2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

A novel Traffic Sign Dataset with Condition Annotations

Hanni Sandhu
Camera-based ADAS & ITS
IAV GmbH
Berlin, Germany
hanni.sandhu@iav.de

Joana Kühne
Camera-based ADAS & ITS
IAV GmbH
Berlin, Germany
joana.kuehne@iav.de

Oliver Sawade
Camera-based ADAS & ITS
IAV GmbH
Berlin, Germany
oliver.sawade@iav.de

Martin Stellmacher
Camera-based ADAS & ITS
IAV GmbH
Berlin, Germany
martin.stellmacher@iav.de

Elmar Matthes
Camera-based ADAS & ITS
IAV GmbH
Berlin, Germany
elmar.matthes@iav.de

Olaf Hellwich
Computer Vision & Remote Sensing
Technical University Berlin
Berlin, Germany
olaf.hellwich@tu-berlin.de

***Abstract*—To develop robust and secure automated transportation systems, Traffic Sign Detection and Recognition (TSDR) is a key part. It plays a crucial role in Advanced Driver Assistance Systems (ADAS), self-driving vehicles and traffic safety. However, the task of TSDR can be challenging due to traffic signs being subject to damages, discoloration, vandalism and occlusion. Even though a lot of progress is made in both research areas of Traffic Sign Detection (TSD) and Traffic Sign Recognition (TSR), no study explicitly deals with the problem of qualitative poor traffic signs appearing in real-world scenarios. This can be assigned to the lack of an extensive traffic sign dataset containing flawless signs as well as imperfect signs. Neural networks trained exclusively on untainted data might fail at detecting flawed signs as they occur in real-world scenarios. Therefore, in this paper, a novel traffic sign dataset with condition annotations is proposed, indicating if a sign is good, discolored, vandalized, dirty or occluded. The custom dataset is created with a semi-supervised approach, in which machine learning models are trained to classify traffic signs in the condition categories. The resulting dataset can be used as basis for more precise traffic sign recognition as well as traffic sign condition classification which can be useful for maintenance planning. The dataset includes approx. 20.000 images of 10 sign classes, where 70% of data is incorporated in the training set, 10% in the validation set and 20% in the test set.**

***Index Terms*—Deep Learning, Image Processing, Transportation and Vehicle Systems**

I. Introduction

Robust and reliable Traffic Sign Detection and Recognition (TSDR) is integrated in today's Advanced Driver Assistance System (ADAS) applications to notify the driver of any upcoming signs and their meaning, ensuring overall road safety and allowing the driver to be a little more at ease while driving on new roads [1, 2]. With the advent of higher automated ADAS and autonomous driving, TSDR becomes even more crucial for safe vehicle operation. However, the task of TSDR can be very challenging because of the varying condition of traffic signs on the roads. Traffic signs can be exposed to discoloration, vandalism, dirt and occlusion, which can hinder the recognition process of sign classes.

Over the years, many studies have been conducted in the research area of TSDR and as an outcome various datasets [3, 4] and deep-learning methods [5, 6] have been proposed. To develop accurate and robust algorithms for TSDR, traffic sign datasets are an essential component. Several publicly available datasets have been created, such as GTSRB (German Traffic Sign Recognition Benchmark) dataset [7], CURE-TSD (Challenging Unreal and Real Environments for Traffic Sign Detection) dataset [3] and Mapillary Traffic Sign Dataset (MTSD) [4], all primarily including images and videos of traffic signs in good visible states and some captured under different weather conditions. However, most of these datasets only provide a small number of images, incomplete annotation information, no dedicated test or validation dataset and most importantly largely do not include vandalized signs [3]. It can be assumed that training a neural network with these datasets would most probably fail in detecting and recognizing traffic signs, which are in poor condition. Since the reliability of autonomous driving systems depends on the reliability of the core technologies that process and analyze sensed information and damaged, faded and vandalized traffic signs occur in the road network, vehicles incorporating such trained Artificial Neural Networks (ANNs) might lead to incorrect movement in the simulated and real-world traffic scenario. Therefore, it is of high importance to train and validate machine learning models also with imperfect signs and additionally determine the condition of traffic signs and replace damaged ones at an early stage to ensure road safety.

