

Synset Signset Germany: A Synthetic Dataset for German Traffic Sign Recognition*

Anne Sielemann¹, Lena Loercher², Max-Lion Schumacher², Stefan Wolf^{3,1},
Masoud Roschani¹, Jens Ziehn¹ and Juergen Beyerer^{1,3}

Abstract—In this paper, we present a synthesis pipeline and dataset for training / testing data in the task of traffic sign recognition that combines the advantages of data-driven and analytical modeling: GAN-based texture generation enables data-driven dirt and wear artifacts, rendering unique and realistic traffic sign surfaces, while the analytical scene modulation achieves physically correct lighting and allows detailed parameterization. In particular, the latter opens up applications in the context of explainable AI (XAI) and robustness tests due to the possibility of evaluating the sensitivity to parameter changes, which we demonstrate with experiments. Our resulting synthetic traffic sign recognition dataset Synset Signset Germany contains a total of 105 500 images of 211 different German traffic sign classes, including newly published (2020) and thus comparatively rare traffic signs. In addition to a mask and a segmentation image, we also provide extensive metadata including the stochastically selected environment and imaging effect parameters for each image. We evaluate the degree of realism of Synset Signset Germany on the real-world German Traffic Sign Recognition Benchmark (GTSRB) and in comparison to CATERED, a state-of-the-art synthetic traffic sign recognition dataset.

I. INTRODUCTION

WITHIN the development of machine learning (ML) and artificial intelligence (AI), and with the substantial advances achieved in the performance, particularly of deep learning, the attention of research and development has shifted to include not only maximum performance of ML and AI, but also properties relating to how this performance is achieved—namely concerning the methods for providing data of sufficient quality, recency and practical costs, and understanding the system behavior w.r.t. the real world, in terms of explainable AI (XAI), robustness, and validation.

Download Synset Signset Germany:
synset.de/datasets/synset-signset-ger/

¹Fraunhofer IOSB, 76131 Karlsruhe, Germany, {anne.sielemann, stefan.wolf, masoud.roschani, jens.ziehn}@iosb.fraunhofer.de

²Fraunhofer IPA, 70569 Stuttgart, Germany, {lena.loercher; max-lion.schumacher}@ipa.fraunhofer.de

³Karlsruhe Institute of Technology (KIT), Vision and Fusion Laboratory (IES), 76131 Karlsruhe, Germany

*This work was supported by the Fraunhofer Internal Programs under Grant No. PREPARE 40-02702 within the “ML4Safety” project, as well as funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) within the program “New Vehicle and System Technologies” as part of the AVEAS research project (www.aveas.org).



Fig. 1: Example images of Synset Signset Germany including challenging conditions as, e.g., noisy, night, overexposed, or shadowed images (lower row).

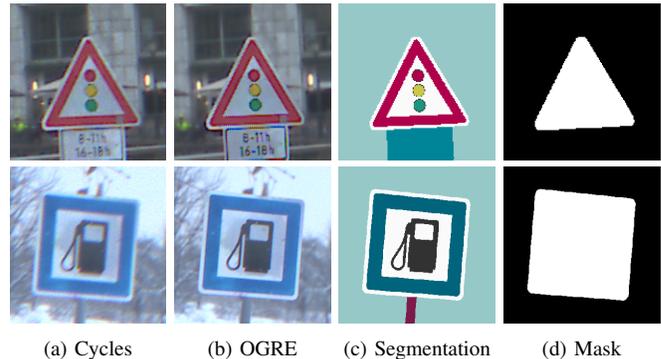


Fig. 2: Two example quartets from the Synset Signset Germany dataset.

In this context, the use of synthetic data has been considered for various roles: Most commonly, synthetic data can reduce costs and effort compared to real-world data acquisition; for example, [1] cites an average of 90 minutes for annotation and quality control of a single image of pixel-level segmentation within the Cityscapes dataset, whereas the same level of annotation can be extracted directly as ground truth from a simulation (e.g., [2], [3]). Here, the main focus is on the substitution of *training data* through synthetic data, primarily because usually larger quantities of training data are used in the development of ML systems compared to during testing. Depending on the application, synthetic data are used as an extension to available real-world training datasets. Simulated images also provide a means of producing data for rare or dangerous scenarios that can hardly be collected under real-world conditions, which can benefit training as well as testing data.

Beyond this, synthetic data can also provide an approach

TABLE I

OVERVIEW OF THE MOST RELEVANT PUBLICLY AVAILABLE TRAFFIC SIGN RECOGNITION DATASETS SORTED BY YEAR OF PUBLICATION. FOR DATASETS OF TYPE REC (RECOGNITION), THE NUMBER OF TRAFFIC SIGN INSTANCES IS EQUAL TO THE NUMBER OF SAMPLES.

