Measuring the Effect of Background on Classification and Feature Importance in Deep Learning for AV Perception

Anne Sielemann*, Valentin Barner*, Stefan Wolf†*, Masoud Roschani*, Jens Ziehn*, and Juergen Beyerer*†

*Fraunhofer IOSB, Email: {anne.sielemann, masoud.roschani, jens.ziehn}@iosb.fraunhofer.de

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

Abstract—Common approaches to explainable AI (XAI) for deep learning focus on analyzing the importance of input features on the classification task in a given model: saliency methods like SHAP and GradCAM are used to measure the impact of spatial regions of the input image on the classification result. Combined with ground truth information about the location of the object in the input image (e.g., a binary mask), it is determined whether object pixels had a high impact on the classification result, or whether the classification focused on background pixels. The former is considered to be a sign of a healthy classifier, whereas the latter is assumed to suggest overfitting on spurious correlations.

A major challenge, however, is that these intuitive interpretations are difficult to test quantitatively, and hence the output of such explanations lacks an explanation itself. One particular reason is that correlations in real-world data are difficult to avoid, and whether they are spurious or legitimate is debatable. Synthetic data in turn can facilitate to actively enable or disable correlations where desired but often lack a sufficient quantification of realism and stochastic properties.

To shed light on this issue and test whether feature importance-based XAI reliably distinguishes between true learning and problematic overfitting, we utilize the task of traffic sign recognition. Based on the synthesis pipeline of the Synset Signset Germany dataset, which demonstrated comparability to real-world data, we show how systematically generated synthetic data can test assumptions about feature importance-based XAI and isolate factors between classification quality and XAI values.

Therefore, we systematically generate six synthetic datasets for the task of traffic sign recognition, which differ only in their degree of camera variation and background correlation. The generated datasets, which we provide for download under a CC-BY license, enable us to quantify the isolated influence of background correlation, different levels of camera variation, and considered traffic sign shapes on the classification performance, as well as background feature importance. A study of this kind is nearly impossible to conduct with real-world data, as real-world data can only be collected with difficulty at this level of comparability and without additional influencing factors. Results include a quantification of when and how much background features gain importance to support the classification task based on changes in the training domain, and show that such metrics can be indicative of complex properties of the training data and task, not purely of learning quality.

I. Introduction

THE principal strength of machine learning (ML) models and especially deep neural networks (DNNs) is their ability to accurately approximate given data distributions. This leads to impressive results, as complex relationships in training data can be learned and mapped [1]; however, in turn, these complex relationships cannot be trivially understood by humans, making DNN decisions incomprehensible; at the same time, the growing scope of AI/ML applications has given rise to a growing number of regulations and standards aiming at transparency and trustworthiness [2]. To overcome this issue, *explainable artificial intelligence* (XAI) methods were introduced to improve transparency, interpretability, and thus error analysis and trustworthiness of ML applications. For computer vision tasks, saliency methods such as, e.g., *Kernel SHAP* [3] and *GradCAM* [4] are predominantly used, providing users with so-called *feature attribution* (FA) maps per input image, which quantify the contribution of each (super-)pixel to the model's prediction.

In case of classification tasks, these methods are used to gauge the validity of a classifier's learned features, such that a healthy classifier is expected to base its decisions primarily on features located on the object to be classified, rather than the background (cf. [5], [6]). Whether, however, low feature attribution on the background does indeed distinguish a healthy classifier from an unhealthy one (namely one that overfitted on spurious background correlations), as intuition may suggest, has thus far not been evaluated systematically. However, the validity of explanations in XAI depends critically on aligning AI/ML properties adequately with human understanding; merely transforming abstract output into a form that invites intuitive but error-prone interpretation will clearly serve no favorable purpose. Therefore, it is pivotal to improve the understanding of XAI metrics to ensure that these, in turn, can contribute to improving the understanding of AI/ML methods.

Such a systematic study of an XAI method is commonly difficult to achieve, as ML tasks and data are usually complex and difficult to control, rendering it difficult to establish a reliable baseline against which the performance of an XAI method could reliably be quantified. To address this challenge, we utilize the traffic sign recognition use case, since it is a well-understood classification task offering classes with and

without human-modeled background correlation. We base the investigations of this work on synthetic / simulative data, since it allows one to use highly accurate labels and selectively introduce specific biases while maintaining, most importantly, that training and testing data can be independent and identically distributed (i.i.d.) in a strict sense, thus enabling systematic evaluations at a level that is usually impossible for real-world data. For data synthetization, we utilize the generation pipeline presented in [5], by which a synthetic twin of the well-known *German Traffic Sign Recognition Benchmark* (GTSRB) [7] was created, named *Synset Signset Germany*. The cross-dataset evaluations in [5], which indicate a good degree of realism for the synthetic data and a relatively narrow domain gap between the provided Synset Signset dataset and GTSRB [7], ensure that the conclusions achieved herein are closely linked to reality rather than merely hypothetical results on simplified toy examples.

