 DOI: 10.3217/978-3-85125-869-1-02

An Evaluation of the Machine Readability of Traffic Sign Pictograms using Synthetic Data Sets*

Alexander Maletzky[1], Stefan Thumfart[1] and Christoph Wruß[2]

***Abstract*— We compare the machine readability of pictograms found on Austrian and German traffic signs. To that end, we train classification models on synthetic data sets and evaluate their classification accuracy in a controlled setting. In particular, we focus on differences between currently deployed pictograms in the two countries, and a set of new pictograms designed to increase human readability. We find that machine-learning models generalize poorly to data sets with pictogram designs they have not been trained on, and conclude that manufacturers of advanced driver-assistance systems (ADAS) must take special care to properly address small visual differences between different traffic sign pictogram designs. Our main contributions are the creation of a vast synthetic data set of traffic sign images, training and evaluating state-of-the-art classification models to assess the machine readability of different pictogram designs, and employing techniques from explainable AI to analyze which image regions are particularly important to the classifiers.**

I. INTRODUCTION

In recent years, the number of semi-autonomous vehicles and advanced driver-assistance systems (ADAS) on our streets has been growing steadily. Even if there are still a lot of problems to be resolved before machines can eventually take over entirely, certain aspects of driving have been successfully automated already. One of them is *traffic sign recognition*, which consists of *detecting* and *classifying* traffic signs in video frames produced by a forward-facing camera. The results of this recognition process can then be used to automatically control the speed of the vehicle, or to display the found traffic signs on the instrument panel to inform the driver about them. In either case, correctly recognizing the traffic signs is of paramount importance for avoiding potentially fatal accidents. In the long-term future human-readable traffic signs will maybe disappear entirely, but in the current mixed-traffic regime machines must still be able to recognize traffic signs tailored to human needs.

State-of-the-art convolutional neural networks (CNNs) achieve near- or even super-human performance in many computer vision benchmark tasks, including traffic sign recognition [25]. However, as prior works illustrates, they may at the same time fail to correctly classify input that slightly deviates from the training distribution [36], [34], [15], [12], [7]. In our experiments we seek to find out whether and how this observation applies to *traffic sign classification models* under varying *pictogram designs*. Concretely, we pose the following questions: (i) Are there significant differences concerning the machine readability of different pictogram designs? In particular, we compare the current Austrian and German designs, as well as a proposed new Austrian design. (ii) How well do models generalize from one design to a new, unseen design? (iii) Which image details and regions are particularly important to classification models, and can this information be used to derive design rules that improve machine readability?

For answering these questions we trained traffic sign classifiers on a vast *synthetic data set*. The reason why we used synthetic- rather than real-world data is twofold: (i) A fair, systematic comparison of the machine readability of different pictogram designs is difficult to realize on real-world data with inherent differences *besides* the actual pictogram design. (ii) No real-world data exists for the proposed new Austrian design. The latter point is of particular importance, because whenever a traffic sign (pictogram) design is replaced by a new one, existing ADAS must be tested on and possibly adapted to the new design despite the lack of real-world training data. Our work demonstrates how this can be accomplished with carefully-crafted synthetic data sets.

An extended version of this paper can be found on arXiv [27].

A. Related Work

There exists a large body of scientific work regarding the automatic detection and classification of traffic signs in real-world as well as synthetic data sets. One of the most widely used real-world data sets is the German Traffic Sign Recognition Benchmark (GTSRB) [35] for classifying small image patches extracted from traffic scenes into one of 43 classes. Similar data sets exist for traffic signs from other countries and territories [29], [16], [43], [37], [40], [18], [8], [24], [42], [31], [14].

Closer to the kind of data sets employed in our experiments are partly synthetic data sets, where photographs or video frames of full traffic scenes are augmented with ultra-realistic weather effects [33], [10], [41]. In addition to these partly synthetic data sets there also exist data bases of fully synthetic 3D renderings of traffic scenes under varying (weather) conditions [4], [32], [2], [3], [5]. All these data sets have in common that they are better suited for object *detection* tasks, though. In [39], [28] and [9], real-world traffic scenes and signs are systematically modified by adding weather effects and other types of corruptions, to evaluate

*This research was funded by FFG (Austrian Research Promotion Agency) under grant 879320 (SafeSign) and supported by the strategic economic research programme "Innovatives OÖ 2020" of the province of Upper Austria.

