Fakultät Wirtschaftswissenschaften

Masterthesis

im Studiengang Wirtschaftsinformatik

zur Erlangung des akademischen Grades
Master of Science (M.Sc.)

Traffic sign classification in an open-set scenario: An experimental study

vorgelegt von:
Manuel Schenk

Ausgabedatum: 01.04.2022
Abgabedatum: 04.10.2022

Erstgutachter: Prof. Dr. Andreas Theissler
Zweitgutachter: Prof. Dr. Manfred Rössle

Table of contents

List of Figures

List of Tables

Abstract

Numerous publications deal with the automatic classification of traffic signs and achieve excellent results, mainly by using convolutional neural networks (CNN). Largely neglected, however, is that a traffic sign classification system deployed in the real world inevitably encounters objects that are not known from training. These can be unknown traffic signs and objects that have similarities due to their shape or color. Therefore, this study compares different open-set recognition (OSR) methods using a self-created dataset. It contains 143,663 images of German traffic signs and similar objects from the real world. The approach that achieves the best open-set and closed-set performance is based on a vision transformer. When a state-of-the-art backbone is utilized for the approaches based on CNNs, the adversarial reciprocal point learning framework achieves competitive open-set performance, and a conventional CNN attains the best closed-set performance.

1 Introduction

Automatic traffic sign recognition (TSR) is an important task with numerous applications. These include autonomous driving, driver assistance systems, as well as traffic sign inventory and maintenance [1]. For almost four decades [2], TSR has been an active research area in the computer vision community [3, 4, 5] and is usually divided into at least two phases: Object detection or traffic sign detection (TSD) and image classification or traffic sign classification (TSC). TSD aims to find traffic signs in an input image showing a road scene. TSC, which is the focus of this work, has the task of determining the type of the traffic sign previously detected [6].

Most publications in the field of TSC use the German Traffic Sign Recognition Benchmark (GTSRB) dataset from 2011, which contains more than 50,000 traffic sign images from 43 classes with different distances, illuminations, weather conditions, partial occlusions, and rotations [7]. Due to the large dataset, convolutional neural networks achieve the best results [8]. For example, accuracies of 99.15 [9], 99.17 [10], 99.46 [11], 99.65 [12], and 99.71 [13] are achieved. In recent years, works have also emerged aiming to improve TSC's everyday usability. These address, for example, its application in real-time [4, 14] or on mobile devices with limited resources [15], as well as the robustness to natural environments such as ice and snow [16]. However, other works show that TSR in everyday use can be vulnerable to adversarial attacks, which can lead to misclassifications [17, 18]. This poses a security risk, especially for autonomous vehicles. In addition, false positives cannot yet be completely avoided in the TSD phase. These can be billboards, car tail lights, tires, satellite dishes or other objects resembling traffic signs due to color or shape [4, 19, 20]. As a result, a TSC system employed under real-world conditions encounters objects that are not traffic signs.

Furthermore, in Germany alone, there are more than 600 different traffic sign types [21], of which the GTSRB dataset covers only 43. If a traffic sign classification system is to be used internationally, extensive datasets for multiple countries would be required. To date, 68 countries have ratified the Vienna Convention on Road Signs and Signals [22] for uniform signs. Nevertheless, there are variations in colors, pictograms, and text [23]. The freely available datasets of other countries, like the GTSRB, do not cover the full range of traffic signs [23]. Even if it is not necessary for an application to recognize all available traffic signs of one or multiple countries, the TSC will still encounter traffic signs in the real world that were not part of the training dataset. It is possible but very time-consuming to create a dataset with sufficient training data for all traffic sign types in one or more countries. However, collecting enough training images for all potential false positives from the TSD phase is not feasible. Therefore, a traffic sign classification system employed in the real world must be able to sort out unknown traffic sign types and traffic sign-like objects while correctly classifying known traffic sign classes. Some work has already addressed the described problem partially, most of them treating it as an out-of-distribution [24, 25, 26, 27, 28] problem. However, solely traffic sign classes not available during training or datasets from other domains are used as unknowns for testing, but no traffic sign-like objects.

In this study, the following research question is investigated: Which machine learning approach is best suited for classifying German traffic signs assuming an open-set scenario? The problem at hand is considered an open-set recognition (OSR) problem since this research area aims to correctly classify samples from known classes while identifying samples from unknown classes. Therefore, several state-of-the-art OSR methods are applied to the described problem and compared using a self-generated dataset.

The main contributions of this work are: (1) This is the first study that applies and compares several state-of-art OSR methods to the traffic sign classification problem under an open-set scenario where both traffic signs not known from training and traffic sign-like objects are used as unknown classes. (2) A new German traffic sign classification dataset with 143,663 images, based on the GTSRB dataset, from which 30,310 images were taken, is presented. It contains 66 traffic sign classes with training and testing data. Additionally, the test set includes traffic signs not present in the training set and a wide range of objects that are similar to traffic signs based on their shape or color.

This paper is structured as follows: Important open-set recognition methods, the current state-of-the-art in German traffic sign classification, and related works are described in the next section. Chapter 3 lists the selected OSR methods and describes implementation details as well as the evaluation procedure. Chapter 4 is devoted to the self-created traffic sign dataset, and chapter 5 summarizes and compares the performance of the OSR methods, which is subsequently discussed. Finally, limitations and potential future research are described, as well as a conclusion is drawn.

2 Research background

2.1 Open-set recognition

Open-set recognition (OSR), out-of-distribution (OOD), anomaly, novelty, and outlier detection all aim to detect test samples that fall outside the training data distribution. As stated in [24, 29, 30, 31, 32, 33, 34], OSR differs from its related research areas in that it aims not only to identify such outliers but also to classify known classes from training correctly. In a multi-class scenario, this can also apply to OOD detection [35], but the focus is often on detecting OOD samples rather than maintaining closed-set performance [32]. Unlike OSR, which divides a dataset into known and unknown classes, OOD detection typically uses one or more other datasets as OOD samples. Often, these come from different domains [36]. Therefore, this paper focuses solely on OSR approaches. Some important ones and those that achieve state-of-the-art performance on the OSR benchmarks are described in this chapter.

Since open-set recognition was formalized by [37], a number of works have been devoted to the topic in recent years. Most of them are collected in [35] and [38]. The first approaches are based on traditional machine learning mainly support vector machines [38]. Bendale and Boult [39] are the first to incorporate deep learning for OSR. They replace softmax with the proposed OpenMax layer, which utilizes the output of the penultimate layer of the network. This is referred to as the activation vector (AV) in the paper. For each class, an average activation vector (MAV) is computed based on the correctly classified training samples. Also the distance of each of these samples to the respective MAV is calculated. A Weibull distribution is then fitted to the largest distances between AV and MAV. This leads to a parameter, which estimates the likelihood that an input is an outlier with respect to the corresponding class. G-OpenMax [40] extends OpenMax by using a conditional GAN to generate synthetic samples from a mixture of the known class distribution. Those images are fed into a pretrained classifier which considers only known classes. The images that are not classified as one of the known mixed classes are included in the training set as samples for the unknown class. Neal et al. [41] propose another approach which generates open-set training samples by utilizing counterfactual image generation. An encoder-decoder GAN architecture is trained to encode training samples into a latent space and decode points from it into realistic images. Those fake examples are generated outside the distribution of the knowns but still inside the common latent space, assuming that known and unknown classes share the same latent space. The generated fake and the known samples are then used to train a classifier where the former represent the unknown class. Additionally, the authors are the first to measure OSR performance with the MNIST, SVHN, CIFAR, and TinyImagenet datasets. In more recent publications, these datasets are still used as benchmarks.

Three important reconstruction-based approaches for OSR are CROSR [42], C2AE [43], and CGDL [44]. Unlike OpenMax or G-OpenMax, CROSR [42] does not only use the network's final prediction for the classification of known and the detection of unknown classes. Contrary, a deep hierarchical reconstruction network is trained to produce a prediction y and a reconstructive latent representation z. The framework uses y for closed-set classification and both y and z for unknown detection. This way, a larger pool of features that may not be discriminative for known classes can be utilized by the unknown detector, and the information in y lost due to deep networks can be supplemented. Oza and Patel [43] employ class conditioned auto-encoders for their approach. An encoder and a classifier are initially trained using cross-entropy loss to address the closed-set problem. Subsequently, in the open-set training, the latent vectors extracted by the encoder are conditioned with so-called label condition vectors. The decoder has the task of reconstructing input perfectly, which is conditioned with a label condition vector that matches the class of an input. If the class does not match, the decoder is supposed to reconstruct the input poorly. In this way, the open set behavior is mimicked. The reconstruction errors are modeled using extreme value theory to find a threshold for unknown detection. During testing, the classification prediction and the reconstruction error are generated conditionally with each label condition vector. The minimum reconstruction error is then used for unknown detection. CGDL [44] addresses the shortcoming of a variational auto-encoder in that it is suitable for unknown detection but cannot provide discriminative representations for classification tasks because all features follow only one distribution. Hence, the proposed method generates class-conditional distributions in the latent space, which are forced to approximate different multivariate Gaussian models. During inference, a sample's extracted latent features and reconstruction error are used to perform unknown detection. For known classes, the latent representation is then fed into a softmax layer to perform closed-set classification. CGDL outperforms SoftMax, OpenMax, G-OpenMax, OSRCI, C2AE, and CROSR.

OpenHybrid, a framework proposed in [33], consists of three components. An encoder, a flow-based

model, and a classifier. During training, the encoder maps images into a latent feature space. Both the flow-based model and the classifier use the output of the encoder, the former for density estimation and the latter for classification learning using cross-entropy loss. During inference, the flow-based model estimates the probability density to distinguish between known and unknown samples and classifies known samples with the classifier. OpenHybrid significantly outperforms previous OSR work such as SoftMax, OpenMax, G-OpenMax, OSRCI, C2AE, and CROSR.

Since prototype learning (PL) was introduced by Yang et al. [45] to improve the robustness of CNNs, several works have applied the approach to open-set recognition [31, 45, 46, 47, 48]. Convolutional prototype network (CPN), a framework introduced in [45], addresses the weakness of conventional CNNs in an open-set setting to divide the entire feature space and assign these regions to known classes without leaving room for unknowns. In CPN, the CNN's feature extractor is retained, but the classifier is omitted in favor of prototypes that are learned during training, each representing a known class in the feature space. During inference, the extracted CNN features are matched with all prototypes. If there is no match, a test sample is classified as unknown. However, prototypes might locate themselves in the space of unknown classes, making it impossible to discriminate between known and unknown classes. Reciprocal point learning (RPL) [46] and its extension adversarial reciprocal point Learning (ARPL) [31] address this issue. The authors suggest that the potential unknown deep space should also be modeled in the training in addition to the known classes. To accomplish this, reciprocal points are used, which are essentially the opposite of prototypes, because the embedding features of a known class should be far from its respective reciprocal point. The objective is to confine unknown samples to an internal bounded embedding space and distribute the reciprocal points around the edge of it. During inference, the likelihood that a sample belongs to one of the known classes is proportional to the distance of the furthest reciprocal point. Test examples from unknown classes are nearer to all of the reciprocal points than instances from known classes. By creating confusing samples that represent potential unknown classes and are supposed to be equally distant from all reciprocal points, ARPL is extended with ARPL+CS. Xia et al. [47] confirm the performance of APRL in an open-set experiment but show that there are still a few unknown classes overlapping with known classes in feature space. Therefore, in their motorial prototype framework (MPF), they leave aside the concept of reciprocal points and reuse the idea of prototypes. MPF is optimized so that the embedding features are close to the corresponding prototype center. In addition, the so-called motorial margin constraint term is added to the loss function to compress the distribution range of the known classes. The authors doubt that MPF works for complex test conditions and therefore introduce AMPF and AMPF++. Both achieve improved performance on the OSR benchmark datasets. As with ARPL+CS, adversarial samples are generated and added to the training to represent unknowns in the embedding space. AMPF++ adds more self-generated data to the training than AMPF, which further improves OSR performance. Another framework that uses prototype learning to address OSR is called prototype mining and learning (PMAL) [48]. The paper claims that current PL approaches, which combine prototype learning and embedding optimization [31, 45, 46, 49], have weaknesses, particularly in complicated situations. This typically includes the undesirable learning of non-discriminative prototypes representing low-quality samples. Redundant prototypes and a lack of diversity within a class are further issues. Therefore, in the so-called prototype mining phase, a prototype set per class is first determined that considers both high quality and diversity. The optimization of the embedding space is only performed in a subsequent phase. PMAL outperforms the other PL approaches on the OSR benchmark datasets with this approach.