Since according to the author's knowledge, there exists no publicly available approach dealing with the problem of traffic signs being in bad shape, the herein presented work discloses a traffic sign dataset with semi-supervised condition annotations that can be used to train and test ANNs for traffic sign condition recognition and further maintenance planning. Furthermore, the study aims to enable more precise TSR by providing the dataset to train and validate traffic sign

978-0-7381-4408-5/23/$31.00 ©2023 IEEE

classification models with high and low quality signs.

The paper makes the following contributions:

- Proposing a traffic sign dataset with sign type and condition annotations[1].
- Providing several classification models with condition annotations. The most precise model achieves an accuracy of 91%.
- Analyzing the performance of TSD and TSR networks when encountering signs of poor quality.

II. Related Work

To train ANNs within the scope of TSD or TSR, traffic sign datasets are required. Stallkamp *et al.* [7] introduced the German Traffic Sign Recognition Benchmark (GTSRB) dataset in 2011 which was created from a video of 10 hours length recorded in Germany during daytime. The dataset provides labels for 43 different sign types, including the distinction between 8 speed limits. The GTSRB dataset generally incorporates 144.769 images, however only 51.840 frames are annotated.

Based on the GTSRB dataset, the German Traffic Sign Detection Benchmark (GTSDB) dataset [8] was built in 2013 and annotates each detected sign in an image with a bounding box and the respective sign type. The dataset includes 900 street images, in which a total of 1.206 signs of 43 sign types are annotated.

Ertler *et al.* [4] proposed the Mapillary Traffic Sign Dataset (MTSD) in 2020, including images and traffic sign classes of six continents with the following percental distribution of images: 20% North America, 20% Europe, 20% Asia, 15% South America, 15% Oceania and 10% Africa. The images are extracted from the street-level imagery platform Mapillary. The dataset holds 100.000 high-resolution images in total, from which 52.000 are fully annotated through manual labeling and 48.000 partially annotated using computer vision technology. MTSD covers 400 known traffic sign classes and other unknown classes.

Temel et al. *et al.* [3] presented the CURE-TSD (Challenging Unreal and Real Environments for Traffic Sign Detection) dataset 2019. The authors focused on images recorded under different challenging conditions, mostly related to weather, and labeled all signs with their respective sign type, challenge type and severity. It is distinguished between 14 sign classes, 12 challenge types and 5 challenge levels, indicating the difficulty in visibility. Challenge level 1 does not affect the visibility of traffic signs from a human perspective and level 5 makes the visibility of small and distant traffic signs nearly impossible. There are 1.72 million images from 5.733 video sequences incorporated, all being labeled.

Furthermore, a lot of studies regarding TSR and TSD have been published over the past years. Wong *et al.* [6] have constructed MicronNet, a highly compact deep convolutional neural network for real-time embedded traffic sign recognition and achieve a top-1 accuracy of 98.9% on the GTSRB dataset.

[1]Custom dataset: https://tubcloud.tu-berlin.de/s/GSwNdQCT4W3smdy

Recently, Sabbir *et al.* [5] proposed a CNN based TSDR framework which can detect traffic signs under different weather conditions. A CNN-based challenge classifier identifies the weather condition in which the image is recorded and forwards it to the Enhance-Net, which enhances the image using an encoder-decoder CNN architecture. Sign detection and classification is performed on the enhanced image using two separate CNN architectures with an overall achieved precision of 91.1%.

Rajendran *et al.* [2] propose a traffic sign recognition system based on YOLOv3. The YOLOv3 detector identifies candidate traffic signs, a bounding box pre-processor enlarges, crops and resizes the detected bounding boxes and feeds them to a CNN-based traffic sign classifier. The classifier is trained on the GTSRB dataset. The YOLOv3 based detector achieves a high accuracy of 92.2% on the GTSDB dataset.

Although various datasets and studies have been disclosed, no research group explicitly focused on imperfect signs appearing in real-world and the influence of such signs on the model performance.

III. Dataset Generation

The task of developing a traffic sign dataset consists mainly of three parts: image acquisition, image annotation and dataset evaluation.