Overall, this publication aims to...

(I) ... systematically examine the XAI assumption of healthy classifiers to predominantly focus on the foreground object, by analyzing which data properties increase the amount of background feature importance and measure their impact on the classification performance,

(II) ... highlight the usefulness of synthetic data for investigating and measuring DNN properties and thus quantitatively evaluating XAI metrics,

(III) ... provide state-of-the-art synthetic datasets for the task of traffic sign recognition and suitable for further XAI metric investigations, containing accurate labels, segmentation images, masks, and specifically modeled biases,

(IV) ... and quantify the influence of environment correlations on the task of traffic sign recognition and thus sensitize DL researchers and developers to the importance and consequences of data selection.

II. State of the Art

A. Synthetic Data (Generation) for AV Perception

The use of synthetic data in computer vision tasks has seen wide application in recent years, where wide overviews of the topic for heterogeneous applications can be found in [8]–[10] for example. In the domain of *automated vehicle* (AV) perception, synthetic datasets have been used successfully for training, including datasets based directly on computer games [11]–[13] and computer game *engines*, such as *Unity*[1] or *Unreal Engine*[2] [14]–[19], including data simulated in the *Carla* Simulator[3]. Other approaches to synthetic data include real-world data modified through augmentation and recombination (e.g., [20]–[23]), or the use of generative AI to create sensor data, such as [24], [25] for the specific task of traffic sign recognition.

---

[1]unity.com
[2]unrealengine.com
[3]carla.org

B. Evaluation of XAI Methods

Evaluation of XAI methods is a challenging task: Qualitative evaluations are subject to human cognitive biases [26] and are therefore not considered sufficiently objective [27]. Hence, the XAI community is anxious to find quantitative and thus more objective evaluation metrics/methods for assessing DNN explanations: [28]–[31] give deeper insights and systematic reviews on this topic. Based on extensive literature research, Nauta et al. [28] identified twelve properties that optimal explanations are desired to fulfill, the so-called *Co-12 properties*, of which six evaluate XAI methods in terms of content. The authors highlight synthetic data as useful for assessing the *correctness* property by a "controlled synthetic data check". However, other works make use of synthetic data to evaluate additional properties, such as, e.g., the *completeness* ("preservation check" [32], "deletion check" [32]) or *contrastivity* ("target sensitivity" [32]).

A general challenge of quantitative evaluation of XAI metrics is the lack of available ground truth [26], as, for example, semantic masks are usually not included in the metadata of classification datasets. To overcome this issue, several synthetic datasets especially designed for the task of evaluating XAI metrics were introduced in recent years: The *Toy Color Dataset* [33] (contains 5×5 pixel images with four possible pixel colors, where DNNs can learn simple color conditions), the *an8Flower* dataset [34] (a dataset of different colored flower parts), or the *FunnyBirds* dataset [32] (includes bird images from which individual object parts can be removed).

III. Influencing Factors of Background Attention

To determine the effect of background on the task of traffic sign recognition, we want to measure (I) the extent to which DNNs trained on datasets with different modeled properties take background into account for their classification decision and (II) how this background consideration affects the classification performance. This enables us to draw conclusions about which data properties encourage DNNs to focus on backgrounds and whether this background attention is spurious or justifiable regarding the achieved classification performance.

We assume three dataset properties to likely influence the amount of background attention:

(I) Correlation of background: In a correlated traffic sign recognition dataset, traffic signs appear mainly in their most probable environment, which turns the background into a source that can provide clues to the traffic sign's class. This offers an incentive to DNNs to also include background features in their classification decisions, resulting in a greater importance of background features.

(II) Degree of camera variation: A higher range of camera variation might encourage DNNs to focus on traffic sign border areas to perceive their optical distortion.

(III) Traffic sign shapes: Depending on the respective task definition, identifying the traffic sign shape (by actively distinguishing foreground and background) can be advantageous

as it is a discriminative feature and allows one to exclude a subset of traffic signs during the classification.

All three possible factors should be considered in the evaluation to review and measure their influence and are thus taken into account during the dataset generation.

IV. Dataset Generation

To generate the needed datasets, we utilized our parameterizable rendering pipeline from our previous work on the Synset Signset Germany dataset [5]. With the pipeline, our goal was to combine the advantages of data-driven and analytical modeling. Therefore, we added a GAN-based texture generation to our self-developed simulation platform OCTAS®[4] (for details, see [5]). For the rendering process, OCTAS® currently supports the usage of the rasterization-based engine OGRE3D[5] as well as the path tracing engine Cycles[6] developed by the Blender project. Since our evaluation results in [5] have shown that there was nearly no difference between the generated data from both approaches, we decided to solely use the less computationally expensive OGRE3D engine for this work.

