards uniform procedures and technical specifications for the type-approval of the automated driving system (ADS) of fully automated vehicles,” Aug. 2022.
- [5] <https://octas.org/>.

- [6] A. Sielemann, L. Loercher, M.-L. Schumacher, S. Wolf, M. Roschani, J. Ziehn, and J. Beyerer, "Synset Signset Germany: A Synthetic Dataset for German Traffic Sign Recognition," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, 2024.
- [7] H. Shu and H. Zhu, "Sensitivity analysis of deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4943–4950.
- [8] N. Drenkow, N. Sani, I. Shpitser, and M. Unberath, "A systematic review of robustness in deep learning for computer vision: Mind the gap?" *arXiv preprint arXiv:2112.00639*, 2021.
- [9] D. Weimer, A. Meyer-Vitali, A. Sielemann, J. Ziehn, J. Zhou, Y. Bulut, M. Graf, and S. Krauß, "DIN SPEC 91527:2025-12, Goals, Methods and Metrics for Automated/Semi-Automated Runtime Monitoring of AI Systems for Non-Adversarial Performance Degradations," Dec. 2025.
- [10] J. Zhou and J. Beyerer, "Impacts of data anonymization on semantic segmentation," in *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2022, pp. 997–1004.
- [11] R. Taylor, V. Ojha, I. Martino, and G. Nicosia, "Sensitivity analysis for deep learning: ranking hyper-parameter influence," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2021, pp. 512–516.
- [12] T. Kalb, M. Roschani, M. Ruf, and J. Beyerer, "Continual learning for class- and domain-incremental semantic segmentation," in *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2021, pp. 1345–1351.
- [13] N. Neis and J. Beyerer, "Sensitivity analysis and extended evaluation of the two-level stochastic model for the lateral movement of vehicles within their lane," in *2024 10th International Conference on Automation, Robotics and Applications (ICARA)*. IEEE, 2024, pp. 1–7.
- [14] C. Steinhäuser, P. Reis, H. Padusinski, J. Lagner, and E. Sax, "Data quality matters: Quantifying image quality impact on machine learning performance," in *2025 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2025, pp. 1564–1570.
- [15] A. A. Al-Dulaimi, M. T. Guneser, A. A. Hameed, F. P. G. Márquez, N. L. Fitriyani, and M. Syafrudin, "Performance analysis of classification and detection for pv panel motion blur images based on deblurring and deep learning techniques," *Sustainability*, vol. 15, no. 2, p. 1150, 2023.
- [16] N. Jose, G. Marchello, and F. Cannella, "Motion blur impact on measurement in autonomous inspection," in *2024 20th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)*. IEEE, 2024, pp. 1–7.
- [17] Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, J. Wang, B. Yu, W. Feng, and Y. Liu, "Watch out! Motion is blurring the vision of your deep neural networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 975–985, 2020.
- [18] X. Xu, T. Zhang, S. Wang, X. Li, Y. Chen, Y. Li, B. Raj, M. Johnson-Roberson, and X. Huang, "From perfect to noisy world simulation: Customizable embodied multi-modal perturbations for slam robustness benchmarking," *arXiv preprint arXiv:2406.16850*, 2024.
- [19] S. I. Nikolenko, "Synthetic data for deep learning," 2019. [Online]. Available: <https://arxiv.org/abs/1909.11512>
- [20] C. M. de Melo, A. Torralba, L. Guibas, J. DiCarlo, R. Chellappa, and J. Hodgins, "Next-generation deep learning based on simulators and synthetic data," *Trends in cognitive sciences*, vol. 26, no. 2, pp. 174–187, 2022.
- [21] S. Jain, G. Seth, A. Paruthi, U. Soni, and G. Kumar, "Synthetic data augmentation for surface defect detection and classification using deep learning," *Journal of Intelligent Manufacturing*, pp. 1–14, 2022.
- [22] T. Golda, A. Blattmann, J. Metzler, and J. Beyerer, "Image domain adaption of simulated data for human pose estimation," in *Artificial intelligence and machine learning in defense applications II*, vol. 11543. SPIE, 2020, pp. 112–127.
- [23] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 102–118.