[1]A. Maletzky and S. Thumfart are with RISC Software GmbH, 4232 Hagenberg, Austria. alexander.maletzky@risc-software.at

[2]Ch. Wruß is with ASFINAG Service GmbH, 1230 Vienna, Austria

Fig. 1: Overview of the experimental setup. Starting from three sets of traffic sign pictograms (each of a different design) a large collection of embedded and corrupted images (pictogram + traffic sign + background) is created. These images are then used to train classification models. Comparing the performance of the models allows to draw conclusions about the machine readability of the initial pictograms.

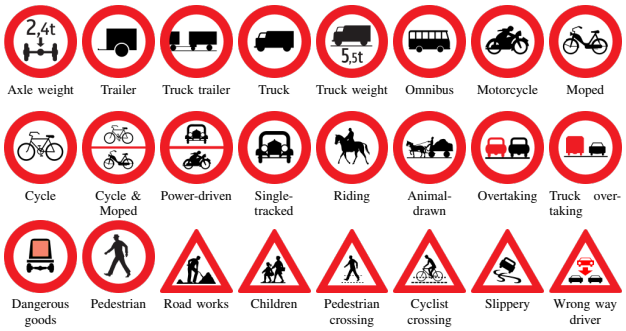


TABLE I: List of the 24 selected traffic sign classes, in the current Austrian pictogram design.

how well traffic sign detectors/classifiers work under such 'challenging conditions'. On the one hand, this resembles the approach we take in our experiments, but on the other hand, the main goal of the cited works is to compare different corruption types, not traffic signs or pictograms.

Similarly, in [20] a corrupted and perturbed version of ImageNet [11] is created. The methods employed there for corrupting images are similar to ours. ImageNet, however, is a general image data base without any particular focus on traffic signs, and the goal of [20] is to evaluate the performance of classification models in general, comparing models trained on 'clean' images to models trained on corrupted versions thereof.

II. METHODS

Fig. 1 presents an overview of the experimental setup. We first created synthetic data sets with traffic sign images and then trained classification models on them. Finally, we evaluated and compared the classification accuracy of these models.

A. Creating the Synthetic Data Sets

For creating the traffic sign images in the synthetic data sets, we started from high-resolution photographs of traffic scenes on Austrian highways in the year 2014 and extracted 14 patches with traffic signs. Seven of these 14 patches contain a prohibitory sign (round with red border), the other seven contain a warning sign (triangular with red border).[1] We then analyzed each of these 14 images w. r. t. color spectrum and perspective, obtaining parameters that allow to automatically replace the displayed pictogram by any given new pictogram in a way that makes the resulting image still look realistic. We then doubled the number of images by flipping them horizontally.

Next, we selected 24 traffic sign classes of the 'prohibitory' (18) and 'warning' (6) categories for our experiments. Pictograms of the current Austrian and German design could simply be downloaded from [6] and [1], respectively. Four of the 24 selected classes exist in Austria but do not have a German counterpart, meaning that we had to craft the corresponding pictograms manually by combining elements from other German pictograms. Pictograms of the proposed new Austrian design were kindly provided by their designer.[2] The complete list of classes is shown in Table I. Note that the selection of the 24 classes was mainly driven by the availability of a new Austrian pictogram design.

We replaced the pictograms in the 28 source images by the pictograms of the 24 selected classes, giving rise to a set of 336 images per pictogram design, with 14 images per class. We resized these images to a uniform size of 64×64 pixels. Finally, we augmented the set of 336 images by applying an arsenal of augmentation methods with varying intensities [21]. In particular, first one out of ten preselected corruption methods, like Gaussian noise, blurring, rain patterns, etc. is applied. Then, the resulting images are down-sampled by first down- and then up-scaling them, to decrease their spatial resolution but keep the size of 64×64 pixels; no extra smoothing is applied. The purpose of down-sampling is to simulate distance, as one of the key properties of well-designed traffic sign pictograms is being readable from large distances. We generated 250 variants for each of the 336 clean images, with five different levels of corruption intensity (50 per level). These intensities only affect the down-sampling factor, i. e., a higher intensity level gives rise to more 'pixelated' images.

Fig. 2 summarizes the whole data generation process. In the lower-left corner, two images per corruption intensity are shown, with intensity increasing from left to right. Eventually, every data set consists of 84,000 images, which are equally distributed across source patches (12,000 per patch), pictogram classes (3,500 per class) and corruption intensity (16,800 per intensity level). This, however, only corresponds to *one* data set, for one pictogram design. Repeating the process outlined above for each of the three designs yields three data sets with 252,000 images, where by construction identical corruptions are applied to the images of each design to enable an unbiased comparison. In order to obtain reliable results and reduce the impact of the randomness inherent to data augmentation on our results, we repeated the entire data generation process as well as the subsequent model training and evaluation three times and then averaged all results over these three independent runs. In total **756,000 images** were

[1] The source images and patches can be provided upon request.