A. Image acquisition

At first, whole image sequences and traffic sign crops were extracted from Mapillary, a street-level imagery platform by Facebook. Within the range of the minimum and maximum latitude and longitude coordinates of Germany, sign crops were derived directly by Mapillary's integrated TSDR functionality. In the process, the search of traffic sign images was limited to 10 sign types, visualized and listed in Fig. 1 with their class ID and name.

Fig. 1: Overview of sign types

B. Image annotation

The image annotation process is performed semi-supervised, where a part of the collected data is at first manually annotated to train and validate several neural networks and subsequently, the best model is used to label all the remaining data. The signs should be labeled into the categories: good, discolored, vandalized, dirty or occluded. An example of each condition is shown in Fig. 2.

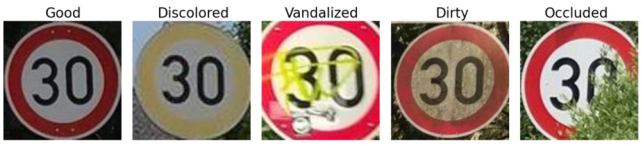


Fig. 2: Condition labels

1) Models: To successfully annotate the condition of all sign crops in a semi-supervised approach, four state of the art architectures are utilized, all being pretrained on the ImageNet dataset. The following four neural networks achieving the smallest error rates on the ImageNet test set are trained and validated with the custom dataset: *ResNeXt-101-32x8d* [9], *Wide ResNet-101-2* [10], *Wide ResNet-50-2* [10] and *ResNet-152* [11]. To train these neural networks with the custom dataset, transfer learning is leveraged, where the network is initialized with pretrained weights and then training is continued with the custom dataset. To adapt the model for the present application, a linear output layer with five neurons is added corresponding to the number of classes.

2) Training process: To train and validate the classification models with condition labels, the training set is constructed with approx. 2.500 images per category (in total 12.500 images), the validation set with approx. 100 images per category (in total 500 images) and the test set with approx. 200 images per category (in total 1.000 images). All crops were manually labeled. Since not many samples were available for training of the categories dirty and occluded, data augmentations are performed on the images of these two categories to enlarge the training data and consequently obtain better model performance. In the scope of that, the following data augmentation techniques were applied: vertical and horizontal flipping, padding and rotations with different angles.

After creating all three datasets, the respective model is trained and evaluated with the training and validation set over 500 epochs and the model state with the best validation accuracy is saved. Finally, inference is conducted with the best model weights on the test set.

To successfully train the classification models, the following parameters are set in the training pipeline:

- batch size: 64
- maximum number of epochs: 500
- learning rate: 0.001 with a stepwise learning rate decay with a factor of 0.1 every 7 epochs
- optimizer: Adam
- applied data augmentations on the training set:
  - Resize each image to 96×96 pixels
  - Gaussian blur with kernel size of (5, 9) and sigma of (0.1, 5)
  - Randomly adjust sharpness by factor 2
  - Randomly rotate by degrees (0, 180)
  - Randomly flip horizontally with probability 0.5

3) Semi-automated image annotation: After training and evaluating the different classification models, 17.044 unlabeled sign crops are automatically annotated with the condition labels good, discolored, vandalized, dirty and occluded using the most precise classifier. Finally, it is manually visually verified whether the signs are categorized correctly.

C. Construction of final dataset and dataset evaluation

The final dataset is created by splitting all annotated data into a training set, validation set and test set with the ratios of 70%, 10% and 20%, respectively.

Lastly, the usage of the generated dataset is evaluated. Since TSD and TSR are important features in ADAS systems and therefore should work precisely error-free, it is examined how TSD and TSR models perform when confronted with signs of poor quality. This step is crucial since among enabling traffic sign condition recognition and thus simplifying the task of traffic maintenance planning, the custom dataset should provide a better base for training and evaluation of machine learning models and lead to more accurate TSR by also classifying the sign type of discolored, vandalized, dirty and occluded signs without any difficulties.

1) Traffic Sign Recognition (TSR): To examine whether the sign type of signs being in a poor condition can be classified correctly, two classification models are used to predict the sign type of the final custom test set containing approx. 3.900 sign crops.