- [24] M. Johnson-Roberson, C. Barto, R. Mehta et al., "Driving in the Matrix: Can Virtual Worlds Replace Human-Generated Annotations for Real World Tasks?" *arXiv preprint arXiv:1610.01983*, 2016, sridhar, Sharath Nittur and Rosaen, Karl and Vasudevan, Ram.
- [25] <https://unity.com>.
- [26] <https://www.unrealengine.com>.
- [27] G. Ros, L. Sellart, J. Materzynska et al., "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 (CVPR)*, Jun. 2016, vazquez, David and Lopez, Antonio M.
- [28] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual Worlds as Proxy for Multi-Object Tracking Analysis," in *CVPR*, 2016.
- [29] Y. Yao, L. Zheng, X. Yang et al., "Simulating content consistent vehicle datasets with attribute descent," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 775–791, naphade, Milind and Gedeon, Tom.
- [30] Y. Cabon, N. Murray, and M. Humenberger, "Virtual KITTI 2," *arXiv preprint arXiv:2001.10773*, 2020.
- [31] T. S. Kim, B. Shim, M. Peven et al., "Learning From Synthetic Vehicles," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, Jan. 2022, pp. 500–508, qiu, Weichao and Yuille, Alan and Hager, Gregory D.
- [32] D. Temel, G. Kwon, M. Prabhushankar, and G. AlRegib, "CURE-TSR: Challenging unreal and real environments for traffic sign recognition," in *Neural Information Processing Systems (NeurIPS) Workshop on Machine Learning for Intelligent Transportation Systems*, 2017.
- [33] <https://carla.org/>.
- [34] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-Column Deep Neural Networks for Image Classification," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012.
- [35] H. K. Ekbatani, O. Pujol, and S. Seguí, "Synthetic data generation for deep learning in counting pedestrians," in *ICPRAM*, 2017, pp. 318–323.
- [36] A. Stergiou, G. Kalliatakis, and C. Chrysoulas, "Traffic sign recognition based on synthesised training data," *Big Data and Cognitive Computing*, vol. 2, no. 3, p. 19, 2018.
- [37] L. Tabelini, R. Berriel, T. M. Paixão, A. F. De Souza, C. Badue, N. Sebe, and T. Oliveira-Santos, "Deep traffic sign detection and recognition without target domain real images," *Machine Vision and Applications*, vol. 33, no. 3, p. 50, 2022.
- [38] V. Shakhuro, B. Faizov, and A. Konushin, "Rare traffic sign recognition using synthetic training data," in *Proceedings of the 3rd International Conference on Video and Image Processing*, 2019, pp. 23–26.
- [39] H. Luo, Q. Kong, and F. Wu, "Traffic sign image synthesis with generative adversarial networks," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2540–2545.
- [40] C. Dewi, R.-C. Chen, Y.-T. Liu, X. Jiang, and K. D. Hartomo, "Yolo V4 for advanced traffic sign recognition with synthetic training data generated by various GAN," *IEEE Access*, vol. 9, 2021.
- [41] A. Sielemann, S. Wolf, M. Roschani, J. Ziehn, and J. Beyerer, "Synset Boulevard: A Synthetic Image Dataset for VMMR," in *2024 International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.
- [42] A. Sielemann, V. Barner, S. Wolf, M. Roschani, J. Ziehn, and J. Beyerer, "Measuring the Effect of Background on Classification and Feature Importance in Deep Learning for AV Perception," in *2025 IEEE International Automated Vehicle Validation Conference (IAVVC)*, 2025.
- [43] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. Computer: Benchmarking Machine Learning Algorithms for Traffic Sign Recognition," *Neural Networks*, vol. 32, pp. 323–332, 2012, selected Papers from IJCNN 2011.
- [44] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
- [45] <https://polyhaven.com>.
- [46] <https://www.ogre3d.org>.
- [47] <https://www.cycles-renderer.org>.
- [48] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," in *European conference on computer vision*. Springer, 2022, pp. 459–479.
- [49] J. Deng, W. Dong et al., "ImageNet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255, socher, Richard and Li, Li-Jia and Li, Kai and Fei-Fei, Li.
- [50] E. D. Cubuk, B. Zoph et al., "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2020, shlens, Jonathon and Le, Quoc V.