[2] Stefan Egger, https://visys.pro/.

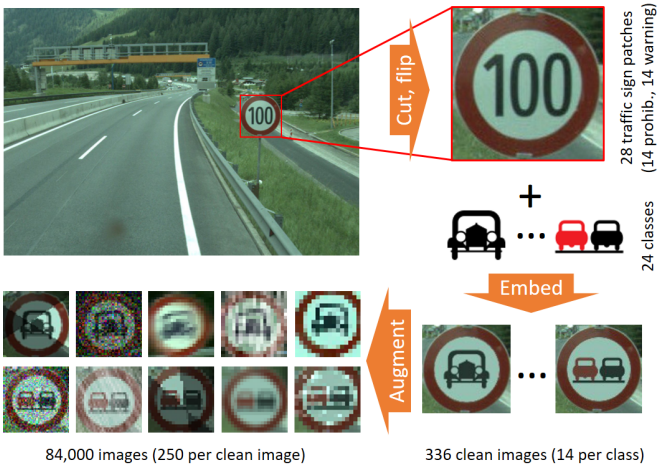


Fig. 2: Data generation process for the synthetic data sets used in our experiments. This process is repeated three times for current Austrian pictograms, proposed new Austrian pictograms, and current German pictograms, yielding nine data sets with a combined total of 756,000 images.

generated for our experiments.

For the sake of brevity, the data set with current Austrian pictogram design will be labeled AT_c, the one with the proposed new Austrian design will be labeled AT_n, and the one with the current German design will be labeled DE in the remainder. The combined data set with all currently deployed designs, i. e., the union of AT_c and DE, will be labeled CUR.

B. Model Training and Evaluation

The classification models were trained separately on each of the three pictogram designs (AT_c, AT_n, DE), as well as jointly on the two current designs (CUR). We considered two deep neural network architectures: a small ResNet architecture [19] with 20 layers and an input size of 64×64 pixels, and the architecture by Li and Wang [25] with an input size of 48×48 pixels. The latter was a natural choice for our experiments, since it represents the state-of-the-art on the GTSRB data set [35], with 99.66% test accuracy.

We split the data into training-, validation- and test sets and trained the models for 60 epochs, using the Adam optimizer [22] with an initial learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is reduced by 80% whenever the validation loss does not improve for ten epochs. In the end, the trained weights of the epoch with the smallest validation loss are taken. Both training- and validation accuracy plateau after only a few (< 10) epochs in each case, so training for a total of 60 epochs is certainly sufficient. The splits into the three sets are based on the 28 source patches all images ultimately originate from, and are identical for each pictogram design.

After training, all models are evaluated on the held-out test sets, using the overall classification accuracy as the main metric of interest. As is common practice, confusion between classes is treated uniformly. Putting less weight on confusion between semantically similar classes (e. g., 'Pedestrian crossing' and 'Cyclist crossing') could be an interesting direction for future research.

	Li-Wang [25]	ResNet [19]
AT_c-AT_c	98.89±0.11	98.48±0.35
AT_n-AT_n	98.68±0.17	98.45±0.09
DE-DE	98.85±0.17	98.23±0.56
CUR-CUR	98.69±0.06	98.28±0.10
AT_c-AT_n	80.18±0.97	77.76±3.97
AT_c-DE	83.94±0.92	80.43±2.88
AT_n-DE	75.33±0.31	74.24±1.07
DE-AT_c	82.03±1.69	77.26±0.26
DE-AT_n	77.35±1.52	72.82±0.82
CUR-AT_n	85.48±1.27	84.17±0.60

TABLE II: Test accuracy (%) of the models trained in our experiments, displayed as *mean*±*SD* over three runs.

First, every model is evaluated on its 'own' test set, i. e., with the same pictogram design as in the set it was trained on. Due to the uniform construction of training-, validation- and test sets, the performance scores thus obtained are feasible for comparing the quality of different models, even if they are trained and evaluated on different pictogram designs. Besides evaluation on the own test set, some models are evaluated on 'foreign' test sets with different pictogram designs, too, to find out how well they generalize to unseen designs. In the remainder, evaluations will be denoted by short identifiers like AT_c-AT_n, where the data set label before the dash indicates the design the model was *trained* on, and the data set label after the dash indicates the design it was *evaluated* on.