The first model is the Wide ResNet-101-2, which is trained with the final custom dataset. The second model is a neural network with three convolutional layers with batch normalization and ELU activation, a dropout layer after each layer and one linear layer at the end, outputting 43 features. The network is pretrained with the GTSRB dataset, which does not contain the sign class crosswalk which therefore cannot be classified.

2) Traffic Sign Detection (TSD): To test whether traffic signs of all five categories can be detected by object detectors, TSD is performed with a pretrained YOLOv5 model and a Faster R-CNN model over approx. 3.000 street images. Since it is only focused on sign crops in the dataset generation, there is a limited amount of whole street images gathered.

Among different pretrained Faster R-CNN models, the pretrained MobileletNetV3 Large Faster R-CNN model is chosen since it gives a good balance between accuracy and speed. The model is trained with the GTSDB dataset. The YOLOv5 model is loaded from different public repositories [12, 13], where initial training was performed with the COCO dataset and training continued with small custom datasets.

IV. Results

A. Performance comparison

Table I provides an overview of the performance of all four models for sign condition classification. The table holds the average validation and test loss, accuracy, F1-Score and ROC AUC (Area under the ROC (Receiver Operating Characteristic) curve) value when conducting multiclass classification. In addition, the number of parameters and number of MAC (Multiply–accumulate operations) of each respective model are stated for a deeper understanding of the results. The term

MAC incorporates a count of all operations done in the model by each layer.

The best performance is accomplished by the two Wide ResNet models, both reaching similar loss and AUC values and the same test accuracy of 0.911, being the highest among all models. Wide-ResNet-50 yields lower loss values and Wide-ResNet-101 higher AUC scores. Based on the higher AUC score, and since the computational resources are not a limiting resource for the dataset generation, Wide-ResNet-101 is chosen for the subsequent semi-automated annotation.

B. Test results of Wide ResNet-101-2

In the following, the predictions conducted on the test set with the trained Wide ResNet-101-2 are displayed in more detail.

Fig. 3 shows that the model classifies samples of the categories good and vandalized the most accurate and misclassifies the most signs of the condition dirty, where 11% of dirty signs are categorized as vandalized. This might correspond to the fact that these signs have dirt spots on them which are recognized as stickers by the classifier, leading to the classification as vandalized.

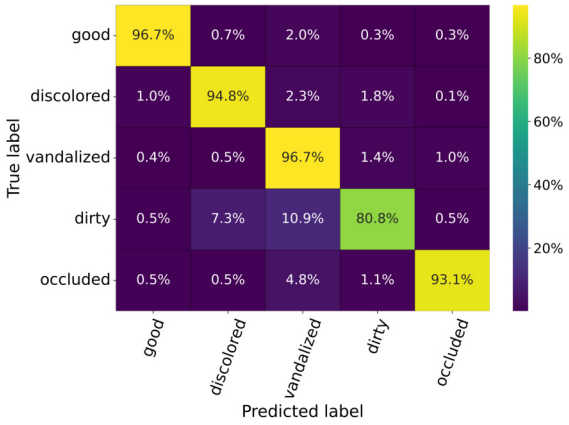


Fig. 3: Test confusion matrix: traffic sign condition recognition

C. Annotation of unlabeled data

The previous evaluations provide an accurate model for classification of the sign condition which can now be used to create the large customary dataset in a semi-supervised fashion. For that, 17.044 unlabeled sign crops are annotated with the condition labels good, discolored, vandalized, dirty and occluded by using the trained Wide ResNet-101-2 model. Afterwards, incorrect labels are corrected manually. In the context of manual labeling adjustments, Fig. 4 is generated, displaying the confusion matrix to evaluate the accuracy of the semi-automated annotation process. Generally, only 697 out of 17.044 signs, approx. 4%, are annotated incorrectly. It can be observed that the least wrong predictions are made with regard to good, discolored and vandalized signs, where only approx. 1% of good signs are not recognized as good by the model, being the largest category with approx. 10.000 images to be labeled. 40% of dirty signs and 22% of occluded signs are identified inaccurately, from which a large part of signs is classified as good because these signs are only dirty or occluded by a small extent and therefore are preferably recognized as being good by the classifier.