Dataset	Year	Type	# Images	# Samples	# Classes	\varnothing Samples/class	Real syn	Region
MASTIF [4]	2009	rec	6 428	6 428	94	68.4	real	Croatia
MASTIF [4]	2010	det & rec	3 889	5 215	86	60.6	real	Croatia
Stereopolis [5]	2010	det & rec	847	251	10	25.1	real	France
MASTIF [4]	2011	det & rec	1 013	1 473	51	28.9	real	Croatia
STS (set 1&2) [6]	2011	det & rec	3 777	6 652	19	350.1	real	Sweden
GTSRB [7]	2011	rec	51 882	51 882	43	1 206.6	real	Germany
LISA [8]	2012	det & rec	6 610	7 855	49	160.3	real	USA
BTSC [9]	2013	rec	7 125	7 125	62	114.9	real	Belgium
TT100K [10]	2016	det & rec	100 000	30 000	221	135.7	real	China
CURE-TSR [11]	2017	rec	2 206 106	2 206 106	14	157 579.0	mixed	Belgium
TSRD [12]	2018	rec	6 164	6 164	58	106.3	real	China
European DS [13]	2018	rec	82 476	82 476	164	502.9	real	Europe
DFG [14]	2019	det & rec	6 957	17 598	200	88.0	real	Slovenia
fully annot. MTSD [15]	2020	det & rec	52 453	257 541	400	643.9	real	Global
part. annot. MTSD [15]	2020	det & rec	53 377	96 613	400	241.5	real	Global
CATERED [16]	2021	rec	94 478	94 478	43	2 197.2	syn	Germany
Synset Signset Ger. (ours)	2024	rec	105 500	105 500	211	500.0	syn	Germany

to dependable AI, by analyzing the performance of ML systems—and particularly their sensitivity to parameters—more systematically and quantitatively, specifically when used as *testing data*. This is particularly important for determining, i.e., the robustness of AI systems. However, to what extent these benefits can be leveraged depends strongly on the degree of realism in the synthetic data. The more pronounced this “sim-to-real” gap is, the less reliable conclusions are, such as conclusions about the performance of a given ML system in the real world, within the intended operational design domain (ODD).

A particularly important regulation on requirements for the use of training and testing data for AI/ML applications is the European AI Act, proposed in 2021 and expected to become law in mid-2024, stating in the texts adopted in the March 2024 resolution “Data sets for training, validation and testing, including the labels, should be relevant, sufficiently representative, and to the best extent possible free of errors and complete in view of the intended purpose of the system” (with the clause “to the best extent possible” added compared to the 2021 proposition) [17]. In this context, the quantitative comparison between domain gaps for a choice of real vs. synthetic data sources is expected to gain highly practical relevance, particularly for high-risk applications identified within the AI Act, such as “AI systems intended to be used as safety components in the management and operation of road traffic” [18, Annex III].

In this context, the task of traffic sign recognition plays multiple roles that extend beyond the immediate purpose of classifying traffic sign images into their legal categories and semantics, for example for driver assistance systems, automated driving, and mapping. Traffic sign recognition

is an extensively researched topic across a wide range of methods [9], [19], [20], [21], [22], [23], spanning a range from completely analytic approaches over classical ML with tailored models and features up to modern deep learning. In this domain, [24] is commonly cited as the first instance where a machine learning approach outperformed humans on a complex computer vision task, with the presented multi-column deep neural networks (MCDNN) achieving half the error rate of humans on the *German Traffic Sign Recognition Benchmark* (GTSRB) dataset [25]. At the same time, with new traffic signs constantly being released and coverage of existing signs in datasets still limited for a distinction of less common classes, the demand for both training and testing data still persists. This connection between a large body of recognition methods with still highly topical applications on a task that provides a relatively controlled scope motivates the choice of traffic sign recognition for an analysis of synthetic data for training and reliability assessment via XAI and robustness checks, and the generation of a novel simulated dataset.

II. STATE OF THE ART

A. Publicly available Traffic Sign Recognition Datasets

Tab. I provides an overview of the most relevant publicly available datasets for the task of traffic sign recognition. Datasets of type “recognition” (“rec”) already contain images cropped to approximately the sign size, while datasets of type “detection and recognition” (“det & rec”) show entire street scenes that must be cropped using specified bounding boxes. Most of the datasets are only valid for certain countries. The best known among the listed datasets are GTSRB [7], *Tsinghua-Tencent 100K* (TT100K) [10], the



Fig. 3: Comparison of real images represented by GTSRB (left) and state-of-the-art image synthesization methods for traffic sign recognition. For achieving a better comparability we cropped the images to a similar area if necessary. The DCGAN, LSGAN, and WGAN samples stemming from [28] result from training 200 epochs respectively.

European Dataset [13] (which includes, i.a., [4], [5], [6], [7], [29]), and the *Mapillary Traffic Sign Dataset* (MTSD) [15].