III. RESULTS

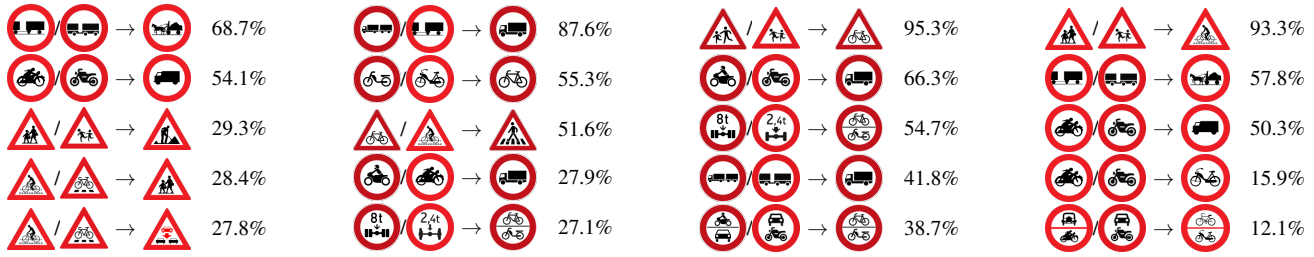
Table II shows the classification accuracy of all models. One can see that there is hardly any difference in the classification accuracy of the models between the three pictogram designs (top-three rows in Table II), and that the Li-Wang models generally tend to outperform the corresponding ResNet models by a small margin.

One can also see very clearly that the classification accuracy of every model drops significantly when evaluated on a 'foreign' test set, with different (albeit similar) pictograms. In fact, the difference between current and proposed new Austrian pictograms seems to be more pronounced than the difference between current Austrian and German pictograms. Models trained on German pictograms generalize only poorly to new Austrian pictograms, and vice versa; this is particularly interesting, since intuitively the design of the new Austrian pictograms resembles the German design much closer than the current Austrian design does, especially w. r. t. stroke width and level of detail.

A. Per-Class Results for Foreign Test Sets

We focus on the Li-Wang models [25] in the remainder of this section. The results of the ResNet models exhibit the same overall tendency as the Li-Wang models, including the frequently confused classes.

Table III lists the pairs of traffic sign classes the models confuse most often if the pictogram design differs from the



(a) AT_c-AT_n (b) DE-AT_c (c) DE-AT_n (d) CUR-AT_n

TABLE III: Frequent confusion of the models when evaluated on 'foreign' test sets. The numbers on the right are the percentages of samples belonging to the class on the left-hand-side of the arrows, which are misclassified as the class on the right-hand-side of the arrows. For better comparison, both training- and evaluation pictograms are shown on the left-hand-side of the arrows; in the case of CUR-AT_n, only current Austrian pictograms are shown, although German pictograms are part of the training set, too.

	AT_c-AT_n	DE-AT_c	DE-AT_n	CUR-AT_n
	11.9% (24)	1.1% (24)	15.7% (22)	26.9% (22)
	29.2% (22)	56.6% (21)	1.6% (24)	26.3% (23)
	59.4% (21)	85.3% (17)	2.8% (23)	3.8% (24)
	28.9% (23)	36.3% (23)	89.3% (15)	84.3% (20)
	64.5% (20)	85.8% (15)	48.3% (20)	81.7% (21)
	83.9% (15)	76.6% (19)	63.7% (19)	88.4% (19)
	87.5% (14)	82.9% (18)	95.6% (13)	95.5% (14)
	90.9% (13)	42.1% (22)	95.9% (11)	97.0% (10)
	98.8% (2)	71.3% (20)	38.9% (21)	97.9% (6)
	69.9% (19)	85.5% (16)	98.4% (2)	97.9% (5)

TABLE IV: Accuracy of selected classes. Numbers in parentheses denote the rank among all 24 classes. Even though only Austrian pictograms are shown in the table, all models are evaluated on the pictogram design indicated in the table header.

design they were trained on. Class 'Truck trailer' seems to cause most problems: DE-AT_c and DE-AT_n often confuse 'Truck trailer' with 'Truck'; AT_c-AT_n hardly ever confuses these two classes, but instead misclassifies 'Truck trailer' as 'Animal-drawn', such that in the total the accuracy of 'Truck trailer' drops as far as 1.1%, as can be seen in Table IV. It can also be seen that in all evaluations 'Motorcycle' is frequently misclassified as 'Truck', which might be owing to the three designs of 'Motorcycle' differing fairly strongly. An analogous statement applies to 'Power-driven'.