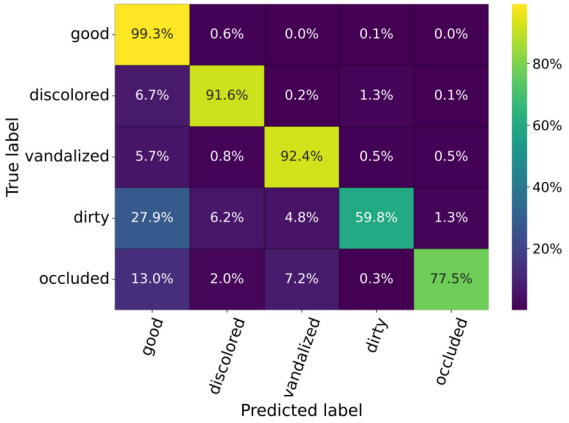


Fig. 4: Automatic labeling of sign condition with the best classification model (Wide ResNet-101-2)

D. Distribution of final dataset

The general distribution of crops in the training, validation and test set is outlined in Table II and is further presented in Fig. 5 in detail for each sign class and all five categories. The crop size (height, width) ranges between (20, 18) and (534, 580). During dataset splitting, it was tried to include a fair amount of crops of each sign class in every set. The training and validation set are balanced with respect to the condition categories, while the test set contains approx. 1.000 sign crops of the conditions good, discolored and vandalized and approx. 400 crops of the categories occluded and dirty.

Dataset	Percentage	Crops per category	Total count
training	70%	2.800	14.061
validation	10%	400	2.009
test	20%	400 - 1.000	3.893

TABLE II: Structure of final dataset

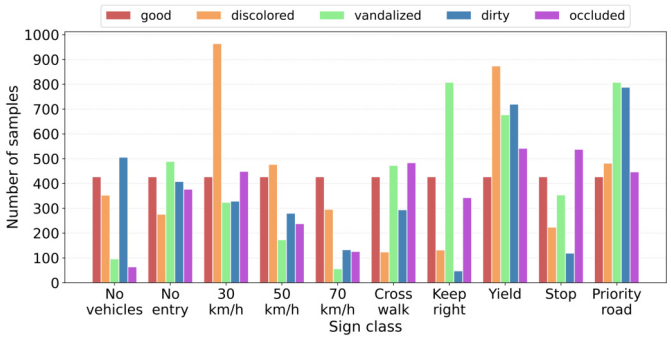


Fig. 5: Distribution of crops in each sign class and all five categories included in the dataset

Model	Loss	Accuracy	F1-Score	ROC AUC	Num of params	Num of MAC
ResNeXt-101-32x8d	Val: 0.326 Test: 0.345	Val: 0.939 Test: 0.907	Val: 0.939 Test: 0.907	Val: 0.893 Test: 0.860	86M	194B
Wide ResNet-101-2	Val: 0.275 Test: 0.304	Val: 0.937 Test: 0.911	Val: 0.937 Test: 0.911	Val: 0.891 Test: 0.910	124M	268B
Wide ResNet-50-2	Val: 0.282 Test: 0.280	Val: 0.939 Test: 0.911	Val: 0.939 Test: 0.911	Val: 0.884 Test: 0.890	66M	134B
ResNet-152	Val: 0.259 Test: 0.346	Val: 0.939 Test: 0.899	Val: 0.939 Test: 0.899	Val: 0.867 Test: 0.867	58M	136B

TABLE I: Performance comparison of different models on validation and test set for traffic sign condition classification

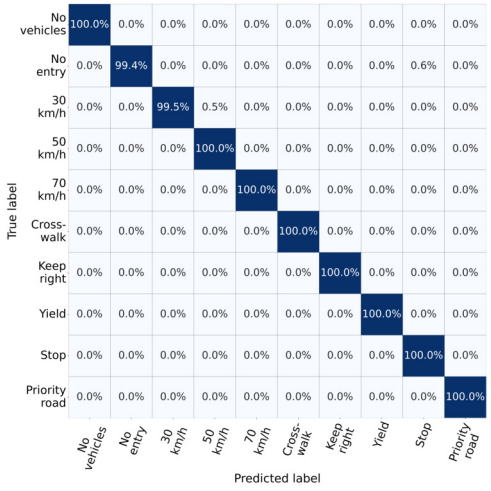
E. Traffic Sign Recognition (TSR)

In the following, the results of performing TSR on good, discolored, vandalized, dirty and occluded signs are presented. It is outlined how accurate TSR models can classify the sign type of signs being in various conditions by recognizing the sign class of all 3.900 sign crops included in the final custom test set.