B. Usage of Synthetic Data

The comparison of synthetic data with real data on fine-grained classification was presented in [30] on the example of the *Synset Boulevard* dataset for the task of vehicle make and model recognition (VMMR). This study found synthetic data to be generally capable of achieving performance comparable to training on the real-world CompCars dataset [31] (cf. [30] also for a broader overview of synthetic data use in mobility). In the specific field of traffic sign recognition, many authors use synthetic data to increase the volume of training and/or test data, especially for rare classes. Commonly applied approaches for such synthetic data generation are:

1) *Image Augmentation*: In general, image augmentation methods (e.g., [26]) implement the following steps: They collect traffic sign templates, apply an affine transformation on them for diversifying the sign rotations and scales, vary the sign hue and/or saturation values by possibly including the background image properties or adapting the background patches, combine templates and backgrounds, and—if applicable—deploy post processing such as blur. In [32], domain randomization is used additionally, while [33] expands this procedure by randomly inserting computer generated traffic signs (for one experiment also with GAN-generated textures) to background images. The authors show that image augmentation approaches are able to expand real-world datasets in a targeted manner, but there are still disadvantages, e.g., that DNNs could overfit on domain differences or insertion edges, and that signs without dirt or wear artifacts oversimplify the classification task.

2) *Simulations*: For the creation of the CATERED dataset [16] the Carla Simulator¹ was utilized. The authors of CURE-TSR [11] expanded their dataset by adding simulated images generated by using the Unreal Engine 4². It is also

conceivable to employ computer games, as already practiced for automotive datasets for object recognition [34] or semantic segmentation [35]. With this approach, it is important to ensure that the simulation environment offers sufficient variance and that the traffic signs are not oversimplified in order to achieve an adequate degree of realism, so that the sim-to-real gap is kept as small as possible and that unrealistic overfitting is prevented.

3) *Generative Adversarial Networks (GANs)*: Other approaches, such as [27] and [28], use generative adversarial networks (GANs) to generate additional training data leading to an improvement of classification results. GANs are able to increase the degree of realism compared to the previous described approaches. However, referring to [27], applying the geometric transformation through the GAN is challenging. This is why the authors therefore implemented traditional methods and used the GANs only for synthesizing the visual appearance. This can also be observed in the results of [28], as the geometric shapes of the signs are partly imprecise. Furthermore, this approach relies on training data, which are difficult to collect for rarely occurring traffic signs.

Fig. 3 compares images of all the approaches mentioned.

III. SYSTEMATIC SYNTHETIZATION: DATASET GENERATION AND COMPOSITION

The dataset was generated through a systematic synthesization approach shown in Fig. 4, distinguishing between factors that require learning distributions from training data and factors that can be modeled analytically. The approach aims to support explainable datasets, where each aspect in the pipeline is associated with a model that is self-contained and specified as clearly as possible w.r.t. assumptions and characteristics.

A. Texture and Defect Generation

The visual appearance of traffic signs is prominently affected by deterioration of the sign surface, through wear, tear, vandalism, or fading of colors. These effects are complex

¹carla.org

²unrealengine.com

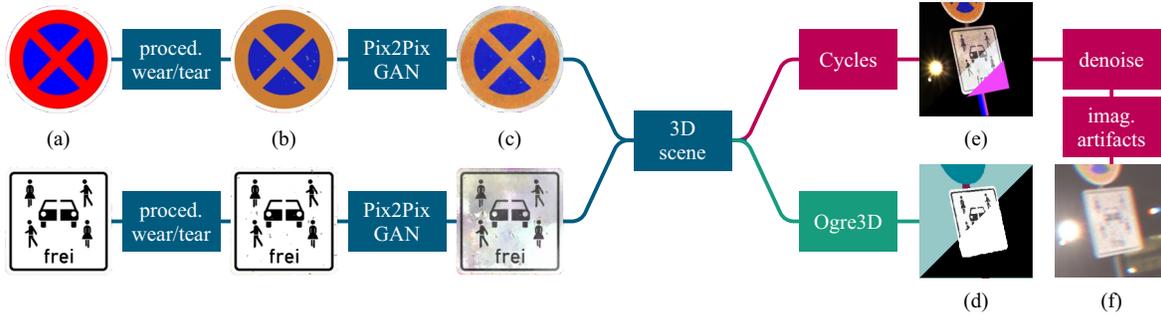


Fig. 4: Overview of the generation pipeline built in OCTANE. Ideal images (a) are procedurally converted to template images (b) defining color degradation and wear/tear masks. A GAN trained on worn traffic signs converts these to diffuse textures (c) that are combined into a 3D scene for physically-based rendering. Segmentation and mask images (d) are rendered using OGRE, while Cycles is used for geometric raytracing of HDR raw image, albedo, and normal image (e). The latter are used to denoise the raytracing samples. Based on this, imaging artifacts are computed on the 2D image data (f).

and difficult to model analytically; hence, a primarily data-driven approach was chosen to introduce realistic defects into textures.