Most of the methods presented in this section perform much better for open-set recognition than a conventional CNN trained with cross-entropy loss, which rejects unknown samples based on the thresholded softmax output. Vaze et al. [50] study the correlation between the closed-set and open-set performance. They, however, demonstrate that the conventional CNN, referred to as Baseline in the OSR literature, can match or even exceed the more intricate state-of-the-art methods by improving its closed-set performance. This is accomplished by utilizing techniques like increased augmentation, improved learning rate schedules, label smoothing, and longer training time. Additionally, they suggest using the maximum logit score rather than the softmax probability to perform the open-set recognition task.

Recently, Cai et al. [51] introduced a novel approach for open-set recognition that makes use of a vision transformer (ViT) pretrained on the ImageNet-21K dataset. For the classification task on the closed set, it utilizes the original ViT architecture proposed in [52], which is trained in a first phase. An additional attached detection head is tasked to represent the known classes in compact clusters, for which a second training phase is performed. The decision of whether or not to reject samples as unknown during inference is made based on the distance between the learned cluster centers of the training classes and the extracted embedding of the respective test image. The authors claim that they achieve new state-of-the-art

performance on several of the common OSR benchmark datasets.

2.2 German traffic sign classification

Many publications in the field of traffic sign classification use the dataset of the German Traffic Sign Recognition Benchmark (GTSRB) [7]. It was part of a competition held at the International Joint Conference on Neural Networks (IJCNN) in 2011. The dataset contains more than 50,000 traffic sign images from 43 classes with different distances, illuminations, weather conditions, partial occlusions, and rotations. Due to the large dataset, approaches incorporating convolutional neural networks (CNN) achieve the best results [8].

Ciresan et al. [9] won the first phase of the GTSRB competition with an accuracy of 98.98 % using an ensemble of CNNs and multi-layer perceptrons. The latter are trained with precalculated HOG features provided by [7]. Sermanet and LeCun [10] finished second in the first stage of the competition achieving an accuracy of 98.97 %. They propose a multiscale CNN that uses both low-level and high-level features for classification. This is done by feeding the results of all feature extraction stages into the classifier. Later they improved the result to 99.17 %. The approach that obtained the best result after the final phase of the IJCNN 2011 is described in [11]. The authors present a multi-column deep neural network (MCDNN) that achieves 99.46 %. An ensemble of 25 CNNs is employed for this, with each five using a different preprocessing method. A single CNN consists of 9 layers, alternating convolutional and max-pooling layers, with two fully connected layers at the end.

In the following years, further approaches that perform traffic sign classification on GTSRB were published. Some important ones are described in [12, 13, 53]. Jin et al. [12] propose a hinge loss stochastic gradient descent (HLSGD) to train an ensemble of CNNs. As in [11], different preprocessing methods are utilized, and for each one, five CNNs are trained. By averaging the output of all 20 CNNs, an accuracy of 99.65 % on the GTSRB is reached. In [53], an approach for both classification and detection of traffic signs is described. For TSC, a supervised three-layer Gaussian-Bernoulli DBM model is first trained. Then, a logistic regression layer is placed on the top hidden layer to build a hierarchical classifier. Using this approach, the authors achieve 99.34 % on GTSRB. Arcos-García et al. [13] propose a convolutional neural network, whose main components are convolutional and spatial transformer modules. Also, SGD without momentum is employed. This way, they correctly classify 99.71 percent of the GTSRB evaluation data set.

Furthermore, work has been done recently to increase TSC's everyday suitability. Some important approaches utilizing the GTSRB are presented in [4, 15, 54, 55, 56]. MicronNet, a lightweight CNN architecture presented by Wong et al. [54], obtains 98.9 percent accuracy on the GTSRB dataset despite having 0.51 million parameters. This is accomplished by leveraging numerical microarchitecture optimization strategies and macroarchitecture design principles. In comparison, MCDNN and HLSGD have 38.5 and 23.2 million parameters, respectively. In [4], a complete traffic sign recognition pipeline consisting of TSD, localization refinement and TSC is presented, which is suitable for real-time deployment. The classification module, which is an efficient CNN with asymmetric kernels, has excellent performance (99.6 %). Two lightweight CNNs are proposed in [56]. Using knowledge distillation, a trained model's knowledge is transferred to a smaller model (0.73 million trainable parameters). The latter, which is used for classifying the evaluation images of the GTSRB, attains an accuracy of 99.61 %. In addition, the student network is compressed utilizing the pruning technique. A network with 0.23 million parameters achieves an accuracy of 99.38 % and one with 0.08 million 99.08 %. Bi et al. [55] reduce the number of parameters of VGG-16 from 33 to 1.15 million by removing convolutional layers. With the proposed approach, an accuracy of 99.21 % on the GTSRB is reached. Zhang et al. [56] present the neural network-based architecture Sill-Net, which increases the robustness under different illumination conditions. The fundamental idea is to separate the illumination and semantic features of the available samples and then augment other training images with the illumination features. This approach shows strong performance (99.68%) on the GTSRB dataset.

In addition to the widely used GTSRB dataset, which covers only German traffic signs, TSC datasets for other countries have been published. These include, for example, Belgium [57], Italy [58], Croatia [59], Sweden [60], and China [61]. Yu et al. [53], Arcos-García et al. [13], Zhang et al. [33], Bendale and Boult [39] and Bi et al. [55] also evaluate their approaches on Belgium traffic signs.

2.3 Related work

Some work has already addressed the issue of a traffic sign classification system encountering test samples that belong to classes not known from training. The majority [24, 25, 26, 27, 28] consider it to be an out-of-distribution detection problem. Iyengar et al. [25] and Guarrera et al. [28] focus solely on the identification of unknown samples, while Masana et al. [24], Chen et al. [26], Schwaiger et al. [27] also report the classification performance within known classes. In [25], various OOD detection methods are applied to the TSC task. 35 GTSRB classes are used as in-distribution (ID) data, 8 GTSRB classes, and a private dataset as out-of-distribution (OOD) data. The latter is made up of 1293 images divided into three classes. Two of these are common traffic sign types, while one class contains region-specific ones. Chen et al. [26] present an algorithm for robust out-of-distribution detection against adversarial perturbations in inputs. In one of three settings, they use GTSRB as ID and the images of CIFAR-10, Textures, Places365, LSUN, and iSUN as OOD samples. Unknown traffic sign classes or objects resembling traffic signs in color or shape are not considered. Guarrera et al. [28] propose an approach to improve the robustness of several OOD detection methods against label shift. For testing, among others, GTSRB is used as ID and several other datasets that do not contain traffic signs or traffic sign-like objects as OOD data. Schwaiger et al. [27] investigate whether uncertainty quantification is useful for detecting out-of-distribution data and compare various methods for doing so. GTSRB is utilized as the in-distribution, while Belgium traffic sign classes [57] that do not have an equivalent in GTSRB form the out-of-distribution dataset. Masana et al. [24] split the Tsinghua traffic sign dataset [61] into two parts: in-distribution and out-of-distribution. The latter is supplemented with Gaussian noise and background patches generated randomly from the Tsinghua dataset's full frames. The decision whether an input is known or unknown is made by measuring euclidean distance in feature space.

Nag et al. [29] present a framework for open-set recognition in a few-shot setting. In two experiments, the authors test this on traffic sign classification. First, 22 GTSRB classes are used as ID and 21 as OOD data. Second, the entire GTSRB and Tsinghua datasets are used as ID and OOD samples, respectively.

Ruiz and Serrat [62] propose a new loss function for hierarchical novelty detection and test it on traffic signs. The goal is to assign classes unseen during training to a previously defined superclass and to classify seen ones correctly. A 'speed limit 30' sign, for example, unknown in training, should be correctly recognized as its superclass 'speed limit'. The Mapillary Traffic Sign dataset [63] and the Tsinghua dataset are utilized in the paper. Objects similar in shape or color to traffic signs are not considered.

Min et al. [20], on the other hand, does not address the problem in the TSC step but tries to limit false positive detections already in the traffic sign detection phase. Only objects in the input image located in a previously determined three-dimensional search region are detected. This method, however, is limited to straight and curved road scenes and is ineffective for complex road scenes such as intersections.

3 Methodology

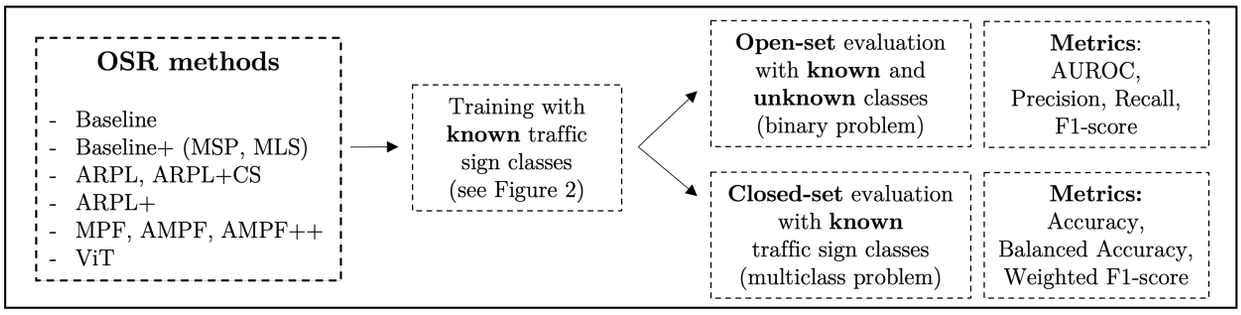

Figure 1: **Methodology**. First, the selected OSR approaches are trained on the known traffic sign classes (see Figure 2 and Appendix, Figures 11 and 12). Then, the open-set and closed-set performance per method is evaluated separately. For the latter, only the known traffic sign classes are used, and for the former, known and unknown traffic sign classes (see Figure 3 and Appendix, Figures 13 and 14) as well as traffic sign-like objects (see Figure 4 and Appendix, Figure 15).

3.1 Selected OSR methods and implementation details

The open-set recognition methods compared in this work, apart from the Baseline method, were selected according to the open-set performance on TinyImageNet reported in the respective paper. It is the most challenging dataset of the standard OSR benchmarks initiated in [41]. Despite the competitive performance, some methods were not considered. No source code is available for OpenHybrid [33] and PMAL [48], and the published implementation of CGDL only supports an image size of 28x28.

Baseline. The standard network in the open-set literature [41] is used as the Baseline method in this work. It is a lightweight CNN (1 million parameters) similar to the VGG architecture, referred to as VGG32 in [50]. The decision of whether an input is classified as known or unknown is made based on the thresholded maximum softmax probability. The network is trained for 100 epochs with a learning rate of 0.01 and cross-entropy loss. As the optimizer, stochastic gradient descent with a momentum of 0.9 is employed.