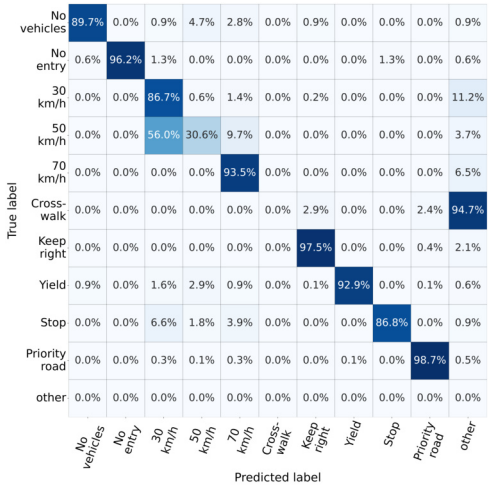
In Fig. 6a, the result of the Wide ResNet-101-2 model trained with the custom dataset is visualized. The test confusion matrix depicts that nearly all signs are classified correctly except for 4 out of 3.893 crops, approx. 0.1%. It can be concluded that the Wide ResNet-101-2 recognizes the sign type of imperfect signs properly when it is trained with high and low quality signs.

The test confusion matrix of the model pretrained with the GTSRB dataset is presented in Fig. 6b. The category other stands for any other sign class besides the sign classes existing in the customary created dataset. 331 out of 3.893 sign crops, approx. 8.5%, are labeled incorrectly, excluding the sign class crosswalk since it is not present in the GTSRB dataset.

The model trained on the custom dataset performs more accurately than the one trained on the GTSRB dataset, however, it has to be considered that the custom dataset includes more training samples per category and focuses on 10 sign classes. The GTSRB dataset involves 43 sign classes. More classes can lead to a higher difficulty level and a greater probability of incorrect predictions. Nevertheless, the outcome indicates how outdated and small the GTSRB dataset is, including 51.840 images and approx. 200 to 1.900 crops per sign class and presenting that a deep learning model trained with the custom dataset can lead to a more accurate TSR.



(a) Wide ResNet-101-2 model trained with custom dataset



(b) Pretrained GTSRB model (crosswalk class not in GTSRB dataset)

Fig. 6: Performing TSR on final custom test set

F. Traffic Sign Detection (TSD)

To examine the performance of object detectors over images containing signs of poor quality, TSD is conducted with a pretrained YOLOv5 and Faster R-CNN model over approx. 3.000 street images. The results of TSD are visualized in Fig. 7. Since Faster R-CNN is trained with the GTSDB dataset and the dataset does not contain the sign class crosswalk, all results of Faster R-CNN are presented disregarding the sign class crosswalk.

It can be observed that Faster R-CNN outperforms YOLOv5 in all categories. The Faster R-CNN model spots almost all good and discolored signs. Both object detectors have the most difficulties in detecting occluded signs. The least accurate performance of YOLOv5 models in all bad

conditions can be explained by means of the fact that the models are mostly not specifically trained on detecting traffic signs, but also recognize, e.g., traffic lights, cars, bicycles and trains since the object detector is initially trained with the COCO dataset.

Furthermore, the class probability in each bounding box, reflecting the probability that the detected object belongs to a particular class, can be analyzed. The confidence score is higher for good and discolored signs, varying between 0.74 and 0.96. It is in average the lowest for vandalized and occluded signs, ranging between 0.29 and 0.82, which indicates the uncertainty of the object detectors in detecting such traffic signs.