A. Correlated vs. Uncorrelated Environment

To compare datasets with correlated environment to those with uncorrelated environment, we need to define both terms:

1) Correlated environment: Each traffic sign is depicted in its most probable environment according to the German traffic code / regulation StVO[7] (Straßenverkehrs-Ordnung) categorized into "urban", "nature", and "urban and nature". For example, a sign warning of wildlife crossing is likely to be set up on a rural road (natural background), while a sign warning of children is probable to be placed in an urban context.

2) Uncorrelated environment: The traffic sign environment is uniformly distributed randomly chosen from the combined set of urban and nature maps, and therefore does not have a semantic connection to the depicted sign.

The rendering pipeline utilizes image-based lighting (IBL). To realize the un-/correlation, we collected all available urban environment maps from Polyhaven[8] and manually restricted the set to only those with a predominantly urban background. Consequently, maps are filtered out that were, e.g., captured in city parks and are therefore predominantly surrounded by nature. A subset of 70 remaining maps results. Polyhaven offers a larger selection of environment maps labeled as nature. However, for a fair comparison we selected 70 environment maps of type nature as well, where we tried to achieve a comparable distribution of day times, although fewer night maps were available. In Tab. I, the property distributions of the urban and nature environment maps are compared.

[4] octane.org, formerly OCTANE
[5] ogre3d.org
[6] cycles-renderer.org
[7] stvo2go.de/verkehrszeichen-wissensnetz
[8] polyhaven.com

TABLE I
COMPARISON OF THE BACKGROUND PROPERTY DISTRIBUTIONS FOR THE SETS OF URBAN AND NATURE ENVIRONMENT MAPS. BOTH SETS INCLUDE 70 MAPS EACH.

		nature		urban	
	morning / aftern.	28	40.0 %	26	37.1 %
	sunrise / sunset	22	31.4 %	18	25.7 %
DAY TIME	midday	16	22.9 %	10	14.3 %
	night	4	5.7 %	13	18.6 %
	not specified	0	0.0 %	3	4.3 %
	clear	21	30.0 %	20	28.6 %
WEATHER	partly cloudy	35	50.0 %	28	40.0 %
	overcast	11	15.7 %	18	25.7 %
	not specified	3	4.3 %	4	5.7 %
	low	28	40.0 %	32	45.7 %
CONTRAST	medium	10	14.3 %	16	22.9 %
	high	32	45.7 %	22	31.4 %
	natural	69	98.6 %	56	80.0 %
LIGHT	artificial	1	1.4 %	13	16.6 %
	not specified	0	0.0 %	1	1.4 %
TOTAL		70	100 %	70	100 %

B. Traffic Sign Selection

We selected the included traffic sign classes with great care, as we wanted some properties to be evenly distributed across the datasets. This comprises:

1) Traffic sign shapes: The generated datasets include 25 circular, triangular, and rectangular traffic signs each. In addition, seven signs of various shapes were added. An overview of the traffic signs included by shape is given in Tab. II (top).

2) The most probable traffic sign environment: We considered including almost the same number of traffic signs to be most probable in an urban and natural environment. Thereby, we regarded the traffic sign shapes round, triangular, and rectangular to also be almost equally distributed in urban and natural environments. There is also a set of traffic signs that are likely to appear in urban as well as natural environments. Tab. II (bottom) shows the included traffic sign classes sorted by the most likely environment of occurrence.

3) Traffic sign colors: For all classes of traffic sign shapes as well as probable environments, our objective was to distribute the appearing colors evenly when possible, to prevent trained networks from overfitting color details. The triangular German warning signs are exclusively red and black with one exception, which explains why the triangular signs are predominantly red. However, the classes round, rectangular, urban, and nature all contain red, gray, and blue traffic signs (cf. Tab. II).

4) Risk of confusion: For all traffic sign shapes, we included traffic sign classes that are likely to be confused with each other. This applies, e.g., to classes that only differ in vertical mirroring (e.g., left curve and right curve) or local details (e.g., pedestrian zone and bike zone).

Note that the 43 GTSRB traffic sign classes [7] are included.

TABLE II
INCLUDED TRAFFIC SIGNS BY SHAPE (TOP) AND MOST PROBABLE ENVIRONMENT (BOTTOM).

TABLE III
OVERVIEW OF THE SIX GENERATED SYNTHETIC DATASET'S PROPERTIES.

dataset	correlation	viewport variation	train	test
UF	uncorrelated	frontal	500	600
UM	uncorrelated	medium	500	600
UH	uncorrelated	high	500	600
CF	correlated	frontal	500	600
CM	correlated	medium	500	600
CH	correlated	high	500	600

C. Stages of Camera Variation

We assume that the degree of camera variation in the training data affects the environmental attention of the resulting DNN. Therefore, we provide each dataset at three levels of camera variation, using normal distributions $\mathcal{N}(\mu, \sigma)$:

1) Frontal (F): No camera variation, roll $\sim \mathcal{N}(0.0°, 0.0°)$, pitch $\sim \mathcal{N}(0.0°, 0.0°)$, and yaw $\sim \mathcal{N}(0.0°, 0.0°)$.