Baseline+ (MSP, MLS) [50]. Training is performed with the identical network, learning rate, and loss as for Baseline. The number of epochs is increased to 200, the Adam optimizer, a cosine annealed learning rate schedule [64] with one restart, and learning rate warmup for 20 epochs are utilized. For testing, Baseline+ (MSP) uses the softmax normalized output of the network, and Baseline+ (MLS) uses the raw outputs of the final layer, called logits. In a second experiment (see section 5.3), VGG32 is replaced by EfficientNetV2M [65].

ARPL, ARPL+CS [31]. The idea of adversarial reciprocal point learning is described in chapter 2.1. For training, the official source code is used. Hyperparameters and learning rate schedule are identical as in [31]. The approach is trained for 100 epochs, the backbone is VGG32, and the optimizer is Adam.

ARPL+ [31, 50]. Identical to ARPL, but like Baseline+, ARPL+ is trained for 200 epochs and leverages a cosine annealed learning rate schedule with one restart as well as a learning rate warmup for 20 epochs. For the results presented in section 5.1 and 5.2 the VGG32 backbone is utilized, and for those in section 5.3 the EfficientNetV2M backbone. ARPL+CS is not fine-tuned with the described changes. The reason is the minimal performance advantage over ARPL (see section 5.3), but the significantly more computationally expensive training.

MPF, AMPF, AMPF++ [47]. The idea of the three approaches is presented in chapter 2.1. For training, the official source code is used. The hyperparameters and learning rate schedule are identical to the paper [47]. The methods are trained for 100 epochs, the backbone is VGG32, and the optimizer is Adam.

ViT [51]. The approach is described in chapter 2.1 and is referred to as ViT in this paper. For training, the official source code is used. As in [51], the ViT-B/16 variant is used as the backbone. The hyperparameters and optimizer are also identical to those in the paper. Both training stages take 50 epochs each.

For all OSR methods, RandAugment [66] is employed for data augmentation. The parameters M and N are set to 9 and 1, respectively. The following augmentation methods and their ranges, indicated in parentheses, are used: AutoContrast, Equalize, Rotate (0, 30), Color (0.1, 1.9), Contrast, (0.1, 1.9), Brightness (0.1, 1.9), Sharpness (0.1, 1.9), ShearX (0., 0.3), ShearY (0., 0.3), CutoutAbs (0, 40), TranslateXabs (0., 100), and TranslateYabs (0., 100). Each training image is augmented with a chance of 0.5. All images of the traffic sign dataset are resized to 128x128, and the pixel values are rescaled to the range from 0 to 1. The batch size is set to 128 for all methods except ViT. For ViT, the batch size of 256 is taken from [51].

3.2 Evaluation metrics

The evaluation of the OSR approaches compared in this work is based on the standard evaluation protocol in the OSR literature [41]. Identical to the protocol, and as can be seen in Figure 1, performance is evaluated separately for the open-set and closed-set tasks. The former is a binary problem in which known classes have to be distinguished from unknown classes. According to the evaluation protocol, the threshold-free area under the receiver-operator curve (AUROC) is used to measure the open-set performance. It indicates the general ability of a model to distinguish between knowns and unknowns. Since the OSR methods are applied to a real-world problem in this work, F1-score, precision, and recall are measured when using a threshold. It is determined using Youden's index, which allows the selection of an optimal threshold based on the ROC curve [67]. The closed-set task is a multi-class problem within the known traffic sign classes. Due to the imbalanced dataset, balanced accuracy and weighted F1 score are utilized in addition to accuracy, which is usually used in the OSR literature for measuring closed-set performance.

4 Dataset

Figure 2: **Known traffic sign classes.** The dataset contains 66 traffic sign classes with training and testing images. Signs with similar meanings were grouped. The groups and the designation per class can be seen in Figures 11 and 12 in the Appendix. Only these classes are used for the training phase and the evaluation of the closed-set performance. For the evaluation of the open-set performance, these are used as knowns.

Figure 3: **Unknown traffic sign classes.** The class designations can be seen in Figures 13 and 14 in the Appendix. These traffic sign types are used as unknowns for evaluating the open-set performance.

4.1 Data collection and annotation

A new German traffic sign dataset was created for this work. About one-fifth of the images derive from the GTSRB dataset, which contains annotated pictures of 43 traffic sign types and is the only public dataset with German traffic signs. It has 30 images of each traffic sign instance, i.e., the same traffic sign, just photographed from a different distance or angle. Therefore, samples with poor data quality, mostly traffic signs captured from a long distance, were manually removed. Most of the new dataset (78.9 %) consists of self-captured images recorded with the vialytics road management system [68] during daytime. This involves placing a smartphone behind the windshield of a car, which then takes a photo of the road scene every four meters while driving. In this way, 280,696 images were collected.

The traffic signs were first located in the images with an object detector and then cut out. For this purpose, a Faster R-CNN [69] with ResNet-101 [70] backbone was trained on the GTSDB [71] and BTSD datasets [72]. The labeling of the cropped images was done in several steps. First, an image classifier

(Xception [73]) was trained solely on GTSRB to sort the pictures that could be assigned to the 43 GTSRB classes. Only inputs for which the network predicted a maximum softmax probability of at least 0.5 were annotated to minimize the number of false labels. The wrong labels were corrected manually. Next, a small portion of each non-GTSRB class was manually collected to train a classifier tasked with annotating the remaining images. Again, a threshold of 0.5 was set, and the correctness was manually verified. This process was repeated until all cropped images were labeled. Images with poor data quality or those where the traffic sign is occluded, and therefore, the corresponding class is not recognizable, were removed. Traffic signs that are rotated, damaged, faded, poorly illuminated or partially occluded but where the class still can be determined were not removed. Unlike GTSRB, the dataset contains a maximum of 12 images from each traffic sign instance. Some traffic sign types that have similar meanings were grouped for training to simplify the classification of the known classes (see Appendix, Figure 11 and 12). False positives, i.e., objects that are not traffic signs but have been detected as such, are used as unknown test samples (see Figure 4 and Appendix, Figure 15).

Figure 4: **Objects that resemble traffic signs because of their color or shape**. These are used in addition to the unknown traffic sign classes as unknowns for evaluating the open-set performance. The images shown here are only a few examples. More examples can be seen in Figure 15 in the Appendix.

4.2 Dataset statistics

The dataset consists of 143,663 images, of which 94,805 (65.99 %) belong to the training set and 48,858 (34.01 %) to the test set. Publicly available traffic sign classification datasets of other countries, such as Belgium [57], Croatia [59], Italy [58], Sweden [60], and China [61], as well as GTSRB, are significantly smaller. In contrast to these datasets, and as can be seen in Figure 5, the test set is divided into knowns (10,413 images), i.e., those classes also present during training, and unknowns (38,445 images). The unknowns are composed of valid traffic signs for which less than 80 samples are available or that are less important (see Figure 3 and Appendix, Figures 13 and 14). Also included are false positives from the object detector, similar to traffic signs due to their color or shape. These are, for example, billboards, satellite dishes, car tires, or tail lights (see Figure 4 and Appendix, Figure 15). Figure 16 in the Appendix shows the class distribution of the unknown test classes.

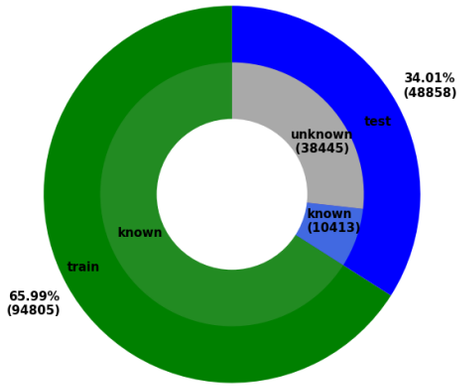

Figure 5: **Proportion of training (green) and test (blue) samples in the entire dataset and proportion of known (blue) and unknown (grey) classes in the test dataset.**

The dataset contains 66 known classes, nine of which are groups of several similar traffic sign types (see Figure 2 and Appendix, Figures 11 and 12). Approximately 90 % of the images were used for training

and ten percent for testing. In some classes, this differs by a few percentage points. This is because only self-captured images are used for testing, even if it is less than ten percent. Furthermore, multiple images of one traffic sign instance are used in their entirety for either the training or the test set. The class distribution of the known classes is shown in Figure 6. It can be seen that this is a highly imbalanced dataset. The largest class contains 9997 training images, and the most minor, 77. The five largest ones in the training dataset each comprise more than 6,000 images and account for 43.81 % of the total training dataset. Seven classes have between 2,000 and 4,000, 17 between 1,000 and 2,000, and 37 less than 1000 samples.

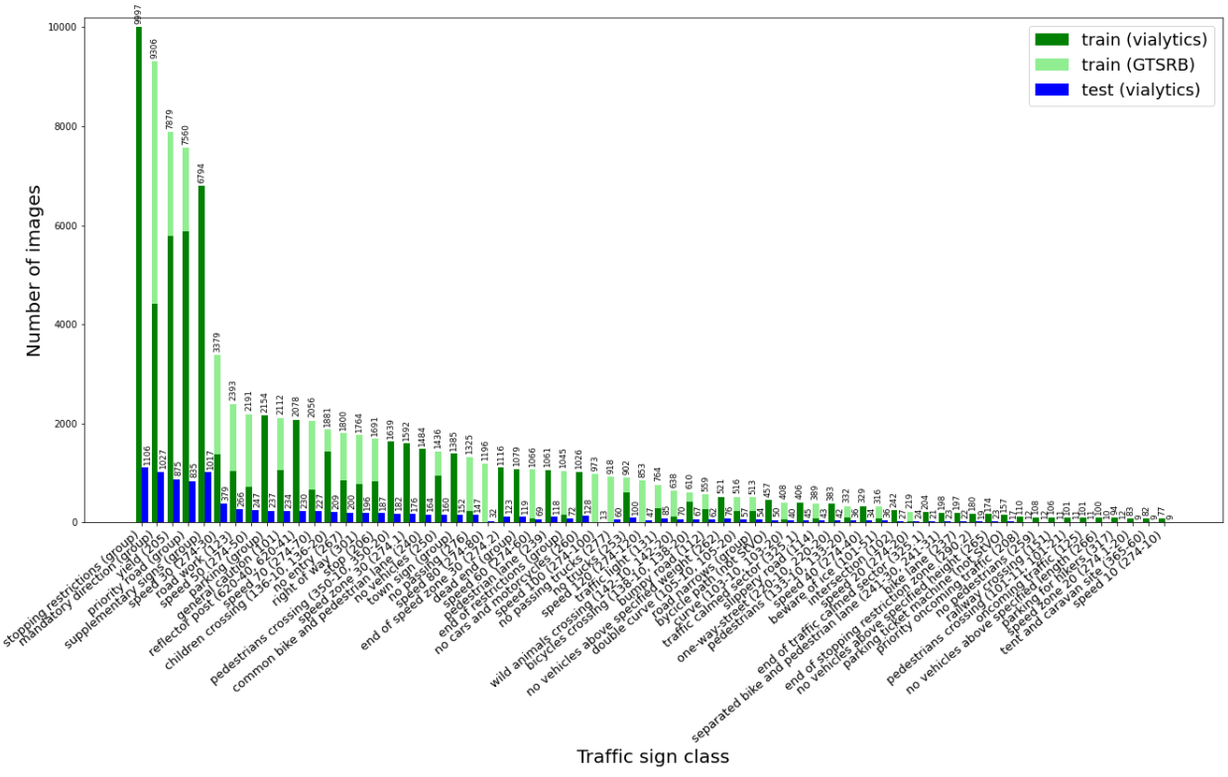

Figure 6: **Class distribution of the known traffic sign classes in the training and testing datasets.** Green bars indicate training data (light green: GTSRB images, green: self-captured images with the vialytics road management system) and blue bars testing data.