The main goal of the particular approach is to achieve a realistic distribution of defects at variable intensities across a potentially unlimited set of traffic signs. Hence a model was designed that can be trained on acquired data but does not depend on particular sign shapes and can use medium-level annotations to selectively apply defects.

1) *GAN-based synthesis from template images:* We apply a Pix2Pix-based generative adversarial model (GAN) [36] without the central $1 \times 1 \times 512$ bottleneck to achieve a fully convolutional layout that can adapt to given input / output dimensions. With this layout, we train the GAN to convert texture patches of arbitrary size at a fixed spatial resolution of 8 px/cm, containing template images of arbitrary shapes, into the equivalent texture patch with defects. Through this, the GAN can generalize towards new physical sign sizes and new shapes; however, possible correlations between sign type (rather than visual shape) and damages (e.g., specific dirt on wild animals crossing signs) will be largely eliminated.

The GAN is trained on 200+ worn traffic signs where the color/dirt templates were extracted through classical image processing (cf. Fig. 5b–c). Color templates support black, white, and saturated colors. Gray spots annotate dirt and scratches—hence, gray is not supported as a sign color, limiting some existing variants of German traffic signs. Retroreflector patterns are excluded and retroreflection is not simulated. Pairs of mask (input) and raw (output) images are generated by randomly cropping and rotating the original images to patches of $256^2 \times 3$ and randomly shuffling the RGB channels to increase the color variation, since yellow, green, cyan, and purple hues are underrepresented or not represented at all in the original dataset.

The output textures are used exclusively as the diffuse component in the PBR (physically-based rendering, cf. Sec. III-C) surfaces.

2) *Generation of template images from sign shapes:* The template images that are used as input to the GAN (Fig. 5d) are generated from the Wikipedia overview of German traffic

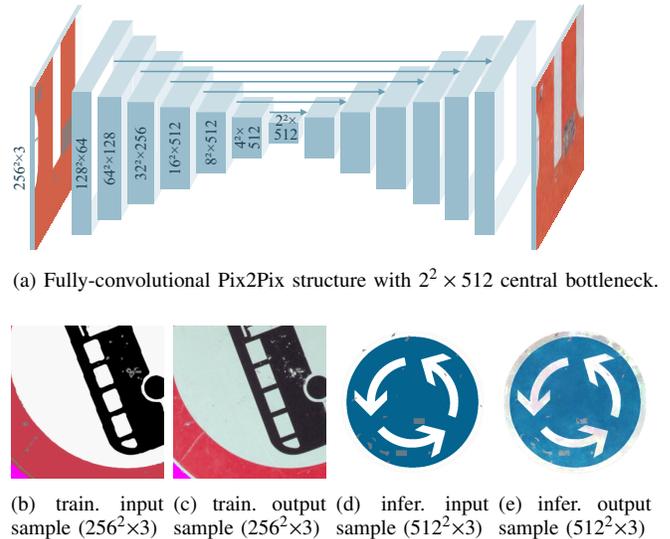


Fig. 5: GAN setup for defect synthetization.

signs³. The signs are separated into black, white, red, orange, yellow, green, and blue components. Each color component is faded stochastically and homogeneously across each sign based on the stochastic distribution of the real sign samples. Subsequently, a gray dirt mask, procedurally generated through a noise process, is overlaid, combining arbitrary shapes and rectangular shapes, the latter representing worn stickers that occur frequently in the real dataset.

B. Scene Variation

The scene variation and rendering of Synset Signset Germany is performed by the Fraunhofer simulation platform OCTANE⁴, written in C++ and following a plugin-based architecture. The following scene variations are applied:

1) *Traffic sign material:* Each traffic sign instance is assigned a unique texture generated as described in Sec. III-A. In addition, the roughness component in the PBR surface is uniformly varied in the interval between 0.2 and 0.4, the specular component between 0.3 and 0.5.

³de.wikipedia.org/wiki/Bildtafel_der_Verkehrszeichen_in_der_Bundesrepublik_Deutschland_seit_2017

⁴octane.org

2) *Traffic sign pole*: We distinguish traffic sign classes into those that are to be exclusively featured on vertical poles, and those that can occur both on vertical and on horizontal poles. In the second case, a horizontal pole is uniformly chosen with a probability of 0.3, a vertical pole otherwise. The pole diameter varies in the vertical case between 8 cm and 12 cm, in horizontal case between 8 cm and 20 cm. The roughness (between 0.4 and 0.6), and the diffuse color ($R=G=B \in [0.25, 0.4]$) of the poles’ PBR surface are also varied.