2) Medium (M): roll $\sim \mathcal{N}(0.0°, 1.5°)$, pitch $\sim \mathcal{N}(0.0°, 5.0°)$, and yaw $\sim \mathcal{N}(0.0°, 13.\overline{3}°)$.

3) High (H): roll $\sim \mathcal{N}(0.0°, 3.0°)$, pitch $\sim \mathcal{N}(0.0°, 10.0°)$, and yaw $\sim \mathcal{N}(0.0°, 26.\overline{6}°)$.

D. Other Dataset Configurations

In contrast to Synset Signset Germany [5], we have abstained from using horizontally oriented traffic sign poles and from adding additional signs to poles in this work. Multiple signs or a change in pole orientation might additionally draw the DNN's attention to areas outside the traffic sign surface in focus, and therefore distort the results. We did not change the generation of template images from sign shapes, the GAN-based texture and defect synthesis from template images, the traffic sign material, traffic sign pole diameter, and pole material variation, as they all proved to be reasonable. In addition, the positioning and orientation of the 3D tree geometry, which is used for occlusions and to cast shadows, remains unchanged. We reused the simulated effects and artifact configuration, but internally improved the white balancing and AEC error calculation to reduce the number of extremely overexposed images. For details on the pipeline, cf. [5].

E. Resulting Datasets

We used the described data generation process and configuration to create a total of six synthetic traffic sign image datasets. Tab. III gives an overview of their abbreviations, high-level configurations, and number of images per class. Each dataset contains 82 classes of traffic signs resulting in 41 000 train and 49 200 test images per dataset, summing up to a total of 541 200 images. Exemplary images are depicted in Fig. 1. We offer all of the six datasets publicly available under the CC-BY license (download link on the first page).

Note that a two-letter abbreviation denotes the whole dataset, e.g., UF$_{\text{train}}$ refers to the complete uncorrelated training dataset captured in frontal view, consisting of 82 classes with 500 images each. In a three-letter abbreviation, the third letter C (circular), T (triangular), or R (rectangular) refers to a dataset's subset containing all signs of a specific shape (cf. Tab. II (top)). So, e.g., the abbreviation CMT$_{\text{test}}$ denotes the correlated test dataset's subset of triangular-shaped traffic signs captured with medium camera variation.

V. EVALUATION

A. Training Setup and Network Instances

For our experiments, we employ a *ConvNeXt-Small* (CNs) [35], *ConvNeXt-Tiny* (CNt) [35], and *ResNet50* (RN50) [36] classification network from *OpenMMLab's* pretraining toolbox *MMPreTrain* [37]. On the one hand, we expect from this selection to gain insight on the influence of network size (CNs vs. CNt) and, on the other hand, to compare the behavior of a state-of-the-art architecture (ConvNeXt) to an older one (RN50). The fundamental training set-up is adopted from Sielemann et al. [38], whereby we refrain from applying random flip augmentation because some traffic signs only differ in vertical mirroring. Our previous experiments on Synset Signset Germany [5] showed that the network trained with a learning rate of 10^{-3} achieved the best in-domain result and therefore is also applied for this work.

Overall, we trained nine different network instances per architecture: one on each of the six generated datasets training sets (82 classes), and, for later conclusions on the influence of the included traffic sign shapes, one on the circular, triangular, and rectangular subset of the CM$_{\text{train}}$ dataset. Each of these subsets comprises 25 classes. For each training, we saved the

(a) correlated frontal (CF_{train})   (b) correlated medium (CM_{train})   (c) correlated high (CH_{train})

(d) uncorrelated frontal (UF_{train})   (e) uncorrelated medium (UM_{train})   (f) uncorrelated high (UH_{train})
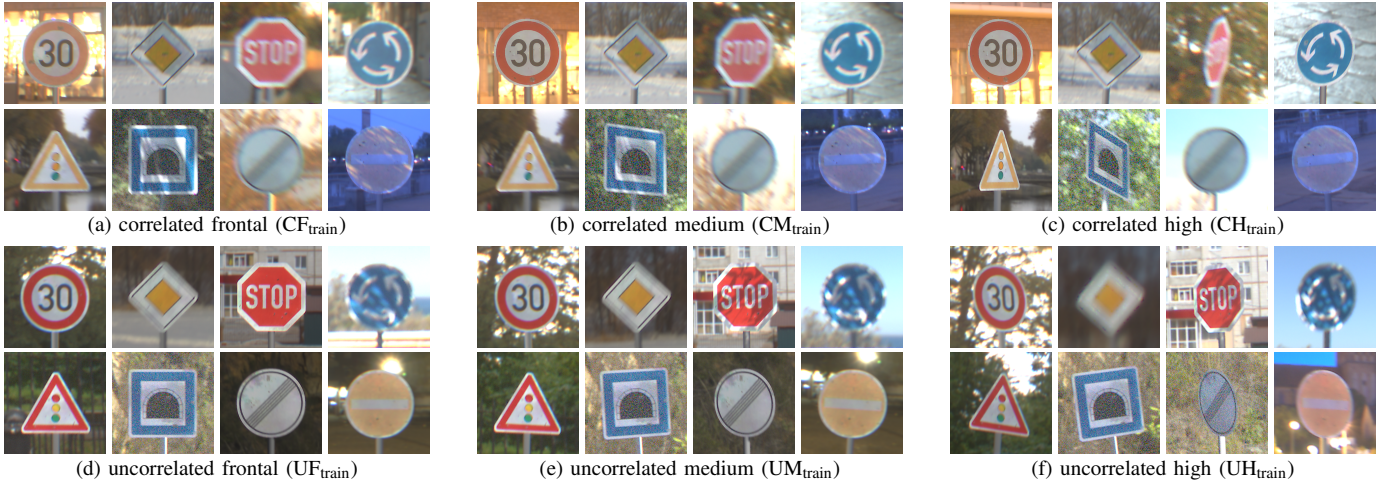
Fig. 1: Example images from the six generated synthetic datasets.