Table IV lists the per-class accuracy of the models for a couple of selected classes. Although the overall classification accuracy of all models drops considerably on foreign pictogram designs, there are blatant inter-class differences. In fact, a big deal of this drop is caused by only a few classes, namely those listed in Table IV; the others are correctly classified most of the time.

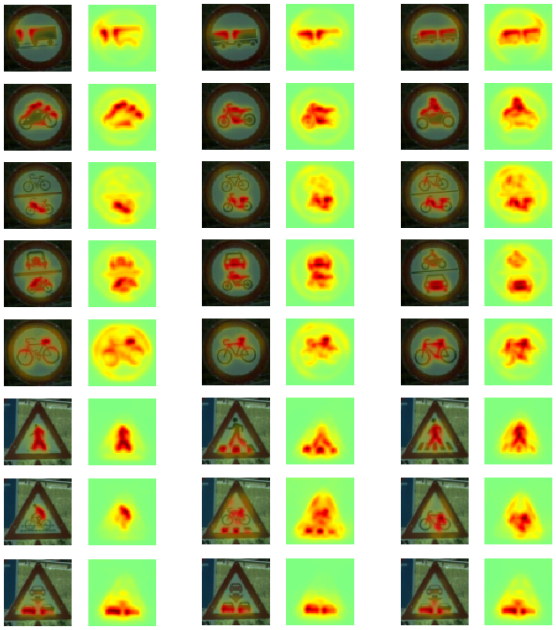
B. *Qualitative Explanations of the Models' Predictions*

We employed *layer-wise relevance propagation* (LRP) [30] for estimating the importance of image regions and -details to the classification models, in order to explain what information they base their predictions on. Among the multitude of possible parameter configurations of LRP we adhered to the one suggested for convolutional neural networks in [23] throughout.

Fig. 3 shows the average explanations of all correctly predicted test images of some selected classes, for the AT_c-AT_c, AT_n-AT_n and DE-DE experiments. Explanations are presented as heatmaps, where the color of a pixel indicates its relevance to the model. For each evaluation, the left column blends the heatmaps with the actual images to facilitate localization, whereas the right column only shows the heatmaps themselves. Evaluations CUR-AT_c and CUR-DE are spared since they exhibit a very similar relevance pattern as AT_c-AT_c and DE-DE, respectively.

As can be seen, all models strongly focus on the pictograms (or parts of them) when classifying a traffic sign image, and only sometimes take the border of the sign into account as well. On the one hand, this means that our models learned to pay attention to the 'right' details of an image and do not base their decisions on spurious artifacts in the background, and on the other hand, it means that the shape of the traffic signs does not really aid the models. This is not surprising, since the uniform circular and triangular shapes carry only little information for classifying the signs – especially if the pictograms alone are sufficient for that purpose. Only in some cases, where the pictograms of prohibitory and warning signs are similar in appearance, taking the shape into account can be beneficial. This phenomenon occurs, for example, with classes 'Cycle' and 'Cyclist crossing': for the models trained on current Austrian pictograms the shape of 'Cycle' seems to be quite important, whereas the other models pay more attention to the shape of 'Cyclist crossing'.

In classes 'Pedestrian crossing' and, to some extent, 'Cyclist crossing', the models trained on proposed new Austrian and German pictograms focus a lot on the zebra crossing



(a) AT_c-AT_c (b) AT_n-AT_n (c) DE-DE

Fig. 3: Average explanations of all correctly predicted test images of some selected classes. Reddish to yellowish hues indicate regions with evidence *in favor* of the predicted class and greenish hues indicate regions without any relevance.

at the bottom. The models trained on the current Austrian pictograms, on the other hand, completely ignore the (only barely visible) zebra crossing and instead focus on the person. This difference in attention might be one of the reasons why in AT_c-AT_n 'Cyclist crossing' and 'Pedestrian crossing' only achieve a comparatively low accuracy of 28.93% and 69.87%, respectively (cf. Table IV).

It can also be observed that sometimes the models only look at certain parts of the pictograms, and not at the whole pictograms. This effect is particularly visible in class 'Wrong way driver', where all models completely ignore the car at the top and almost entirely ignore the arrow as well. Apparently, the two cars at the bottom are sufficient for robustly distinguishing this class from the other 23 traffic sign classes in our experiments. Likewise, in class 'Cycle & Moped' the moped receives a lot more attention than the cycle, especially in AT_c-AT_c. Interestingly, in class 'Power-driven', whose pictogram is similarly split into an upper and a lower part, the relevance is distributed much more evenly across the car and the motorcycle.