3) *Number of signs per pole*: For each traffic sign in our dataset, we manually labeled permissible possible upper and lower signs by taking the German traffic code / regulation StVO⁵ (Straßenverkehrs-Ordnung) and real-world examples into account. Thereby, we have not only considered the 211 traffic signs contained in Synset Signset Germany, but also 135 additional supplementary traffic signs. Additional traffic signs are only added to vertical poles with a probability of 0.5 to increase the dataset’s level of difficulty.

4) *Camera orientation*: We choose the camera orientation as follows: In case of a vertical pole, roll $\sim \mathcal{N}(0.0^\circ, 2.0^\circ)$, pitch $\sim \mathcal{N}(5.0^\circ, 10.0^\circ)$, and yaw $\sim \mathcal{N}(0.0^\circ, 21.0^\circ)$. For horizontal poles, we define a smaller yaw orientation range but higher pitch mean, namely roll $\sim \mathcal{N}(0.0^\circ, 2.0^\circ)$, pitch $\sim \mathcal{N}(30.0^\circ, 10.0^\circ)$, and yaw $\sim \mathcal{N}(0.0^\circ, 16.0^\circ)$. The camera is positioned so that it is centered on the traffic sign.

5) *Environment*: To modulate the environment and lighting, our approach uses image-based lighting (IBL) based on 327 uniformly sampled environment maps collected from Polyhaven⁶. Moreover, their azimuth is also varied uniformly.

6) *Occlusion object*: To cast shadows, a 3D tree object is randomly placed in the scene for $3/4$ of the images. Whether the shadow is visible on the sign also depends on the random position of the sun.

C. Optical Simulation / Rendering Pipeline

The optical simulation in OCTANE follows the general framework of physically-based rendering (PBR) [37] which provides approximately consistent models for light transport in the scene using a common set of properties. This enables the exchange of “solvers” for image generation, for which OCTANE currently supports the rasterization-based engine OGRE⁷ as well as the path tracing engine Cycles⁸ from the Blender project. We provide all 105 500 Synset Signset Germany images rendered by Cycles and OGRE respectively. For the XAI and robustness analysis the rendering was performed by OGRE to be able to test more configurations due to the reduced amount of render time. The segmentation masks and mask images were created by using OGRE.

As an approximation for the complex light transport in the scene, the modeling separates into an idealized geometric light tracing in the scene purely based on ray / surface interactions, and the computation of convolutional effects

and degradations based on the resulting high dynamic range raster images.

Thus, subsequent computations after the geometric rendering include the following:

- Stochastic errors in automatic exposure control (AEC) and white balance (WB) as presented in [30].
- Simulation of the *point spread function* (PSF) based on a Tamron M112FM35 35 mm lens to represent focusing, lens optics, and diffraction through a mixture-of-Gaussian model as presented in [30].
- Simulation of *lens flares* for visible light sources and lighting-dependent *noise*, each as presented in [30].
- Simulation of *motion blur* and *chromatic aberration* through linear convolution kernels in arbitrary direction with uniformly distributed length $\sim \mathcal{U}(0 \text{ px}, 10 \text{ px})$.
- Simulation of *digital image sharpening* effects using unsharp masking.
- Addition of artifacts from Bayer BGGR bilinear *demo-saicing* as in [30].

For all simulated effects and artifacts, the stochastically selected parameter values are given per individual image in the dataset.

D. Dataset Statistics

Our resulting dataset *Synset Signset Germany* contains 211 traffic sign classes depicted in Fig. 6. The dataset is balanced with 500 images per class, resulting in a total of 105 500 images. Thereby, the traffic sign classes can be grouped as follows:

17	Speed limit signs	45	Danger signs
33	Other prohibitory signs	21	Derestriction signs
12	Stop, wait, and parking signs	28	Information signs
13	Driving lane control signs	4	Priority signs
13	Special zones and way signs	4	Highway signs
13	Additional road signs	8	Other signs

The image resolutions in our dataset vary between the maximum resolution of 389×389 pixels and the minimum resolution of 22×22 pixels.

IV. SYNTHETIC DATA AS TRAINING DATA

To determine the degree of realism in the synthetic data, we evaluate our dataset in comparison to the GTSRB dataset [7] based on the subset of the first 43 classes in Synset Signset that overlaps with GTSRB. Additionally, we utilize the CATERED [16], [38] dataset for training and evaluation as a synthetic reference dataset. For all experiments, we employ a ConvNeXt-Small [39] network with similar settings as Sielemann et al. [30]. We only refrain from applying random flip augmentation since the orientation of some signs are a distinguishing feature, and we utilize different learning rates. For the evaluation on CATERED, we additionally remove the center crop and instead directly resize to 224×224 since the images in CATERED are already tightly cropped.