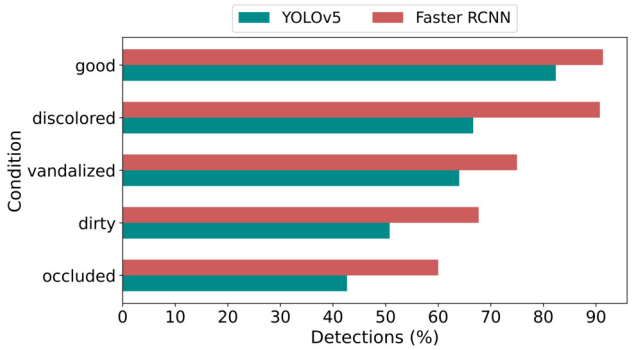


Fig. 7: Performing TSD with YOLOv5 and Faster R-CNN

V. Conclusion

In the scope of this paper, a traffic sign dataset with semi-supervised condition annotations is generated that includes approx. 20.000 images and covers 10 sign classes. A condition annotation indicates whether the sign is good, discolored, vandalized, dirty or occluded. The dataset is created to enable more precise TSR by using the dataset to train and validate traffic sign classification models with signs of good and several bad states. Furthermore, the dataset can be utilized specifically to train and test deep learning models for traffic sign condition recognition and thereby simplify the task of traffic maintenance.

As future work, the most accurate model trained with the custom dataset could be used in a real-world scenario with a TSD system to verify that the condition of traffic signs can be identified correctly and that the sign type of signs being in a bad state can be classified precisely. Furthermore, since all sign crops, included in the custom dataset, are collected by using the TSDR functionality of Mapillary, it is not guaranteed that all signs of poor quality occurring in Germany are included in the custom dataset. Signs of sufficiently poor quality might not have been detected at all by the detector. In the future, the dataset could be enlarged with more signs of very poor quality, which could be gathered through other effective methods such as synthetic data or controlled sign deterioration in lab environments. Moreover, the dataset could be extended by more sign types and sign classes of other countries. Another important step would be the incorporation of complete scenery images with traffic signs of poor quality as a basis for the improvement of TSD models.

References

[1] Aashrith Vennelakanti et al. “Traffic sign detection and recognition using a CNN ensemble”. *2019 IEEE international conference on consumer electronics (ICCE)*. IEEE. 2019, pp. 1–4.

[2] Shehan P Rajendran et al. “Real-time traffic sign recognition using YOLOv3 based detector”. *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE. 2019, pp. 1–7.

[3] Dogancan Temel et al. “Challenging environments for traffic sign detection: Reliability assessment under inclement conditions”. *arXiv preprint arXiv:1902.06857* (2019).

[4] Christian Ertler et al. “The mapillary traffic sign dataset for detection and classification on a global scale”. *European Conference on Computer Vision*. Springer. 2020, pp. 68–84.

[5] Sabbir Ahmed, Uday Kamal, and Md Kamrul Hasan. “DFR-TSD: A deep learning based framework for robust traffic sign detection under challenging weather conditions”. *IEEE Transactions on Intelligent Transportation Systems* (2021).

[6] Alexander Wong, Mohammad Javad Shafiee, and Michael St Jules. “MicronNet: a highly compact deep convolutional neural network architecture for real-time embedded traffic sign classification”. *IEEE Access* 6 (2018), pp. 59803–59810.

[7] Johannes Stallkamp et al. “The German traffic sign recognition benchmark: a multi-class classification competition”. *The 2011 international joint conference on neural networks*. IEEE. 2011, pp. 1453–1460.

[8] Sebastian Houben et al. “Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark”. *The 2013 international joint conference on neural networks (IJCNN)*. Ieee. 2013, pp. 1–8.

[9] Saining Xie et al. “Aggregated residual transformations for deep neural networks”. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1492–1500.

[10] Sergey Zagoruyko and Nikos Komodakis. “Wide residual networks”. *arXiv preprint arXiv:1605.07146* (2016).

[11] Kaiming He et al. “Deep residual learning for image recognition”. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[12] Anant Mishra. *Road-sign-detection*. https://github.com/Anant-mishra1729/Road-sign-detection. 2022.

[13] Yutong. *vehicle-and-traffic-sign-detection*. https://github.com/Yutong-gannis/vehicle-and-traffic-sign-detection-depended-on-yolov5-and-cascade-classifier. 2022.