weights after the 100th epoch, as well as the best configuration validated on the respective corresponding test dataset. We denote the networks by DNN(dataset), so, e.g., $CNs_{100}(UF_{train})$ refers to the ConvNeXt-Small instance trained for 100 epochs on the training set of the uncorrelated frontal dataset.

B. Background Correlation Effect on Feature Importance

For calculating *feature attributions* (FA), we apply *Kernel SHAP* (KS) [3], as it is model agnostic and has proved to be well suited to explain classifications on *ImageNet* [29], [39], and the widely known *GradCAM* (GC) [4] method, both from the *Captum* library[9]. However, in principle, the metrics could also be replaced by another suitable saliency method. We define the KS hyper-parameters "baseline" to zero, the "number of samples" to 1 000, and group areas of 32×32 pixels into superpixels, leading to a FA resolution of 7×7. Based on the FAs, the *pixel ratio* can be calculated, as for the robustness analysis conducted in [5]. It is defined as the ratio of positive attributing features within the traffic sign image area relative to the absolute value of attributing features in the whole image. Fig. 2 provides a visualization of the components used for the pixel ratio computation using KS: The KS values are computed based on the examined network instance in addition to an input image. Related to each image, the rendering pipeline outputs a binary mask which is used to crop the positive KS values to the traffic sign dimensions. These are summed and divided by the total sum of positive pixel attributions, resulting in the pixel ratio.

In the first step, we calculate for both methods the mean pixel ratios for all considered network architectures trained on the six synthetically generated training datasets for 100 epochs and evaluated on CF_{test}. But due to the high computing time requirements of KS, we make two compromises:

(I) We restrict the feature importance experiments to evaluate on one of the generated test sets. The choice has been made for correlated, as the real world is assumed to be correlated,
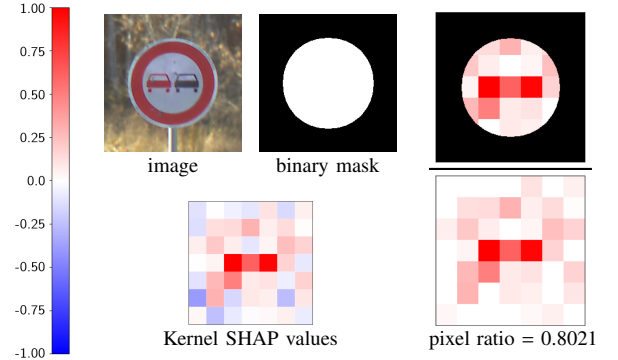
[9]captum.ai

Fig. 2: Visual explanation of the pixel ratio calculation using Kernel SHAP.

and for frontal, because all network instances should be familiar with this camera perspective.

(II) We apply both methods only to the first 200 images of each class in the test datasets instead of all 600 images per class, which we denote by $^1/_3\,CF_{test}$.

Furthermore, we exemplarily provide for KS the mean pixel ratios calculated on the shape-based subsets CFC_{test}, CFT_{test}, and CFR_{test} and CNs to give an impression of the shape-based differences. The KS findings are listed in Tab. IV, for GC in Tab. V respectively. For better interpretability, the results were column-wise colored blue (min value) over green to white (max value).