Regarding the learning rates, we train models with learning rates of 10^{-4} , 10^{-3} , and 10^{-2} and choose the learning rate

⁵stvo2go.de/verkehrszeichen-wissensnetz

⁶polyhaven.com

⁷ogre3d.org

⁸cycles-renderer.org



Fig. 6: Overview of the signs in Synset Signset Germany. The first row of 43 signs corresponds to the classes in GTSRB. Sign shapes are based on the Wikipedia overview of German traffic signs from 2017 onwards (cf. footnote 3).

TABLE II

TOP-1 ACCURACY OF EACH COMBINATION OF TRAINING AND TESTING ON THE CONSIDERED DATASETS. THE RESULTS INDICATE THE HIGH EFFECTIVENESS OF SIGNSET FOR TRAINING AS WELL AS FOR EVALUATION PURPOSES.

Eval. ▶ Train. ▼	Signset Cycles	Signset OGRE	GTSRB	CATERED
Signset Cycles	99.5%	99.4%	98.3%	84.4%
Signset OGRE	99.6%	99.6%	98.2%	84.6%
GTSRB	89.4%	87.4%	99.9%	77.1%
CATERED	50.0%	48.6%	76.4%	86.1%

with the best in-domain result for each of the datasets. This is done to choose the learning rate which is most appropriate for training on each dataset while reducing the risk of overfitting in the cross-domain evaluations. We apply an 80–20 split to Signset for extracting a training and a validation set while utilizing the official train–validation splits for the other datasets. The results as measured with top-1 accuracy are shown in Tab. II. They show an accuracy above 80% for the evaluation on all three datasets when training on Synset Signset Germany. For the evaluation in the cross-dataset scenarios, the scores are just closely behind the in-domain trainings, only lacking 1.2 percentage points when evaluating on the real-world GTSRB dataset, while the evaluation on Signset shows the highest score by a large margin. This highlights the usefulness of our dataset for training classification models. Moreover, the large margin between training on Signset compared to training on one of the other datasets for an evaluation on Signset indicates the challenge of the training dataset, and thus, a high usefulness for evaluation purposes considering the saturation of possible improvements on the GTSRB. It additionally provides a significantly higher value due to the inclusion of a total of 211 instead of 43 classes, with a model trained on all classes still achieving a score of 99.6% for both full Signset dataset versions, Cycles and OGRE.

V. SYNTHETIC DATA FOR XAI AND ROBUSTNESS ANALYSIS

Synthetic data can also play a relevant role in the investigation of ML models in terms of robustness (stability of the model prediction performance w.r.t. input perturbations) and explainability, which attempts to explain model decisions in order to increase their comprehensibility. Both of these aspects are important in assessing the reliability of a trained AI/ML system, especially if that system is to be used in a

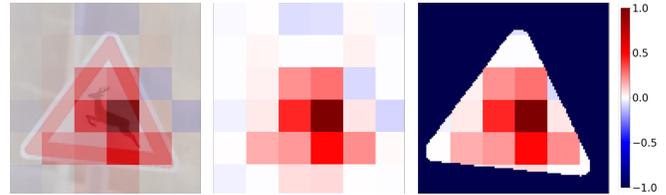


Fig. 7: Example image with its corresponding FA map overlaid on the left. The center image shows the pure FA map, and the right image shows only the features (i.e., the image pixels) with increased attribution value that make up the traffic sign.

safety-relevant context.

In this regard, we use the Synset Signset Germany synthesis pipeline to evaluate specific explainability and global robustness measures. By introducing certain parametric perturbations, we can ascertain the explanation quality and robustness level for arbitrary parameters. Explanation quality is quantified via the *pixel ratio*, under the premise that a “good” explanation should predominantly highlight the object of interest rather than something in the background. In the experiments described, explanations are generated using local saliency methods that yield *feature attribution* (FA) maps per image. These maps, when combined with binary mask images (cf. Fig. 2, right), enable us to assess the amount of attributed features (i.e., pixels with a positive attribution value) that belong to the traffic sign. Specifically, the pixel ratio is defined as the proportion of positively attributed pixels—weighted by their attribution value—constituting the traffic sign relative to those in the entire image. This definition is illustrated in Fig. 7: The pixel ratio is the ratio of the positive FA in the right image to the positive FA in the center image. The higher its value, the better the explanation. The FA was computed using the KernelSHAP method from the Captum library⁹.

Global robustness is evaluated through the method described in [40]. The evaluation is based on a sequence of hypothesis tests certifying a specific level of global robustness given a required confidence level. In contrast to the aforementioned method, instead of comparing the model prediction for the original input with the prediction for the perturbed input, here we compared the prediction for the original input with the ground-truth label. In that sense prediction performance for different perturbation intensities is used to assess robustness.