5 Results

5.1 Open-set performance

Table 1 summarizes the open-set results of the selected OSR methods on the self-generated German traffic sign dataset. The best results are highlighted in bold. ViT has the highest AUROC score (97.7 %) and, when applying a threshold, the highest F1 score for the known and (82.5 %) and unknown class (94.7 %). If the goal is only to detect the most unknowns, Baseline+ (MSP) is best. At the same time, however, this method achieves the lowest recall value of all methods for the known class. ARPL+ outperforms all other approaches except ViT in terms of AUROC and F1 score, and ARPL+CS is slightly better than ARPL. Baseline+ (MLS) achieves a better AUROC score than Baseline, and regarding the F1 scores, both are almost identical. MPF, AMPF, and AMPF++, have the worst results for almost all metrics. Only AMPF++ is slightly better at detecting unknowns than the Baseline.

Figure 7 illustrates the distribution of the predictions for the three best-performing methods. For ViT, the anomaly score, for ARPL+, the maximum logit score, which corresponds to the distance of a sample to the farthest reciprocal point, and for Baseline+ MSP, the maximum softmax probability per input image is shown. For all three approaches, an overlap between the samples of the known and unknown classes can be seen. The following briefly explains for ViT, ARPL+, and Baseline+ (MSP) what kind of unknown images

Methods	AUROC	known (with threshold)			unknown (with threshold)		
		Precision	Recall	F1	Precision	Recall	F1
MPF [47]	86.3	47.3	80.8	59.6	93.6	75.6	83.6
AMPF [47]	88.6	51.1	83.6	63.4	94.6	78.3	85.7
AMPF++ [47]	90.7	56.8	84.0	67.8	95.0	82.7	88.5
Baseline	91.6	56.6	86.7	68.5	95.8	82.0	88.4
Baseline+ (MLS) [50]	93.0	56.1	88.4	68.6	96.3	81.3	88.1
ARPL [31]	93.3	60.7	86.1	71.2	95.8	84.9	90.0
ARPL+CS [31]	93.4	64.0	84.2	72.7	95.3	87.2	91.1
Baseline+ (MSP)	93.6	**76.8**	72.9	74.8	92.8	**94.0**	93.4
ARPL+ [31, 50]	95.8	72.1	88.1	79.3	96.6	90.8	93.6
ViT [51]	**97.7**	75.9	**90.3**	**82.5**	**97.2**	92.2	**94.7**

Table 1: **Open-set performance**. AUROC is a threshold-independent metric that indicates the general ability of the OSR method to distinguish between knowns and unknowns. For Precision, Recall, and F1 score, a threshold determines whether an input sample is classified as known or unknown. This threshold is determined by Youden's index.

are predicted as one of the known classes when applying a threshold determined by Youden's index. In addition, Figures 17, 19, and 21 in the Appendix show confusion matrices for these three methods under the assumption of an open test set.

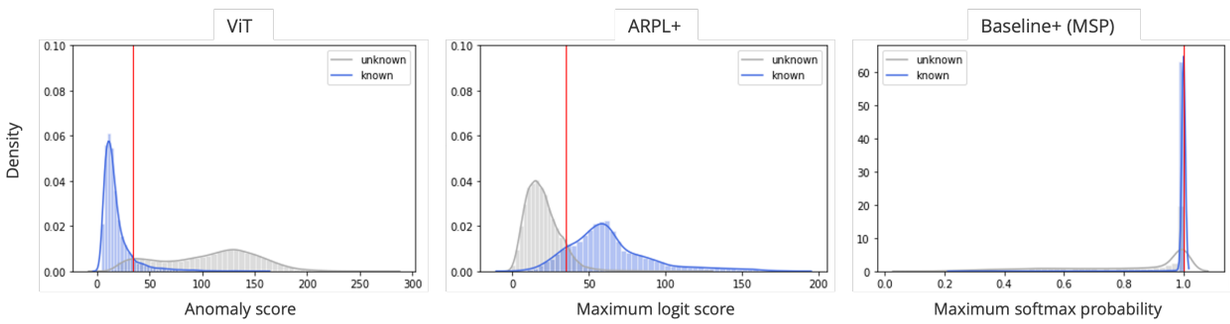

Figure 7: **Distribution of the predictions for the three best-performing methods**. The optimal threshold, determined by Youden's index, is indicated by a red vertical line.

ViT. In total, 2,986 of the 38,445 unknown images are incorrectly classified as one of 16 known classes. Most of them as *'supplementary signs (group)'* (2,118 images), *'town sign (group)'* (626 images), and *'parking (group)'* (198 images). Some examples can be seen in Figure 10. The ones misclassified as *'supplementary sign (group)'* come from the unknown classes *'direction sign (white) (535)'*, *'direction sign bicycle path (not StVO)'*, *'hydrant (sign)'*, *'panels (white) (500-599)'*, *'sign with another sign'*, and *'street name (white) (437)'*, or are other traffic sign-like objects which have a white or light background and text. The unknowns predicted to be class *'town sign (group)'* are all yellow and contain text. They belong to unknown classes *'direction sign (arrow shape, yellow) (415, 418, 419, 454)'* and *'direction sign (yellow) (434, 438, 439, 453)'*, or are other traffic sign-like objects, such as billboards, that also have these attributes. The images misclassified as *'parking (group)'* are all from unknown class *'parking (not StVO)'*. These are very similar to the pictures of *'parking (group)'*. They also have a white letter P on a blue background. Still, they do not correspond exactly to the traffic signs according to the German Road Traffic Regulations because other numbers, arrows, or texts are also depicted.

ARPL+. 3,553 unknown images are incorrectly predicted as 53 different known classes. Most of them as *'supplementary signs (group)'* (787 images), *'priority road (group)'* (687 images), *'parking (group)'* (630 images), *'town sign (group)'* (286 images), *'mandatory direction'* (130 images), *'stop (206)'* (108 images), *'yield (205)'* (108 images), *'traffic calmed sector (325.1)'* (70 images), and *'stopping restrictions (group)'* (58 images). Again, images incorrectly classified as *'supplementary signs (group)'* or *'town sign (group)'* are mainly traffic signs or similar objects with text and white or yellow background. Red and white objects with text are erroneously recognized as class *'stop (206)'*. For *'priority road (group)'*, the misclassifications come from the unknown class *'barriers, beacon, direction in curves, etc. (600, 605, 625, 626, 628, 629,*

630)', or are also diamond-shaped. Images of the unknown class *'parking (not StVO)'* and other blue-white traffic signs or traffic sign-like objects are predicted as *'parking (group)'*. The latter is also valid for the classes *'mandatory direction (group)'* and *'traffic calmed sector (325.1)'*. The unknown images assigned to *'yield (205)'* all have the same triangular shape as the actual traffic sign. Most belong to the unknown class *'nature reserve (not StVO)'*. Pictures misclassified as *'stopping restrictions (group)'* either have the same colors as the actual sign or are from class *'taxi (229)'*. Among the remaining misclassifications, it is also noticeable that unknown traffic sign classes with a red border and white background with either a round or triangular shape are assigned to known classes to which the same attributes apply.

Baseline+ (MSP). 2,293 unknown images are predicted as 14 known classes. The majority as *'supplementary signs (group)'* (1,844 images), *'parking (group)'* (157 images), *'yield (205)'* (81 images), *'reflector post (620-40, 620-41)'* (70 images) *'priority road (group)'* (59 images), and *'town sign (group)'* (33 images). As with ViT and ARPL+, white or yellow signs with text are wrongly recognized as *'supplementary signs (group)'* and *'town sign (group)'*, and the images of the unknown class *'parking (not StVO)'* are misclassified as *'parking (group)'*. Images predicted to be *'yield (205)'* mainly belong to class *'nature reserve (not StVO)'*. Photos misidentified as class *'reflector posts (620-40, 620-41)'* are mostly from the unknown group *'barriers, beacon, direction in curves, etc. (600, 605, 625, 626, 628, 629, 630)'*, and those detected to be *'priority road (group)'* are diamond-shaped like the actual traffic sign.

5.2 Closed-set performance

Table 2 shows the closed-set performance per OSR method. ViT is the best performing method with an accuracy of 99.6 %, a balanced accuracy of 98.0 %, and a weighted F1 score of 99.6 %. Baseline+ (MLS) and Baseline+ (MSP) are second best, with only slightly worse results than ViT. ARPL+, ARPL+CS, and ARPL follow them. As with the open test set, ARPL+CS performs better than ARPL, and ARPL+ outperforms both. All three methods also do better than the baseline model. MPF, AMPF, and AMPF++ achieve the worst results. Figures 18, 20, and 22 in the Appendix show confusion matrices for ViT, Baseline+ (MSP, MLS), and ARPL+ when only known classes are used for testing.

Methods	Accuracy	Balanced Accuracy	Weighted F1
MPF [47]	84.9	65.3	84.1
AMPF++ [47]	94.6	83.6	94.4
AMPF [47]	94.9	84.6	94.5
Baseline	98.1	90.9	97.8
ARPL [31]	98.3	93.2	98.1
ARPL+CS [31]	98.8	93.9	98.7
ARPL+ [31, 50]	99.0	94.5	98.9
Baseline+ (MLS) [50]	99.4	97.5	99.4
Baseline+ (MSP) [50]	99.4	97.5	99.4
ViT [51]	**99.6**	**98.0**	**99.6**

Table 2: **Closed-set performance**: Accuracy, balanced accuracy, and weighted F1 score are used to evaluate the performance on the multi-class problem, when only known classes are used for testing. The class with the maximum softmax probability or logit score (depending on the method) was predicted.

Figure 8 illustrates the F1 score per OSR method and traffic sign class, sorted in the same order as in Figure 6. MPF, AMPF, and AMPF++ were removed for clarity as they perform significantly worse. Baseline+ (MSP) and Baseline+ (MLS) were combined because they have identical closed-set results.