When comparing uncorrelated to correlated across the architectures (cf. Tab. IV left (KS) and Tab. V (GC)), the DNNs trained on uncorrelated data predominantly achieve higher pixel ratios, meaning they are less focused on the background than their correlated counterparts. The only exceptions are for both methods RN50 evaluated on high and for GC CNs evaluated on frontal camera variation stage. In all other cases, the pixel ratios for training on uncorrelated are between 0.0049 (min) and 0.0257 (max) for KS and 0.0038 (min) and 0.1284 (max) for GC greater than those of training on correlated data. Although this trend might not seem strongly pronounced, the

| | | Eval Network Architectures | | | | | | Eval CNs on Shapes | | | | | |
| | | CNs$_{100}$ on $\frac{1}{3}$ CF$_{\text{test}}$ | | CNt$_{100}$ on $\frac{1}{3}$ CF$_{\text{test}}$ | | RN50$_{100}$ on $\frac{1}{3}$ CF$_{\text{test}}$ | | CNs$_{100}$ on $\frac{1}{3}$ CFC$_{\text{test}}$ | | CNs$_{100}$ on $\frac{1}{3}$ CFT$_{\text{test}}$ | | CNs$_{100}$ on $\frac{1}{3}$ CFR$_{\text{test}}$ | |
		μ	CI$_{95\%}$	μ	CI$_{95\%}$	μ	CI$_{95\%}$	μ	CI$_{95\%}$	μ	CI$_{95\%}$	μ	CI$_{95\%}$
Training Dataset	UF$_{\text{train}}$	0.6589	0.0032	0.6365	0.0033	0.7418	0.0033	0.6685	0.0020	0.6236	0.0017	0.7381	0.0022
	CF$_{\text{train}}$	0.6333	0.0035	0.6191	0.0035	0.7312	0.0031	0.6472	0.0021	0.5889	0.0020	0.7200	0.0024
	UM$_{\text{train}}$	0.6536	0.0032	0.6364	0.0033	0.7362	0.0031	0.6674	0.0020	0.6160	0.0017	0.7298	0.0022
	CM$_{\text{train}}$	0.6326	0.0034	0.6247	0.0035	0.7303	0.0031	0.6466	0.0021	0.5825	0.0020	0.7224	0.0024
	UH$_{\text{train}}$	0.6437	0.0031	0.6426	0.0032	0.7374	0.0030	0.6511	0.0020	0.6116	0.0017	0.7191	0.0022
	CH$_{\text{train}}$	0.6387	0.0033	0.6288	0.0034	0.7405	0.0029	0.6518	0.0021	0.5920	0.0019	0.7229	0.0023

| | | Eval Network Architectures | | | | | |
| | | CNs$_{100}$ on $\frac{1}{3}$ CF$_{\text{test}}$ | | CNt$_{100}$ on $\frac{1}{3}$ CF$_{\text{test}}$ | | RN50$_{100}$ on $\frac{1}{3}$ CF$_{\text{test}}$ | |
		μ	CI$_{95\%}$	μ	CI$_{95\%}$	μ	CI$_{95\%}$
Training Dataset	UF$_{\text{train}}$	0,5466	0.0060	0,8153	0.0048	0,9185	0.0029
	CF$_{\text{train}}$	0,5873	0.0070	0,6869	0.0062	0,9042	0.0030
	UM$_{\text{train}}$	0,6269	0.0061	0,8162	0.0045	0,9145	0.0030
	CM$_{\text{train}}$	0,5881	0.0069	0,7008	0.0065	0,9201	0.0028
	UH$_{\text{train}}$	0,5609	0.0056	0,8440	0.0040	0,9202	0.0028
	CH$_{\text{train}}$	0,5109	0.0068	0,7666	0.0050	0,9164	0.0028

| | Eval on $^{1}/_3$ CFX_{test} | | | | | |
| | CNs$_{100}$ | | CNt$_{100}$ | | RN50$_{100}$ | |
Train ▼	μ	CI$_{95\%}$	μ	CI$_{95\%}$	μ	CI$_{95\%}$
CMC$_{\text{train}}$	0.6885	0.0021	0.6757	0.0020	0.7933	0.0015
	+0.0419		*+0.0330*		*+0.0272*	
CMT$_{\text{train}}$	0.6565	0.0019	0.6439	0.0018	0.7540	0.0020
	+0.0740		*+0.0801*		*+0.0637*	
CMR$_{\text{train}}$	0.7528	0.0019	0.7492	0.0020	0.8403	0.0018
	+0.0304		*+0.0303*		*+0.0475*	

consideration of the 95 % confidence intervals (CI), where the largest is ±0.0035 (KS) / ±0.0070 (GC), indicates it to be significant. With regard to differences between the architectures, it can be observed that with a rising number of trainable parameters, the average pixel ratio decreases, so the background attention rises. However, in comparison to the CN architectures, RN50 shows only small differences between training on correlated and uncorrelated data, which indicates that this architecture is not able to draw many helpful classification clues from the background. In contrast to our assumptions, no clear tendency can be identified concerning the different stages of camera variation. The break down according to shapes (cf. Tab. IV right) shows rectangular signs to have the highest mean pixel ratio, followed by circular signs, while triangular signs result in the lowest values. To ensure this not to be a consequence of the chosen FA resolution, we also evaluated this part with a FA resolution of 14 × 14 and 28 × 28 (not given in table due to limited space). With higher resolutions, the mean pixel ratio drops, but the observed trend remains.