To demonstrate the approach, we consider a ResNet-18 [41] trained on the GTSRB dataset [7]. However, the procedure works for any model. We focus on one specific perturbation, namely motion blur, and vary its intensity

⁹captum.ai

TABLE III
RESULTS OF THE XAI AND ROBUSTNESS ANALYSIS ON THREE DIFFERENT DATASETS WITH IMAGES OF VARYING MOTION BLUR INTENSITY.

Dataset	Pixel Ratio	Global Robustness
No motion blur	0.74	0.9
Mid-level motion blur	0.63	0.88
High-level motion blur	0.58	0.84



Fig. 8: Example images for the three different datasets. On the left is an image with no motion blur, in the center an image with mid-level motion blur, and on the right an image with high-level motion blur.

parameter in order to create three distinct traffic sign datasets for our experiments. These datasets correspond to none, mid-level, and high-level motion blur, as depicted in Fig. 8. The results of the experiments are reported in Tab. III.

As expected, prediction performance as well as the quality of explanations (i.e., the pixel ratio) decrease with increasing motion blur intensity. From this it can be concluded that the model’s performance diminishes when processing images with increasing levels of motion blur.

Overall, the benefit of Synset Signset Germany for XAI and robustness analysis lies in the ability to obtain quantitative measures for specific desired perturbation intensities. Furthermore, the inverse problem of finding intensity parameters to a required level of robustness and explanation quality can be addressed.

VI. CONCLUSION AND OUTLOOK

We have presented the Synset Signset Germany dataset, a synthetic dataset for the task of traffic sign recognition, containing a total of 105 500 images of 211 different German traffic sign classes, including comparatively rare and very recent traffic signs. A subset of 43 classes in the dataset aims to represent a “synthetic twin” of the GTSRB dataset [7] with similar imaging parameters.

For each sign, detailed, stochastically chosen synthetization parameters are provided, along with additional binary mask and a segmentation mask label images. This is intended to support both the use of the dataset to understand machine learning effects on real data due to known “ground truth” parameters in the simulated images, and to understand the impact of different simulation methods on dataset quality.

Through this, the resulting dataset is among the largest and most diverse datasets for traffic sign recognition and, to the best of our knowledge, one of the first publicly available large-scale synthetic datasets for this task.

Our implemented synthesis pipeline proved to combine the advantages of data-driven and analytical modeling. Com-

pared to the purely analytically simulated CATERED dataset, Synset Signset Germany achieves an approximately 20% better top-1 accuracy, which, together with the high cross-dataset scores, indicates a good generalization ability probably due to the increased level of realism resulting from the GAN generated textures.

Outlook

One of the main advantages in the use of synthetic data generation is its scalability towards further applications. Hence, based on the work, an important next step is to abandon the current limitation on German traffic signs and provide extensions towards international traffic signs.

A widely acknowledged limitation of synthetic data, in turn, is the sim-to-real domain gap. While the practical experiments indicate that the domain gap is sufficiently low for practical applications as training data, and that the data/metadata composition is well suited XAI, the requirements for the use as testing data are considerably higher. Here, the dataset not only has to cover the target domain sufficiently to train adequate generalization capabilities, but instead must also enable the quantitative performance estimation of trained models by relating effects on synthetic data to those on (yet unseen) real-world data. While the extensive annotation provided with the dataset is expected to support research in this area, i.e., by conducting XAI and robustness analyses for the remaining parameters, there is still considerable demand for future research.

The GAN-based defect synthesis so far uses only a very simple concept that does not distinguish between types of dirt and damage/wear and lacks representation of features such as gray sign areas and retroreflectors. Future work should improve on these limitations.

Furthermore, the number of occlusion objects should be increased in order to achieve more complex and diverse shadow casts and occlusions of the traffic signs. If the size of the synthetic dataset is to be significantly increased, it would be advisable to collect more environment maps to gain a higher data variance.

Eventually, the choice of recognition models in this paper is limited to few deep learning models within the state of the art. A more extensive analysis including more diverse models and potentially also some classical, non-deep learning approaches would substantiate the findings and extend the understanding of the applicability of the dataset.

ACKNOWLEDGMENT

We would like to thank the civil engineering dept. of the city of Karlsruhe, Germany (Tiefbauamt Karlsruhe) for their kind support in creating the dataset of worn traffic signs.