For two thirds of the classes, comparable and almost perfect performance prevails for ViT, ARPL+, ARPL+CS, ARPL, Baseline+ (MSP, MLS), and Baseline. ViT is the best approach in 52 classes. 45 times at least one other method is equally good, and in 7 classes, ViT alone is best. Baseline+ performs best 44 times with others and six times alone, ARPL+ 36 and two times. In some classes, the performance differs significantly per approach. In *'speed 80 (274-80)'*, ARPL+ achieves 1.0, ViT 0.92, Baseline+ 0.77, ARPL+cs 0.69, ARPL 0.48, and Baseline 0.17 F1 score. For class *'speed 100 (274-100)'*, ViT has a score of 1.0 and Baseline+ of 0.87, while all other methods do not correctly classify any image. In class *'speed 120 (274-120)'*, ViT, ARPL+, and Baseline+ all achieve at least 0.97 F1 score, while ARPL+CS, ARPL, and the Baseline method only reach 0.53, 0.29, and 0.19. For *'speed 40 (274-40)'*, ViT, ARPL+cs, and Baseline+ attain not less than 0.94. On the contrary, Baseline, ARPL, and ARPL+ only reach 0.69, 0.77,

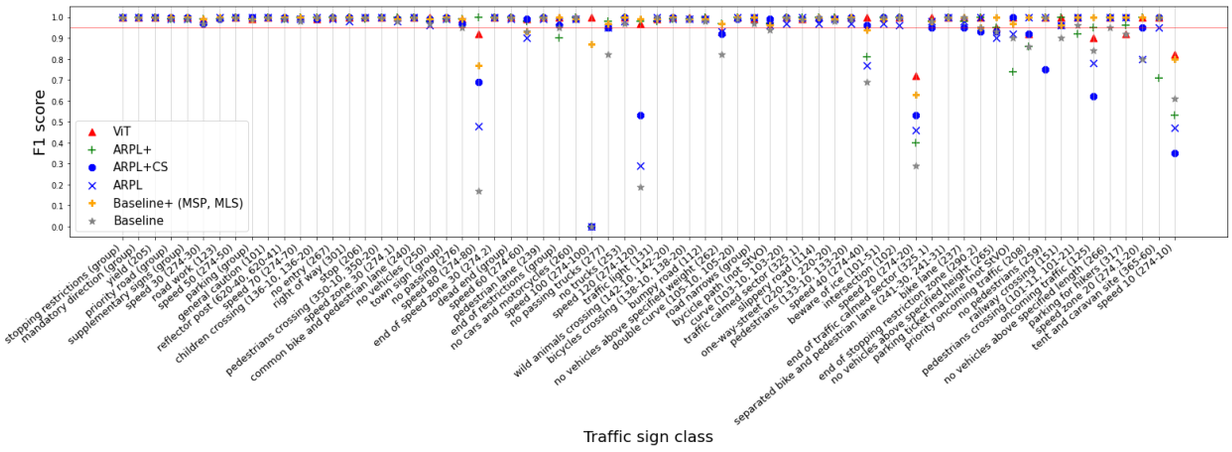

Figure 8: **F1 score per traffic sign class and OSR method**. MPF, AMPF, and AMPF++ were removed for clarity as they perform significantly worse. Baseline+ (MSP) and Baseline+ (MLS) were combined because they have identical closed-set results. The red horizontal line indicates an F1 score of 0.95.

and 0.81, respectively. In the classes *'speed 20 (274-20)'* and *'speed 10 (274-10)'*, no method achieves a result greater than 0.82. ViT is best in both cases. The worst performing methods achieve 0.29 (baseline) and 0.35 (ARPL+CS). Also worth mentioning is the class *'oncoming traffic (125)'*, where the F1 value varies greatly depending on the OSR approach, from 1.0 (Baseline+) to 0.62 (ARPL+CS).

5.3 Results with state-of-the-art CNN architecture

Table 3 lists the open-set and Table 4 the closed-set results for Baseline+ (MSP), Baseline+ (MLS), and ARPL+ with EfficientNetV2M backbone.

Open-set performance. For ARPL+, the better backbone results in a 1.2 % higher AUROC score, and the threshold-dependent metrics were also improved. In terms of AUROC score, ViT is still better than ARPL+, but for the threshold-dependent metrics, the performance is almost identical. Contrary to the results with VGG32 backbone, Baseline+ (MLS) performs now better than Baseline+ (MSP). The former has improved in all metrics and has an almost identical AUROC score as ARPL+ with VGG32 backbone. However, it cannot match this approach for the threshold-dependent metrics. For Baseline+ (MSP), the EfficientNetV2M backbone results in a deterioration. The AUROC value is even worse than for AMPF++. However, when applying a threshold, Baseline+ (MSP) has the highest precision for detecting unknowns, but this results in finding fewer of them. For both ARPL+ and Baseline+ (MLS), unknowns are mainly misclassified as the known classes *'supplementary signs (group)'*, *'town sign (group)'*, and *'parking (group)'*.

Methods	AUROC	known (with threshold)			unknown (with threshold)		
		Precision	Recall	F1	Precision	Recall	F1
Baseline+ (MSP)	89.9	59.4	**92.9**	72.5	**97.7**	82.8	89.7
Baseline+ (MLS)	95.7	68.5	89.1	77.5	96.8	88.9	92.7
ARPL+	97.0	**75.9**	90.4	**82.5**	97.2	**92.2**	**94.7**
ViT	**97.7**	**75.9**	90.3	**82.5**	97.2	**92.2**	**94.7**

Table 3: **Open-set performance of Baseline+ (MSP, MLS) and ARPL+ with EfficientNetV2M backbone compared to ViT.**

Closed-set performance. The two Baseline+ methods again have identical results. They outperform the previous best method, ViT. ARPL+ improves with EfficientNetV2 backbone in all metrics, especially in balanced accuracy. It is thus almost identical to Baseline+ (MSP, MLS) with VGG32.

Methods	Accuracy	B. Accuracy	Weighted F1
ARPL+	99.4	97.6	99.3
ViT	99.6	98.0	99.6
Baseline+ (MSP)	**99.8**	**98.9**	**99.8**
Baseline+ (MLS)	**99.8**	**98.9**	**99.8**

Table 4: **Closed-set performance of Baseline+ (MSP, MLS) and ARPL+ with EfficientNetV2M backbone compared to ViT.**

6 Discussion

The OSR approach that shows the best performance in this work for both the closed-set and open-set tasks, when compared to the standard implementations of other methods, is a vision transformer (ViT-B/16) with an attached detection head [51] pre-trained on ImageNet-21K [74]. This confirms the results presented in [51], where new state-of-the-art performance for open-set recognition is claimed. Similarly, the benefits of vision transformers for OSR and OOD detection were stated in [50] and [75, 76], respectively. However, Salehi et al. [35] trace these successes back to the pretraining on ImageNet. It has a large intersection with the training and testing data used in the experiments, which means that the unknown classes are not truly unknown. This is not the case for the dataset utilized in this work and can therefore be disproved for the traffic sign classification problem.

It should be noted, however, that the comparison with the other OSR methods is not fair. For the CNN based approaches, the obsolete architecture VGG32 (1.0 million parameters) is used to evaluate the methods against the standard OSR benchmarks. In contrast, vision transformers have recently been proposed as a replacement for CNNs [52]. Also, the ViT architecture used in this work has 86 times more parameters than VGG32. For the experiments with larger datasets in [31, 47, 48, 50], VGG32 is exchanged with ResNet-50 [70], but as Cai et al. [51] noted, more recent CNN architectures could further improve the performance. The experiments presented in Section 5.3 confirm this conjecture. ARPL+ with an EfficientNetV2M backbone (54.4 million parameters) [65] achieves competitive results for the open and closed-set tasks compared to the vision transformer. Replacing the backbone for Baseline+ (MLS, MSP) outperforms ViT in the closed-set task. Therefore, if vision transformers are increasingly used for OSR in the future, fair comparison conditions should be established.

The results presented in [31] regarding adversarial reciprocal point learning, coincide with the results in this work. ARPL has better open and closed-set performance than Baseline, and ARPL+CS can enhance the performance even further.

The open-set results reported in [50] for the standard OSR benchmark datasets are partially consistent with the results for the traffic sign dataset. In this work, the Baseline model's performance could also be boosted by training for a longer time and using a custom learning rate schedule. This way, as in [50], the performance of the much more sophisticated approaches ARPL and ARPL+CS could be surpassed. Analogous to [50], these techniques were also applied to the adversarial reciprocal point learning framework (ARPL+). Contrary to the results on the OSR benchmark datasets reported in [50], this leads to a significantly better open-set performance than the improved baseline model achieves. The finding in [50] that for a conventional CNN, the maximum logit score is better suited than the maximum softmax probability to distinguish unknown from known classes applies in this work exclusively to the model with EfficientNetV2M backbone. Not, however, to the one with VGG32 backbone. Vaze et al. [50] report closed-set performance only on the self-proposed semantic shift benchmark, which includes three fine-grained classification datasets and ImageNet. Baseline+ MLS is superior in all four cases. This is also the case for the traffic sign dataset.

The poor performance of MPF, AMPF, and AMPF++ is unexpected and does not coincide with the results in [47]. The paper reports better open-set performance of AMPF and AMPF++ than ARPL and ARPL+CS. However, the statement that adding adversarial samples to the training improves the results, can be confirmed. Regarding the closed-set performance, it should be noted that the model checkpoints were selected according to the best AUROC result. However, comparable validation accuracy to ARPL and ARPL+ was achieved when training the models, but the AUROC result was significantly worse for those checkpoints.

As described in section 5.2, there are some classes where bad closed-set performance is achieved across

multiple OSR methods. Figure 9a) shows the relation between the average F1 score across several OSR approaches and the number of training images per traffic sign class. Figure 9b) illustrates the average F1 score and the proportion of GTSRB training images per class. No overall correlation can be found for both cases. However, for classes *'speed 10 (274-10)'* and *'speed 20 (274-20)'*, dependence between the F1 score and the number of training images can be assumed. An attempt to solve this problem by applying class weights did not lead to any improvements. Also, a dependency between the proportion of GTSRB training images and the F1 score of the classes *'speed 20 (274-20)'*, *'speed 80 (274-80)'*, *'speed 100 (274-100)'*, and *'speed 120 (274-120)'* can be assumed. This could be because the GTSRB dataset, as described in section 4.1, contains many images per traffic sign instance and thus has a lot of similar samples.

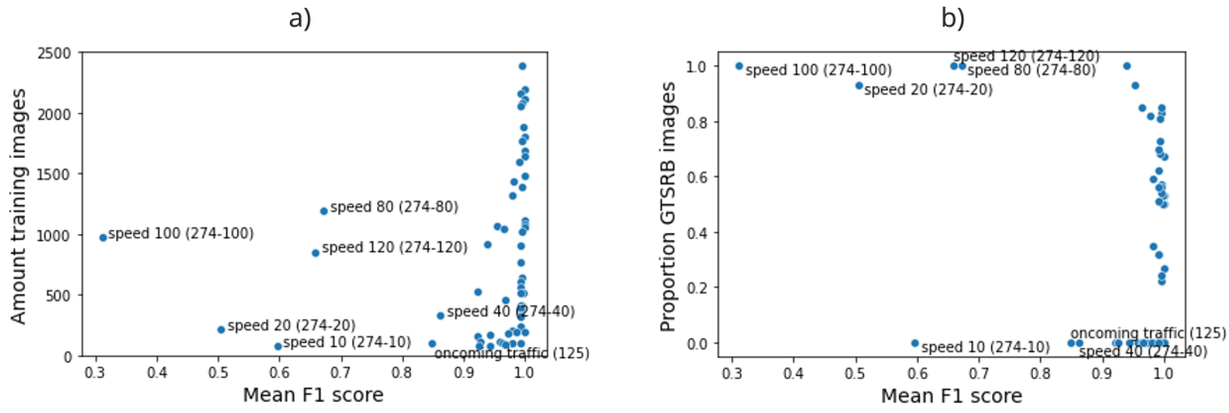

Figure 9: **Mean F1 score in relation to the number of training images (a) and the proportion of GTSRB training images (b) per traffic sign class**. The mean F1 score is calculated from the results of ViT, ARPL+, ARPL+CS, ARPL, Baseline, and Baseline+ (MSP, MLS) presented in section 5.2. MPF, AMPF, and AMPF++ were not considered, due to their significantly worse closed-set performance. In plot a), five classes with more than 2500 images (which all have nearly perfect closed-set performance) are not included for clarity.