As second step, we repeat the pixel ratio evaluation for the three network instances per architecture which were exclusively trained on traffic signs of the same shape (circular, triangular, and rectangular). The results are presented in Tab. VI. Compared to the pixel ratio mean of CN$_{100}$(CM$_{\text{train}}$) evaluated on CFX_{test} (cf. Tab. IV), the means notably increased. This implies that the CNs trained on traffic signs of only one shape are clearly less focused on the background features for their classification. In summary, this experiment thus provides the insight of the task definition having a comparably high impact on background feature importance.

C. Background Correlation Effect on Classification

For investigating the effect of background correlation on the classification performance, we conducted two additional training runs to the existing ones with differing random seeds for choosing the initial network weights, to be able to report the mean and standard deviation of three runs overall. The results are provided in Tab. VII.

The results show that the CNs architecture is able to benefit from correlated data: Independently of the stage of camera variation, the instance trained on correlated data is always able to achieve the highest top-1 accuracy evaluated on the correlated test set. However, with increasing camera variation also the difference between the highest (white) and lowest (blue) top-1 accuracy increases. For the camera variation stages F and H, the worst result is obtained by the networks trained on correlated, but tested on uncorrelated data. This may be attributed to the fact that mixed-up backgrounds are likely

TABLE VII

Top-1 accuracy for training on each of the 6 generated synthetic datasets, evaluated on the respectively corresponding correlated and uncorrelated (counterpart) test set for the best as well as the 100th epoch. For better interpretability, the results were colored by 4×4 squares from blue (min value) over green to white (max value). Please note that the coloring is based on the full precision values. The reported mean and standard deviations are based on three training runs each, with different random seeds.

| | | ConvNeXt-Small | | | | ConvNeXt-Tiny | | | | ResNet50 | | | |
| | | Best Epoch | | 100th Epoch | | Best Epoch | | 100th Epoch | | Best Epoch | | 100th Epoch | |
		UX_{test}	CX_{test}	UX_{test}	CX_{test}	UX_{test}	CX_{test}	UX_{test}	CX_{test}	UX_{test}	CX_{test}	UX_{test}	CX_{test}
Training Dataset	UF_{train}	99.83 ±0.007	99.87 ±0.020	99.83 ±0.010	99.88 ±0.015	99.80 ±0.007	99.85 ±0.014	99.80 ±0.007	99.85 ±0.012	99.77 ±0.015	99.80 ±0.007	99.77 ±0.012	99.81 ±0.005
	CF_{train}	99.80 ±0.012	99.88 ±0.009	99.80 ±0.006	99.89 ±0.004	99.75 ±0.018	99.85 ±0.018	99.75 ±0.008	99.86 ±0.010	99.73 ±0.005	99.82 ±0.023	99.74 ±0.008	99.83 ±0.009
	UM_{train}	99.82 ±0.014	99.81 ±0.027	99.84 ±0.005	99.84 ±0.006	99.79 ±0.026	99.81 ±0.012	99.82 ±0.011	99.82 ±0.014	99.76 ±0.012	99.74 ±0.012	99.77 ±0.008	99.75 ±0.010
	CM_{train}	99.81 ±0.008	99.85 ±0.004	99.81 ±0.011	99.85 ±0.004	99.77 ±0.007	99.83 ±0.007	99.78 ±0.005	99.83 ±0.013	99.72 ±0.022	99.77 ±0.033	99.73 ±0.010	99.78 ±0.011
	UH_{train}	99.39 ±0.010	99.33 ±0.020	99.38 ±0.021	99.34 ±0.017	99.32 ±0.010	99.29 ±0.000	99.31 ±0.012	99.28 ±0.003	99.22 ±0.023	99.15 ±0.017	99.21 ±0.022	99.15 ±0.016
	CH_{train}	99.32 ±0.005	99.37 ±0.018	99.30 ±0.008	99.35 ±0.003	99.27 ±0.012	99.31 ±0.028	99.24 ±0.026	99.31 ±0.025	99.20 ±0.021	99.17 ±0.041	99.20 ±0.018	99.19 ±0.018

to confuse the network, which should have learned a correlated context. An exception is the camera variation stage M, where CNs(UM_{train}) evaluated on CM_{test} achieved the worst result. For the instances trained on uncorrelated data, it holds that their performance is always better or at least equal when tested on the uncorrelated test set in comparison to the correlated one.

For the remaining architectures CNt and RN50, the observed trends are less pronounced. In ten out of twelve cases, the instances trained on correlated data perform worst when evaluated on uncorrelated data. This implies that predominantly instances trained on correlated data perform better when tested on correlated data; in seven out of twelve cases, they perform best. Furthermore, it can be noticed that the differences in top-1 accuracy between networks trained on uncorrelated data and tested on uncorrelated vs. correlated data are in some cases notably higher than observed for CNs. The fact that the results of RN50(XH_{train}) trained for 100 epochs deviate from the trends shown by the other instances in the KS experiments as well as in the classification performance experiments might indicate that overfitting occurred in this case.