REFERENCES

- [1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The Cityscapes Dataset for Semantic Urban Scene Understanding,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

- [2] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?" *arXiv preprint arXiv:1610.01983*, 2016.
- [3] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4340–4349.
- [4] S. Šegvić, K. Brkić, Z. Kalafatić, V. Stanislavljević, M. Ševrović, D. Budimir, and I. Dadić, "A computer vision assisted geoinformation inventory for traffic infrastructure," in *13th international IEEE conference on intelligent transportation systems*. IEEE, 2010, pp. 66–73.
- [5] R. Belaroussi, P. Foucher, J.-P. Tarel, B. Soheilian, P. Charbonnier, and N. Paparoditis, "Road sign detection in images: A case study," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 484–488.
- [6] F. Larsson and M. Felsberg, "Using Fourier descriptors and spatial models for traffic sign recognition," in *Image Analysis: 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, May 2011. Proceedings 17*. Springer, 2011, pp. 238–249.
- [7] 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.
- [8] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *IEEE transactions on intelligent transportation systems*, vol. 13, no. 4, pp. 1484–1497, 2012.
- [9] M. Mathias, R. Timofte, R. Benenson, and L. Van Gool, "Traffic sign recognition—How far are we from the solution?" in *The 2013 international joint conference on Neural networks (IJCNN)*. IEEE, 2013, pp. 1–8.
- [10] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-Sign Detection and Classification in the Wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [11] 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.
- [12] L. Huang, "Chinese Traffic Sign Database," <https://nlpr.ia.ac.cn/pal/trafficdata/recognition.html>, 2018.
- [13] C. G. Serna and Y. Ruichek, "Classification of traffic signs: The european dataset," *IEEE Access*, vol. 6, pp. 78 136–78 148, 2018.
- [14] D. Tabernik and D. Škočaj, "Deep Learning for Large-Scale Traffic-Sign Detection and Recognition," *IEEE Transactions on Intelligent Transportation Systems*, 2019.
- [15] C. Ertler, J. Mislej, T. Ollmann, L. Porzi, and Y. Kuang, "Traffic Sign Detection and Classification around the World," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [16] I. Sinioglou, P. Sarigiannidis, Y. Spyridis, A. Khadka, G. Efstathiopoulos, and T. Lagkas, "Synthetic Traffic Signs Dataset for Traffic Sign Detection & Recognition In Distributed Smart Systems," in *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2021, pp. 302–308.
- [17] European Commission, "European Parliament legislative resolution of 13 March 2024 on the proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (Texts Adopted, COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))," Mar. 2024.
- [18] —, "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM(2021) 206 final, ANNEXES 1 to 9)," Apr. 2021.
- [19] A. De La Escalera, L. E. Moreno, M. A. Salichs, and J. M. Armingol, "Road traffic sign detection and classification," *IEEE transactions on industrial electronics*, vol. 44, no. 6, pp. 848–859, 1997.
- [20] A. De la Escalera, J. M. Armingol, and M. Mata, "Traffic sign recognition and analysis for intelligent vehicles," *Image and vision computing*, vol. 21, no. 3, pp. 247–258, 2003.
- [21] C. Bahlmann, Y. Zhu, V. Ramesh, M. Pellkofer, and T. Koehler, "A system for traffic sign detection, tracking, and recognition using color, shape, and motion information," in *IEEE Proceedings. Intelligent Vehicles Symposium, 2005*. IEEE, 2005, pp. 255–260.
- [22] S. Maldonado-Bascón, S. Lafuente-Arroyo, P. Gil-Jimenez, H. Gómez-Moreno, and F. López-Ferreras, "Road-sign detection and recognition based on support vector machines," *IEEE transactions on intelligent transportation systems*, vol. 8, no. 2, pp. 264–278, 2007.
- [23] M.-Y. Fu and Y.-S. Huang, "A survey of traffic sign recognition," in *2010 International conference on wavelet analysis and pattern recognition*. IEEE, 2010, pp. 119–124.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] C. Grigorescu and N. Petkov, "Distance sets for shape filters and shape recognition," *IEEE transactions on image processing*, vol. 12, no. 10, pp. 1274–1286, 2003.
- [30] 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.
- [31] L. Yang, P. Luo, et al., "A Large-Scale Car Dataset for Fine-Grained Categorization and Verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, change Loy, Chen and Tang, Xiaou.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] M. Pharr, W. Jakob, and G. Humphreys, *Physically based rendering: From theory to implementation*. MIT Press, 2023.
- [38] B. Villarini, P. Radoglou-Grammatikis, T. Lagkas, P. Sarigiannidis, and V. Argyriou, "Detection of physical adversarial attacks on traffic signs for autonomous vehicles," in *2023 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*. IEEE, 2023, pp. 31–37.
- [39] Z. Liu, H. Mao, C.-Y. Wu, et al., "A ConvNet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11 976–11 986, feichtenhofer, Christoph and Darrell, Trevor and Xie, Saining.
- [40] M.-L. Schumacher and M. F. Huber, "Probabilistic Global Robustness Verification of Arbitrary Supervised Machine Learning Models," in *International Conference on Machine Learning (ICML), 2nd Workshop on Formal Verification of Machine Learning*, 2023.
- [41] K. He, X. Zhang, et al., "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778, ren, Shaoqing and Sun, Jian.