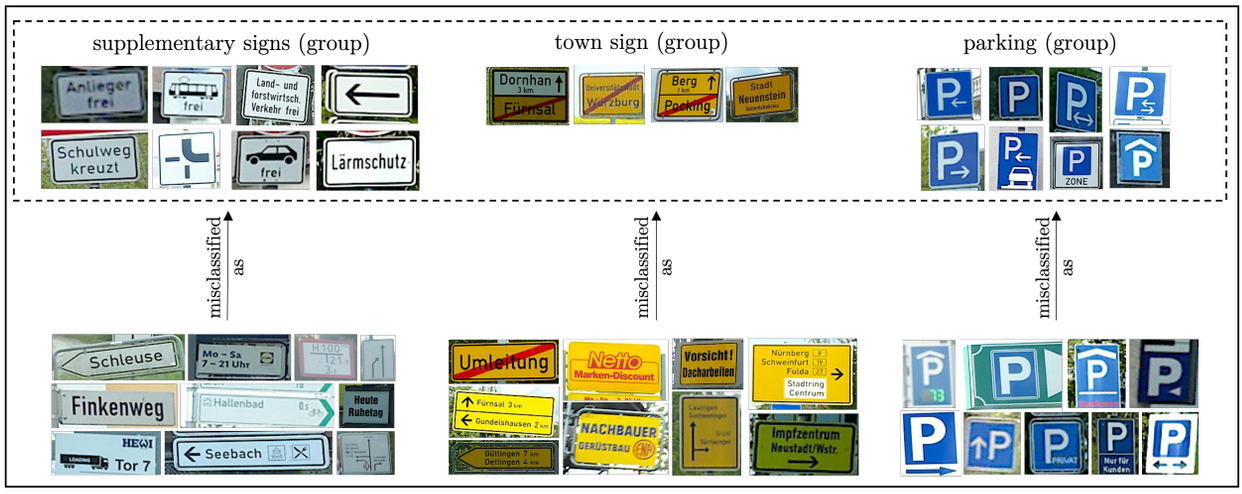

Figure 10: **Unknown images misclassified by ViT as one of the known classes.**

As described in Chapter 5, for multiple OSR methods, many of the misclassified unknown images are predicted as the classes *'supplementary signs (group)'*, *'town sign (group)'*, and *'parking (group)'*. As can be seen in Figure 10, those are all groups of several similar but not identical traffic sign types. The Figure also shows some misclassifications of ViT, which are very similar to the traffic signs in the respective groups. In such cases, it might be a problem that for ARPL, ARPL+CS, and ARPL+, a single reciprocal point, and for ViT, a single class center has the task of representing an entire class. For the two classes of *'town sign (group)'* and *'parking (group)'*, the idea of the prototype-based OSR approach described in [48] of having multiple prototypes per class could be a possible solution to this problem. The class *'supplementary signs*

(group)' probably has too much intra-class diversity, as it contains over 70 different subcategories. They all have a rectangular shape and a white background color in common. However, on some, only text or a pictogram is displayed, and on others, both. It might be helpful here to build subcategories with lower intra-class diversity first.

7 Limitations and future work

This study is limited to traffic signs and traffic sign-like objects in Germany during daylight hours. The dataset created for this contains, besides the known German traffic sign classes, a wide range of images of unknown traffic signs and objects similar to traffic signs based on their shape or color. It should be noted, however, that the real world can still not be fully represented with this.

The distribution of known traffic sign classes is very unbalanced. As discussed, dependence between the low number of training images and the bad performance in some classes can be assumed. Therefore, balancing the distribution of known training classes should be considered in future work. One approach is the synthetic minority oversampling technique (SMOTE) [77]. Regarding the classes with poor performance and a high proportion of GTSRB images, it should be considered to collect more training data since many of the GTSRB images are very similar.

Moreover, this work grouped different traffic sign types with similar meanings to simplify the problem. However, as mentioned earlier, this most likely weakened the ability to distinguish between known and unknown signs with high similarity. It seems reasonable to resolve this grouping in the future or to apply methods to the problem that address intra-class diversity, such as [48]. Another limitation of this study, as described in Chapter 3, is that some state-of-the-art open-set recognition methods could not be incorporated because the source code is not available or usable. In particular, PMAL [48] and OpenHybrid [33] should be considered in future studies. Also, for the selected OSR methods, only the default hyperparameters reported in the respective papers were used. A finetuning of these could further improve the performance. Finally, in addition to the OSR methods, state-of-the-art out-of-distribution detection approaches might also be suited for the problem and should therefore be applied to the dataset in the future.

8 Conclusion

In the literature, MNIST, SVHN, CIFAR, and TinyImageNet are primarily used to evaluate open-set recognition approaches. For this, these are split into known and unknown classes. Newer methods have additionally been evaluated on larger, fine-grained, and long-tailed datasets. In this work, however, selected OSR approaches were applied to a real-world problem: the classification of German traffic signs under an open-set scenario. Actual road signs that are not present in the training dataset and real-world objects that have similarities due to their color or shape were used as unknown classes. For example, billboards, car tail lights, satellite dishes, or house roofs.

The following can essentially be summarized for the open-set task, i.e., the distinction between the test classes known and unknown from the training phase. All compared OSR methods, except the (adversarial) motorial prototype framework, achieve better results than a conventional CNN under the same training conditions. This shows that the research successes in open-set recognition can, in principle, be transferred to real-world applications. However, unknown images that show high similarity to known classes which have high intra-class diversity are often confused by several methods. The approach with the best open-set performance, which was recently proposed for OSR, is based on a vision transformer. However, competitive open-set performance can be achieved with the OSR framework of adversarial reciprocal point learning, when the outdated CNN architecture is replaced with EfficientNetV2M, a custom learning rate schedule is applied, and training time is increased.

For the closed-set task, i.e., solely classifying test images from the traffic sign classes available during the training phase, the following can be concluded. The vision transformer performs best when the other methods utilize the outdated CNN architecture VGG32. When replacing this with an EfficientNetV2M, applying a custom learning rate schedule, and training for more epochs, the Baseline method outperforms the vision transformer. Across multiple methods, it can be observed that classes with poor classification performance either have a low number of training images, a high proportion of GTSRB training images, or both.

References

[1] A. de la Escalera, J. Armingol, and M. Mata, "Traffic sign recognition and analysis for intelligent vehicles," *Image and Vision Computing*, vol. 21, no. 3, pp. 247–258, 2003.

[2] H. Akatsuka and S. Imai, "Road signposts recognition system," in *SAE Technical Paper Series*. SAE International, 1987.

[3] M.-Y. Fu and Y.-S. Huang, "A survey of traffic sign recognition," in *2010 International Conference on Wavelet Analysis and Pattern Recognition*. IEEE, 2010.

[4] 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.

[5] R.-Q. Qian, Y. Yue, F. Coenen, and B.-L. Zhang, "Traffic sign recognition using visual attribute learning and convolutional neural network," in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*. IEEE, 2016.

[6] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1484–1497, 2012.

[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.

[8] ——, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Networks*, vol. 32, 2012.

[9] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," in *The 2011 International Joint Conference on Neural Networks*. IEEE, 2011.

[10] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in *The 2011 International Joint Conference on Neural Networks*. IEEE, 2011.

[11] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," *Neural Networks*, vol. 32, pp. 333–338, 2012.

[12] J. Jin, K. Fu, and C. Zhang, "Traffic sign recognition with hinge loss trained convolutional neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 1991–2000, 2014.

[13] Á. Arcos-García, J. A. Álvarez-García, and L. M. Soria-Morillo, "Deep neural network for traffic sign recognition systems: An analysis of spatial transformers and stochastic optimisation methods," *Neural Networks*, vol. 99, pp. 158–165, 2018.

[14] A. Shustanov and P. Yakimov, "CNN design for real-time traffic sign recognition," *Procedia Engineering*, vol. 201, pp. 718–725, 2017.

[15] J. Zhang, W. Wang, C. Lu, J. Wang, and A. K. Sangaiah, "Lightweight deep network for traffic sign classification," *Annals of Telecommunications*, vol. 75, no. 7, pp. 369–379, 2020.

[16] S. Zhou, C. Deng, Z. Piao, and B. Zhao, "Few-shot traffic sign recognition with clustering inductive bias and random neural network," *Pattern Recognition*, vol. 100, p. 107160, 2020.

[17] F. Woitschek and G. Schneider, "Physical adversarial attacks on deep neural networks for traffic sign recognition: A feasibility study," in *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2021.

[18] C. Sitawarin, A. N. Bhagoji, A. Mosenia, P. Mittal, and M. Chiang, "Rogue signs: Deceiving traffic sign recognition with malicious ads and logos," 2018.

[19] Z. Wang, J. Wang, Y. Li, and S. Wang, "Traffic sign recognition with lightweight two-stage model in complex scenes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 1121–1131, 2022.

[20] W. Min, R. Liu, D. He, Q. Han, Q. Wei, and Q. Wang, "Traffic Sign Recognition Based on Semantic Scene Understanding and Structural Traffic Sign Location," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2022.

[21] Bundesanstalt für Straßenwesen, "Liste aller verkehrszeichen des verkehrszeichenkatalogs," 2021.

[22] E. C. for Europe-Inland Tansport Committee *et al.*, "Convention on road signs and signals," *United Nations Treaty Series*, vol. 1091, p. 3, 1968.

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

[24] M. Masana, I. Ruiz, J. Serrat, J. van de Weijer, and A. M. Lopez, "Metric learning for novelty and anomaly detection," 2018.

[25] M. Iyengar, M. Opitz, and Bischof, Horst , "Detecting out-of-distribution traffic signs," 2019.

[26] J. Chen, Y. Li, X. Wu, Y. Liang, and S. Jha, "Robust out-of-distribution detection for neural networks," *arXiv preprint arXiv:2003.09711*, 2020.

[27] A. Schwaiger, P. Sinhamahapatra, J. Gansloser, and K. Roscher, "Is uncertainty quantification in deep learning sufficient for out-of-distribution detection?" in *AISafety@ IJCAI*, 2020.

[28] M. Guarrera, B. Jin, T.-W. Lin, M. A. Zuluaga, Y. Chen, and A. Sangiovanni-Vincentelli, "Class-wise thresholding for robust out-of-distribution detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 2836–2845.

[29] S. Nag, D. S. Raychaudhuri, S. Paul, and A. K. Roy-Chowdhury, "Learning few-shot open-set classifiers using exemplar reconstruction," *arXiv preprint arXiv:2108.00340*, 2021.

[30] D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Learning placeholders for open-set recognition," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4399–4408.

[31] G. Chen, P. Peng, X. Wang, and Y. Tian, "Adversarial reciprocal points learning for open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[32] S. Esmaeilpour, L. Shu, and B. Liu, "Open-set recognition via augmentation-based similarity learning," 2022.

[33] H. Zhang, A. Li, J. Guo, and Y. Guo, "Hybrid models for open set recognition," in *Computer Vision – ECCV 2020*. Springer International Publishing, 2020, pp. 102–117.

[34] T. E. Boult, S. Cruz, A. Dhamija, M. Gunther, J. Henrydoss, and W. Scheirer, "Learning and the unknown: Surveying steps toward open world recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 9801–9807, 2019.

[35] M. Salehi, H. Mirzaei, D. Hendrycks, Y. Li, M. H. Rohban, and M. Sabokrou, "A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges," 2021.

[36] Z. Yue, T. Wang, Q. Sun, X.-S. Hua, and H. Zhang, "Counterfactual zero-shot and open-set visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 404–15 414.

[37] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, 2013.

[38] C. Geng, S.-J. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3614–3631, 2021.

[39] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1563–1572.

[40] Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi, "Generative openmax for multi-class open set classification," 2017.

[41] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open set learning with counterfactual images," in *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 620–635.

[42] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, and T. Naemura, "Classification-reconstruction learning for open-set recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019.

[43] P. Oza and V. M. Patel, "C2ae: Class conditioned auto-encoder for open-set recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[44] X. Sun, Z. Yang, C. Zhang, K.-V. Ling, and G. Peng, "Conditional gaussian distribution learning for open set recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 477–13 486.

[45] H.-M. Yang, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Robust classification with convolutional prototype learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3474–3482.