Classes 'Truck trailer' and 'Motorcycle' allow us to speculate why the models fail to generalize to other pictogram designs in some cases. Namely, both classes differ between the three design groups in certain aspects the models pay a lot attention to. The fact that the truck in 'Truck trailer' is visible in its entirety in the German design seems to be important to the models, since quite some relevance is assigned to the front part of the truck. When comparing the two Austrian designs of this class one can also observe a subtle difference in the relevance pattern: only a small part of the truck is visible in the current Austrian design, leading to a vertical relevance pattern; in contrast, the proposed new Austrian design displays a slightly larger part of the truck, leading to a more horizontal pattern. Similarly, the fact that the current Austrian and German designs feature a person riding the motorcycle in class 'Motorcycle' seems to be important to the models. The proposed new Austrian design lacks a rider, which the models seem to compensate by paying more attention to the front wheel.

It must be noted, though, that further experiments are necessary to confirm the hypotheses expressed in the preceding paragraphs. The relevance patterns constructed by LRP or any other feature attribution method are only meant to illustrate which parts of an image are important to a model, but one must be careful when trying to draw conclusions why the model fails to classify some class correctly.

IV. DISCUSSION

The objective of our work was to answer three research questions regarding the machine readability of traffic signs:

1) Is there any significant difference between different pictogram designs (AT_c, AT_n, DE) in terms of machine readability?
2) Can traffic sign classification models trained on one pictogram design be safely deployed to traffic signs featuring a different design?
3) Can general 'design rules' for pictograms be formulated to improve machine readability?[3]

The first question can be answered readily: even though there *are* small differences in the observed model accuracies in AT_c-AT_c, AT_n-AT_n and DE-DE (cf. Table II), these differences are not significant. Hence, all three pictogram designs are equally well machine-readable.

The answer to the second question is also negative: if any of the models trained on one pictogram design is applied to a different design, its classification accuracy drops significantly, by about 15–23 percentage points. In this regard, it is particularly interesting to note that a few classes cause massive problems, whereas most of the others can still be classified accurately. Even more surprising is the fact that our models consistently generalize best between German and *current* Austrian pictograms (in both directions), although a human would probably find more similarities between German and *new* Austrian pictograms. We do not have any explanation for this phenomenon. Models trained on CUR generalize better to the unseen design AT_n than models trained on AT_c and DE alone. However, even here the performance drop of more than 13 percentage points, from 98.69% accuracy in CUR-CUR to 85.48% in CUR-AT_n, is significant. Hence, training on a more diverse set of pictogram designs leads to some improvement, but is still far from optimal. A larger study involving even more pictogram designs remains a possible subject for future research.

[3]Adding something like QR-codes would be an obvious affirmative answer, but in this work we focus on traffic infrastructure that can be processed by machines and humans alike.

Answering the third question is more intricate. Although we generated explanations for the models' predictions in Section III-B, formulating design rules for pictograms based on them is difficult. Still, what *can* be said is that deep neural networks perceive traffic signs differently than humans: humans try to *understand* the meaning of pictograms in order to classify them, machines only try to *distinguish* them. This is characteristic for discriminative methods, as it is exactly what they are meant to do. Distinguishing a fixed set of pictograms, however, might be possible based on small, semantically meaningless details. We hypothesize that this is the main reason why the models in our experiment fail to generalize to 'foreign' pictogram designs. Unfortunately, it is hardly possible to predict a-priori which details will be important to a classification model. The only general rule that can be formulated in this regard concerns the visibility of pictogram elements: the zebra crossings in classes 'Cyclist crossing' and 'Pedestrian crossing' of the current Austrian design consist of small, thin line segments that quickly become imperceptible when the images are corrupted, and hence the models do not pay attention to them. In the proposed new Austrian design the zebra crossings are far more pronounced and thus better visible, and the models *do* take them into account. Thin lines and overly small patches of ink should therefore be avoided.

Summarizing, the main takeaways of our work are as follows:

- Machines can handle different pictogram designs equally well, provided they have been trained on them.
- In the realistic scenario that an ADAS should correctly recognize traffics signs with different pictogram designs, the models must be trained appropriately. This can be achieved by either training one single classifier on a data set encompassing many different designs or by training a separate classifier for each design.
- If existing pictograms are replaced by a new design, classification models will likely have to be updated. Since acquiring large real-world data sets is time-consuming and only possible once the actual traffic signs have been replaced, it might be necessary to resort to synthetic data sets as presented in this paper, instead.