VI. Conclusion and Outlook

In this work, we measured and systematically analyzed the effect of background on the feature importance and classification performance for the use case of traffic sign recognition. Therefore, we generated six synthetic datasets based on the synthesis pipeline of the Synset Signset Germany dataset [5], which differ with respect to their background correlation and stages of camera variation. We evaluated the datasets concerning their degree of background feature importance by determining the mean pixel ratio of Kernel SHAPs and classification performance when used as training data represented by the top-1 accuracy.

To conclude this work, we look back at our initial question of whether feature importance-based XAI reliably distinguishes between true learning and problematic overfitting with regard to our three formulated and investigated hypotheses on influencing factors of background attention:

(I) *The correlation of background:* Our results show a significant trend of correlated training data leading to the network giving a higher importance to background features.

(II) *Stages of camera variation:* Our findings do not support the hypothesis that higher camera variation results in more background attention.

(III) *Traffic sign shapes:* A higher variety of traffic sign shapes included in the classification task results in an increased importance of background features, as recognizing the shape against the background is a discriminative feature.

Contrary to the general assumption of lower background attention being an attribute of a well-performing, healthy classifier, our performance evaluation showed especially the modern ConvNeXt-Small architecture, which has the largest number of trainable parameters compared to the investigated architectures, being able to benefit from background correlation when tested on data of the same domain despite higher background attention. The results also demonstrate that background attention may increase through other factors, namely as it serves to distinguish the shape of foreground objects, which again is not indicative of deficient training.

The presented study has only shed light on the problem with regard to one use case. To widen the view, it would certainly be necessary to investigate more use cases, a greater variety of DNN architectures, and a larger selection of saliency methods. Nevertheless, some general findings can be identified from this work and applied to other use cases:

Our study shows synthetic data to be a helpful tool for investigating XAI metrics with respect to particular data properties, provided that they exhibit a sufficient degree of realism. In future work, it might be beneficial to include more corner cases in the data (for the use case of traffic sign recognition, e.g., rain, snow, greater overexposure, ...) to increase the dataset difficulty. This could further highlight performance differences.

References

[1] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *2012 IEEE Conference on Com-*

puter Vision and Pattern Recognition (CVPR). IEEE, 2012, pp. 3642–3649.

[2] S. Genovesi, M. Haimerl et al., "Evaluating Dimensions of AI Transparency: A Comparative Study of Standards, Guidelines, and the EU AI Act," in Symposium on Scaling AI Assessments (SAIA 2024). Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2025.

[3] S. Lundberg, "A unified approach to interpreting model predictions," arXiv preprint arXiv:1705.07874, 2017.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[5] 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.

[6] J. Stodt, C. Reich, and N. Clarke, "A novel metric for xai evaluation incorporating pixel analysis and distance measurement," in 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2023, pp. 1–9.

[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] S. I. Nikolenko, "Synthetic data for deep learning," 2019. [Online]. Available: https://arxiv.org/abs/1909.11512

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual Worlds as Proxy for Multi-Object Tracking Analysis," in CVPR, 2016.

[16] 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.

[17] Y. Cabon, N. Murray, and M. Humenberger, "Virtual KITTI 2," arXiv preprint arXiv:2001.10773, 2020.

[18] 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.

[19] 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.

[20] H. K. Ekbatani, O. Pujol, and S. Segui, "Synthetic data generation for deep learning in counting pedestrians." in ICPRAM, 2017, pp. 318–323.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] A. Arias-Duart, F. Parés, D. Garcia-Gasulla, and V. Gimenez-Abalos, "Focus! rating xai methods and finding biases," in 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, 2022, pp. 1–8.

[27] A. Binder, L. Weber, S. Lapuschkin, G. Montavon, K.-R. Müller, and W. Samek, "Shortcomings of top-down randomization-based sanity checks for evaluations of deep neural network explanations," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 16 143–16 152.

[28] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," ACM Computing Surveys, vol. 55, no. 13s, pp. 1–42, 2023.

[29] B. Fresz, L. Lörcher, and M. Huber, "Classification metrics for image explanations: Towards building reliable xai-evaluations," in The 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 1–19.

[30] L. Coroama and A. Groza, "Evaluation metrics in explainable artificial intelligence (xai)," in International conference on advanced research in technologies, information, innovation and sustainability. Springer, 2022, pp. 401–413.

[31] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence," Information fusion, vol. 99, p. 101805, 2023.

[32] R. Hesse, S. Schaub-Meyer, and S. Roth, "Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2023, pp. 3981–3991.

[33] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," arXiv preprint arXiv:1703.03717, 2017.

[34] J. Oramas, K. Wang, and T. Tuytelaars, "Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks," arXiv preprint arXiv:1712.06302, 2017.

[35] 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), Jun. 2022, pp. 11 976–11 986, feichtenhofer, Christoph and Darrell, Trevor and Xie, Saining.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[37] M. Contributors, "Openmmlab's pre-training toolbox and benchmark," https://github.com/open-mmlab/mmpretrain, 2023.

[38] 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.

[39] 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.