[46] G. Chen, L. Qiao, Y. Shi, P. Peng, J. Li, T. Huang, S. Pu, and Y. Tian, "Learning open set network with discriminative reciprocal points," in *Computer Vision – ECCV 2020*. Springer International Publishing, 2020, pp. 507–522.

[47] Z. Xia, P. Wang, G. Dong, and H. Liu, "Adversarial motorial prototype framework for open set recognition," 2021.

[48] J. Lu, Y. Xu, H. Li, Z. Cheng, and Y. Niu, "Pmal: Open set recognition via robust prototype mining," *arXiv preprint arXiv:2203.08569*, 2022.

[49] H.-M. Yang, X.-Y. Zhang, F. Yin, Q. Yang, and C.-L. Liu, "Convolutional prototype network for open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[50] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: a good closed-set classifier is all you need?" 2021.

[51] F. Cai, Z. Zhang, J. Liu, and X. Koutsoukos, "Open set recognition using vision transformer with an additional detection head," 2022.

[52] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[53] Y. Yu, J. Li, C. Wen, H. Guan, H. Luo, and C. Wang, "Bag-of-visual-phrases and hierarchical deep models for traffic sign detection and recognition in mobile laser scanning data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 113, pp. 106–123, 2016.

[54] A. Wong, M. J. Shafiee, and M. S. Jules, "MicronNet: A highly compact deep convolutional neural network architecture for real-time embedded traffic sign classification," *IEEE Access*, vol. 6, pp. 59 803–59 810, 2018.

[55] Z. Bi, L. Yu, H. Gao, P. Zhou, and H. Yao, "Improved VGG model-based efficient traffic sign recognition for safe driving in 5g scenarios," *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 11, pp. 3069–3080, 2020.

[56] H. Zhang, Z. Cao, Z. Yan, and C. Zhang, "Sill-net: Feature augmentation with separated illumination representation," *arXiv preprint arXiv:2102.03539*, 2021.

[57] R. Timofte, K. Zimmermann, and L. V. Gool, "Multi-view traffic sign detection, recognition, and 3d localisation," *Machine Vision and Applications*, vol. 25, no. 3, pp. 633–647, 2011.

[58] A. Youssef, D. Albani, D. Nardi, and D. D. Bloisi, "Fast traffic sign recognition using color segmentation and deep convolutional networks," in *Advanced Concepts for Intelligent Vision Systems*. Springer International Publishing, 2016, pp. 205–216.

[59] F. Jurišić, I. Filković, and Z. Kalafatić, "Multiple-dataset traffic sign classification with onecnn," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 614–618.

[60] F. Larsson and M. Felsberg, "Using fourier descriptors and spatial models for traffic sign recognition," in *Scandinavian conference on image analysis*. Springer, 2011, pp. 238–249.

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

[62] I. Ruiz and J. Serrat, "Hierarchical novelty detection for traffic sign recognition," *Sensors*, vol. 22, no. 12, p. 4389, 2022.

[63] 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*. Springer International Publishing, 2020, pp. 68–84.

[64] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," 2016.

[65] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 10 096–10 106.

[66] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[67] R. Fluss, D. Faraggi, and B. Reiser, "Estimation of the youden index and its associated cutoff point," *Biometrical Journal*, vol. 47, no. 4, pp. 458–472, 2005.

[68] vialytics GmbH, "vialytics - the intelligent road management system," 2022, https://www.vialytics.com/.

[69] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[70] 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.

[71] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark," in *International Joint Conference on Neural Networks*, no. 1288, 2013.

[72] 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.

[73] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[74] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[75] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7068–7081, 2021.

[76] R. Koner, P. Sinhamahapatra, K. Roscher, S. Günnemann, and V. Tresp, "Oodformer: Out-of-distribution detection transformer," *arXiv preprint arXiv:2107.08976*, 2021.

[77] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority oversampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

Appendix

Known training and test classes

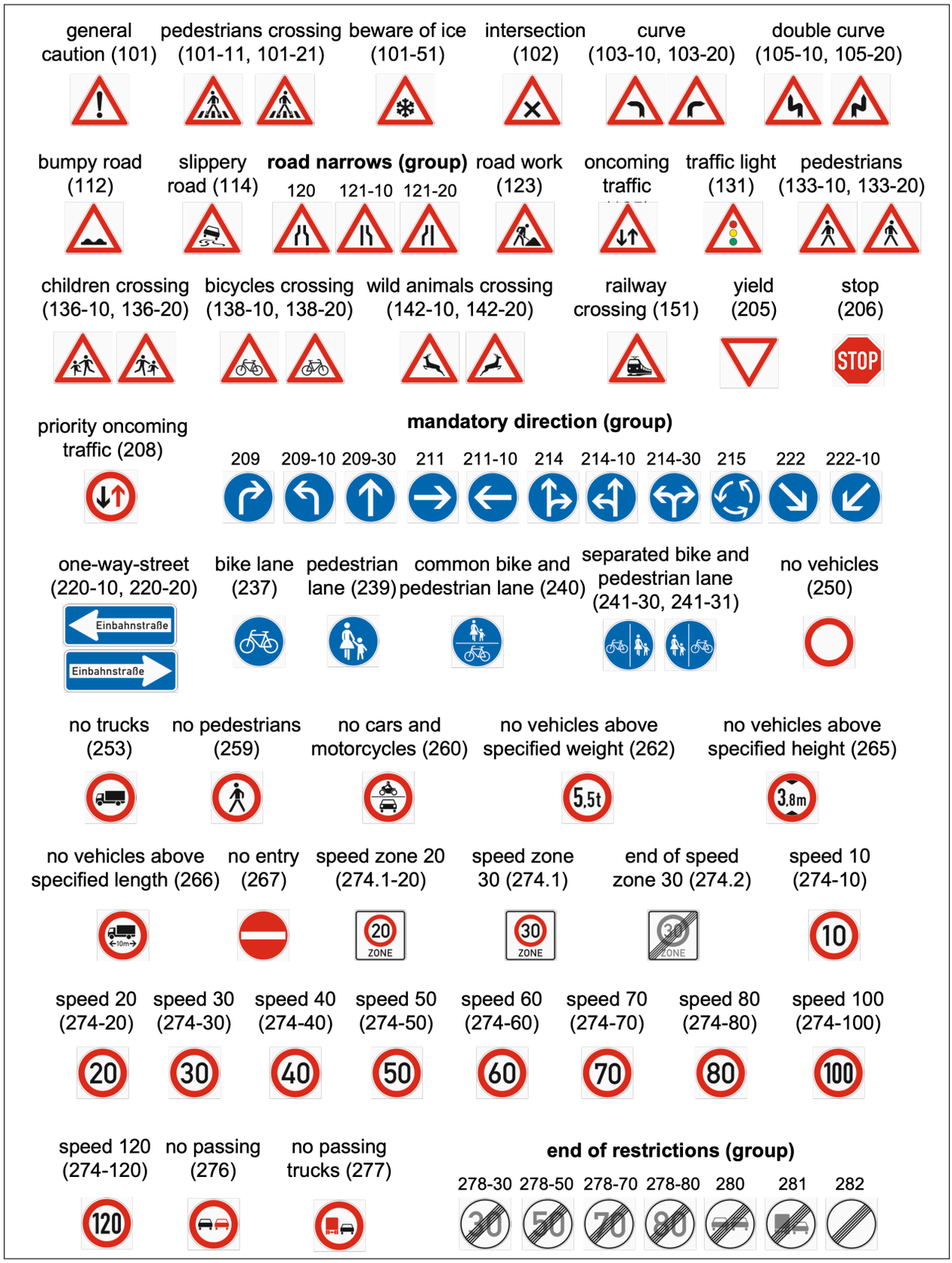

Figure 11: **Known training and test classes** (1/2)

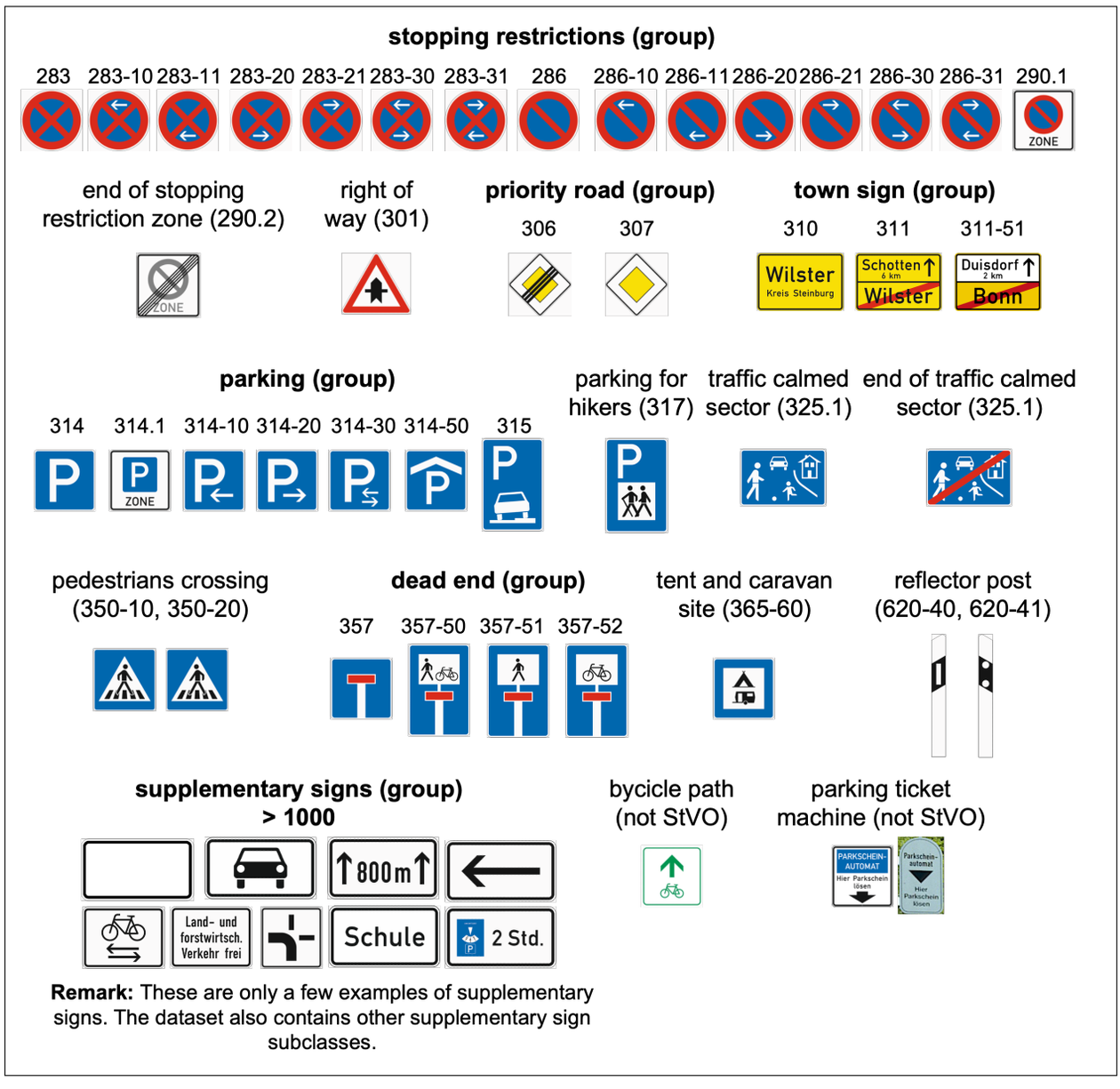

Figure 12: **Known training and test classes** (2/2)