V. Limitations and Future Work

When defining our experimental setup, we had to fix certain parameter values that are up to discussion and could be revised in future extensions of our experiments. First, we only considered 24 traffic sign classes from categories 'prohibitory' and 'warning'. Actual classifiers deployed in ADAS must be trained on a much wider variety of classes and may hence exhibit a different behavior w. r. t. sensitivity to pictogram design, frequently confused classes, and attention patterns. Still, we believe that our reduced setting approximates reality sufficiently well for making our findings hold more generally. A similar statement applies to the investigated model architectures. Extending the experiments to more architectures, like Vision Transformers [13], for obtaining more reliable results is certainly possible. Furthermore, traffic sign recognition systems deployed in ADAS are not disclosed to the scientific community, so one could question whether our findings are even applicable to them. Indeed, we merely want to encourage developers of ADAS to consider our experimental results and, if appropriate, conduct similar experiments with their own traffic sign classifiers. Yet, we think that the highly similar results of two fairly distinct architectures present strong evidence that our findings are not limited to the concrete architectures under consideration.

Another point of discussion concerns the image corruption strategy. From the vast space of conceivable corruption methods we picked some that we deemed either realistic or particularly interesting, but many others would have been at our disposal, too. In future experiments, one could in particular try to incorporate corruptions that are specific to traffic signs, like some kind of 'over-exposure' where, due to the production process and reflectivity of the traffic sign foil, brighter areas seem to 'grow' and hide parts of darker neighboring areas, making small and fine pictogram elements seemingly disappear. Furthermore, we focused on simulating distance by spatially downsampling the images at varying degrees, but we did not apply other geometric transformations like rotations and perspective distortions. In addition to the degree of downsampling, one could systematically vary the intensities of the 'secondary' corruptions (rain, noise, blur, etc.) as well. A complementary augmentation strategy could specifically target the pictograms themselves, for instance, by systematically (re)moving vertices in vectorial versions of the pictograms.

As discussed above, the models we obtained are not very robust w. r. t. 'foreign' pictogram designs. One way to counter this could be forcing the models to pay more attention to the global shape of the pictograms, instead of small details. This, in turn, can perhaps be achieved by borrowing ideas from current research on *adversarial attacks* [36], [17], like *adversarial training* [26], [38]. Alternatively, one could also try to preprocess the images before training and applying a model, by applying a low-pass- or bilateral filter that destroys high-frequency information and thereby biases the model towards low-frequency shape information. Repeating our experiments with adversarially trained models or said input preprocessing could be an interesting direction for future research.

Finally, it would be interesting to see how well the models trained on our purely synthetic data would perform on real-world data, like GTSRB. This could serve as sort of a 'sanity check' to ensure that the synthetic data sets resemble reality sufficiently well. Of the 24 classes considered in our experiments only seven are included in GTSRB, though, rendering an exhaustive evaluation impossible.

Acknowledgment

We thank Stefan Egger for providing us with the proposed new Austrian pictogram designs and for his suggestions regarding our experimental setup. We also thank Isabell Ganitzer for carefully proofreading a draft version of this paper, and the anonymous reviewers for their valuable comments.

REFERENCES

[1] "VzKat 2017," http://www.vzkat.de, 2017, online.

[2] "AI.Reverie," https://aireverie.com/, 2021, online.

[3] "Anyverse," https://anyverse.ai/, 2021, online.

[4] "Cognata Traffic Sign Datasets," https://www.cognata.com/traffic-sign-datasets/, 2021, online.

[5] "CVEDIA," https://www.cvedia.com/, 2021, online.

[6] Austrian Federal Ministry for Digital and Economic Affairs, "Straßenverkehrsordnung 1960, Fassung vom 12.11.2018," https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10011336, §§ 50 and 52, 2018, online.

[7] A. Azulay and Y. Weiss, "Why do deep convolutional networks generalize so poorly to small image transformations?" *Journal of Machine Learning Research*, vol. 20, no. 184, pp. 1–25, 2019.

[8] R. Belaroussi, P. Foucher, J.-P. Tarel, B. Soheilian, P. Charbonnier, and N. Paparoditis, "Road sign detection in images: A case study," in *International Conference on Pattern Recognition*, 2010, pp. 484–488.

[9] C. Berghoff, P. Bielik, M. Neu, P. Tsankov, and A. von Twickel, "Robustness testing of AI systems: A case study for traffic sign recognition," in *Artificial Intelligence Applications and Innovations*, ser. IFIP Advances in Information and Communication Technology, I. Maglogiannis, J. Macintyre, and L. Iliadis, Eds. Springer, 2021, pp. 256–267.

[10] A. v. Bernuth, G. Volk, and O. Bringmann, "Simulating photo-realistic snow and fog on existing images for enhanced CNN training and evaluation," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 41–46.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[12] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2020.

[14] C. Ertler, J. Mislej, T. Ollmann, L. Porzi, G. Neuhold, and Y. Kuang, "The mapillary traffic sign dataset for detection and classification on a global scale," in *Computer Vision – ECCV 2020*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Springer, 2020, pp. 68–84.

[15] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 7538–7550.

[16] C. Gámez Serna and Y. Ruichek, "Classification of traffic signs: The european dataset," *IEEE Access*, vol. 6, pp. 78 136–78 148, 2018, conference Name: IEEE Access.

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[18] 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.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[20] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations (ICLR)*, 2019.

[21] A. B. Jung, K. Wada, J. Crall, S. Tanaka, J. Graving, C. Reinders, S. Yadav, J. Banerjee, G. Vecsei, A. Kraft, Z. Rui, J. Borovec, C. Vallentin, S. Zhydenko, K. Pfeiffer, B. Cook, I. Fernández, F.-M. De Rainville, C.-H. Weng, A. Ayala-Acevedo, R. Meudec, M. Laporte, *et al.*, "imgaug," 2020. [Online]. Available: https://github.com/aleju/imgaug

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[23] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lapuschkin, "Towards best practice in explaining neural network decisions with lrp," in *International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.

[24] F. Larsson and M. Felsberg, "Using fourier descriptors and spatial models for traffic sign recognition," in *Image Analysis*, ser. Lecture Notes in Computer Science, A. Heyden and F. Kahl, Eds. Springer, 2011, pp. 238–249.

[25] J. Li and Z. Wang, "Real-time traffic sign recognition based on efficient CNNs in the wild," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 975–984, 2019.

[26] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018.

[27] A. Maletzky, S. Thumfart, and C. Wruß, "Comparing the machine readability of traffic sign pictograms in Austria and Germany," *arXiv:2109.02362 [cs.CV]*, 2021.

[28] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel, "Benchmarking robustness in object detection: Autonomous driving when winter is coming," *arXiv:1907.07484 [cs, stat]*, 2019.

[29] A. Møgelmose, 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.

[30] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: An overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ser. Lecture Notes in Computer Science, vol. 11700, 2019, pp. 193–209.

[31] A. L. Pavlov, P. A. Karpyshev, G. V. Ovchinnikov, I. V. Oseledets, and D. Tsetserukou, "IceVisionSet: lossless video dataset collected on russian winter roads with traffic sign annotations," in *International Conference on Robotics and Automation (ICRA)*, 2019, pp. 9597–9602, ISSN: 2577-087X.

[32] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3234–3243.

[33] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.

[34] C. Sitawarin, A. N. Bhagoji, A. Mosenia, M. Chiang, and P. Mittal, "DARTS: Deceiving autonomous cars with toxic signs," *arXiv:1802.06430 [cs]*, 2018.

[35] 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.

[36] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[37] D. Tabernik and D. Skočaj, "Deep learning for large-scale traffic-sign detection and recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1427–1440, 2020.

[38] J. Tack, S. Yu, J. Jeong, M. Kim, S. J. Hwang, and J. Shin, "Consistency regularization for adversarial robustness," *arXiv:2103.04623 [cs]*, 2021.

[39] D. Temel, M.-H. Chen, and G. AlRegib, "Traffic sign detection under challenging conditions: A deeper look into performance variations and spectral characteristics," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2019.

[40] R. Timofte, K. Zimmermann, and L. Van Gool, "Multi-view traffic sign detection, recognition, and 3d localisation," *Machine Vision and Applications*, vol. 25, no. 3, pp. 633–647, 2014.

[41] G. Volk, S. Müller, A. v. Bernuth, D. Hospach, and O. Bringmann, "Towards robust CNN-based object detection through augmentation with synthetic rain variations," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 285–292.

[42] Y. Yang, H. Luo, H. Xu, and F. Wu, "Towards real-time traffic sign detection and classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 2022–2031, 2016.

[43] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2110–2118.