Unknown test classes

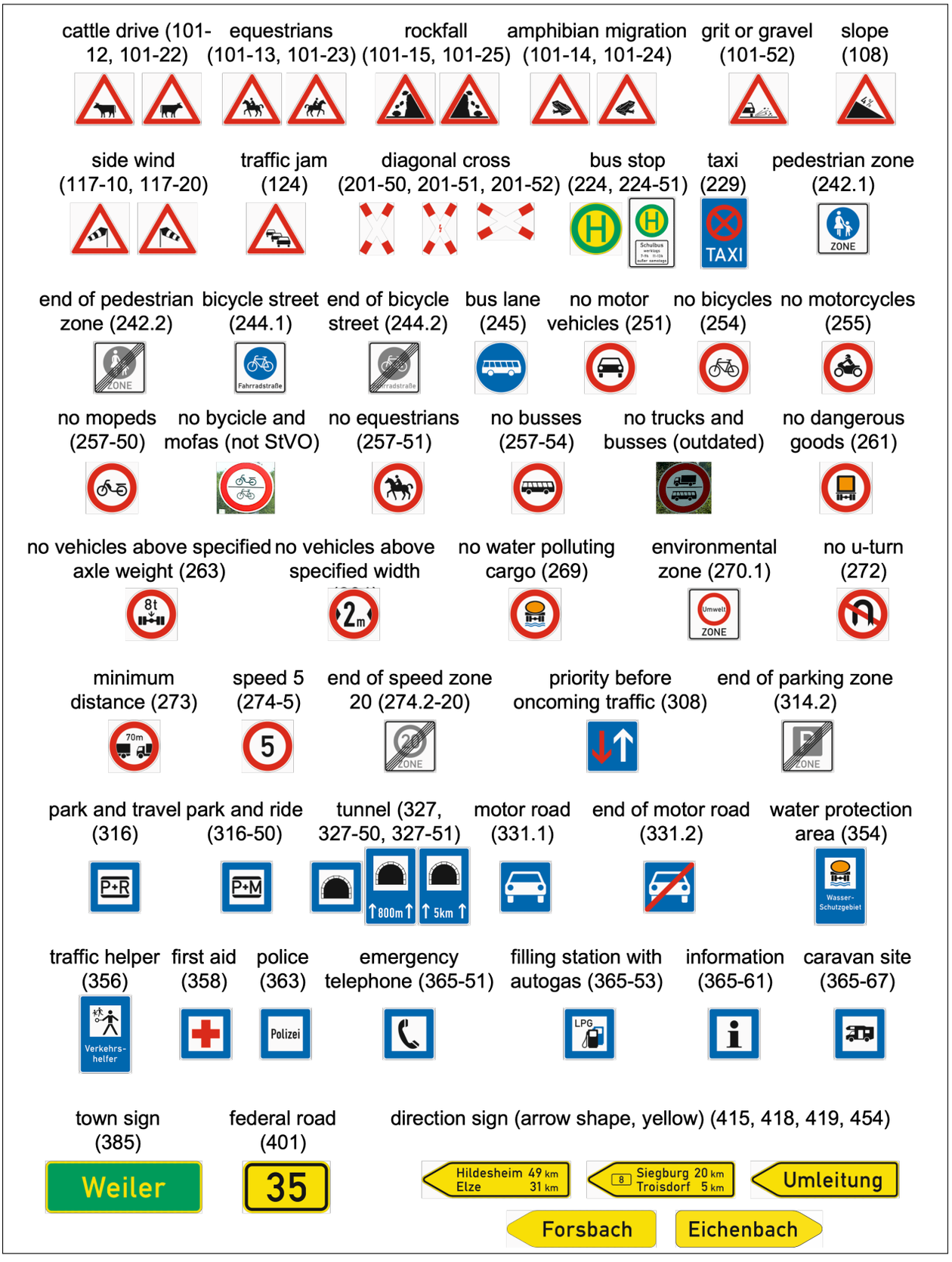

Figure 13: **Unknown test classes** (actual traffic signs) (1/3)

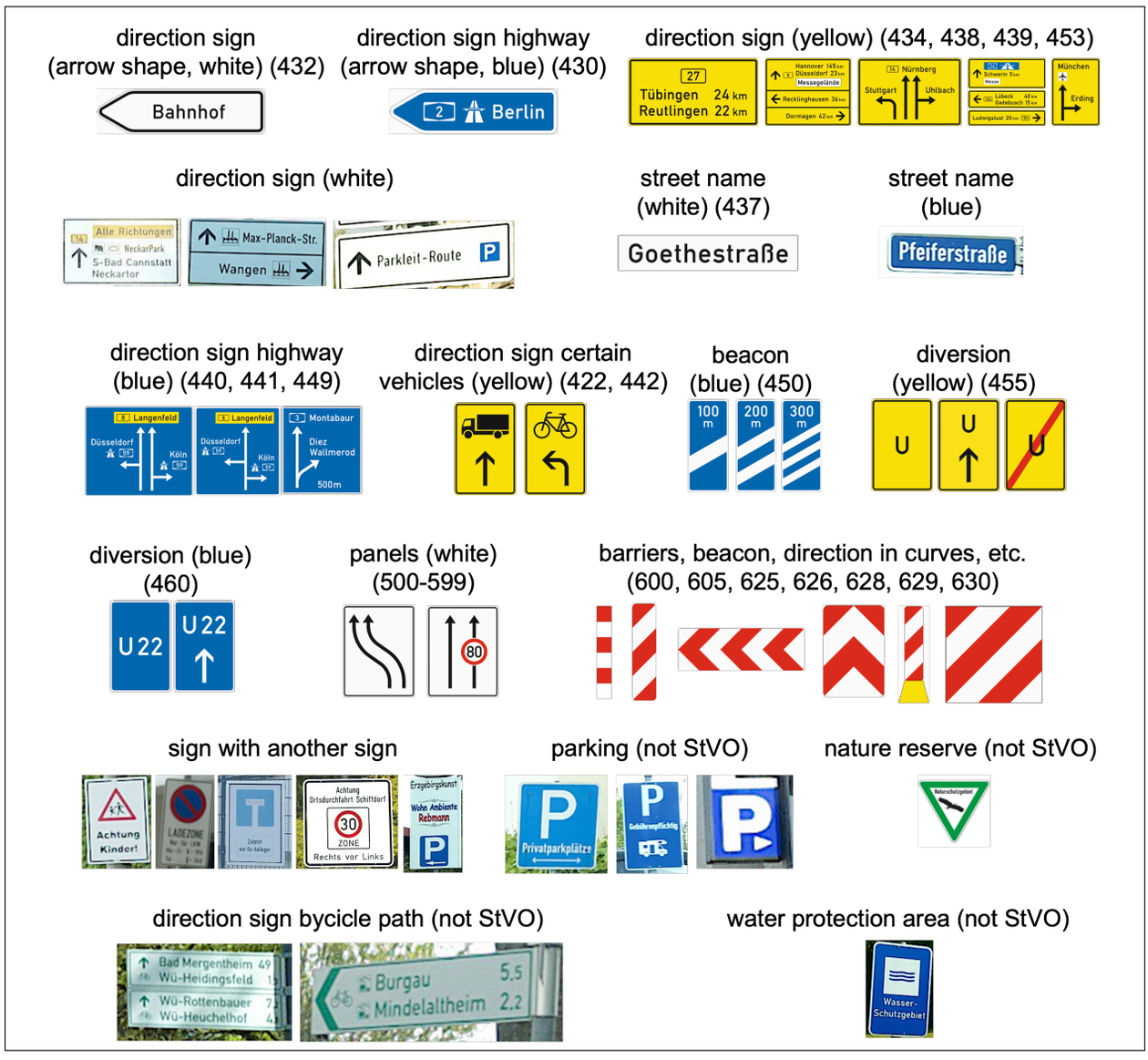

Figure 14: **Unknown test classes** (actual traffic signs) (2/3)

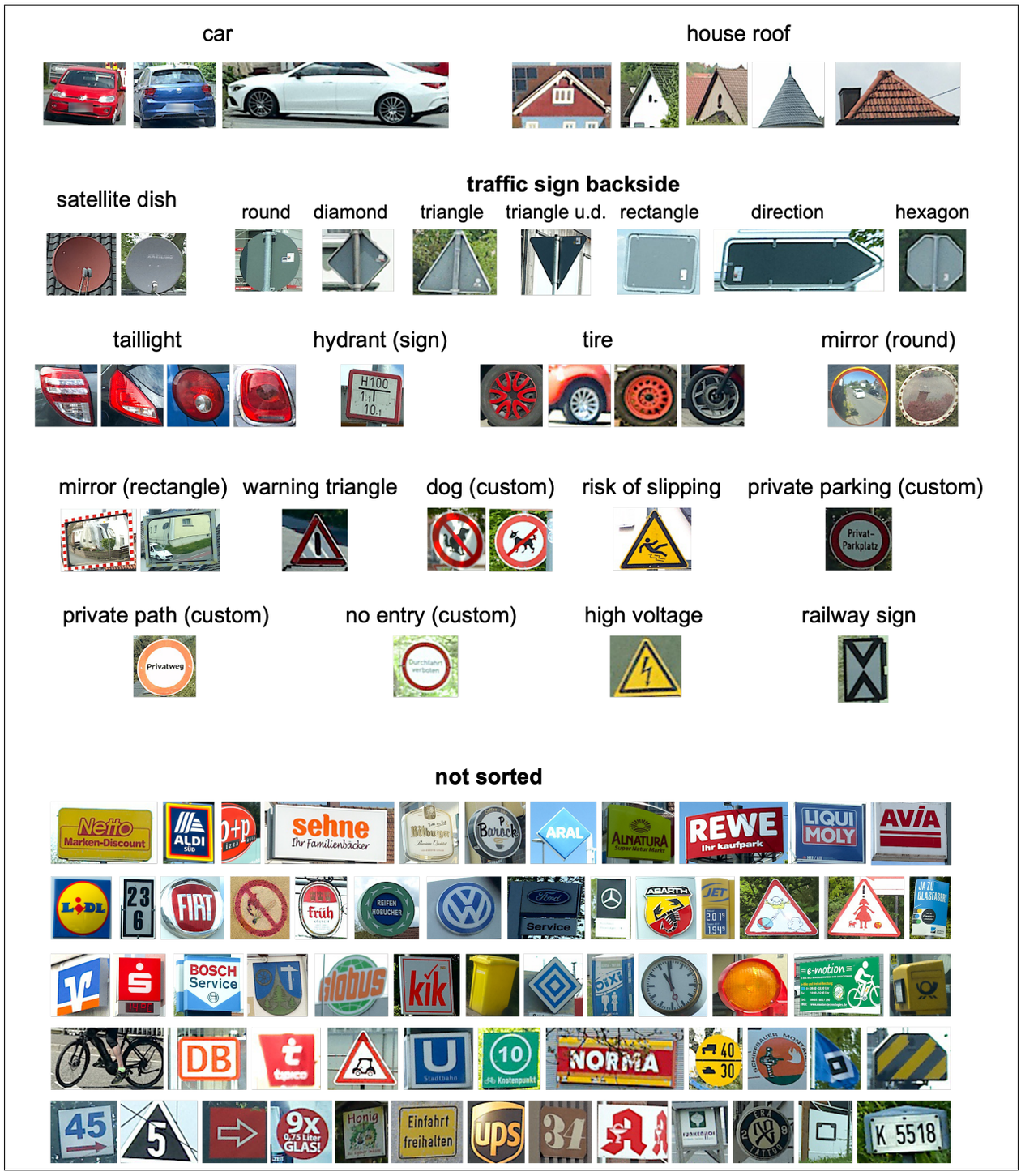

Figure 15: **Unknown test classes** (traffic sign similar objects) (3/3)

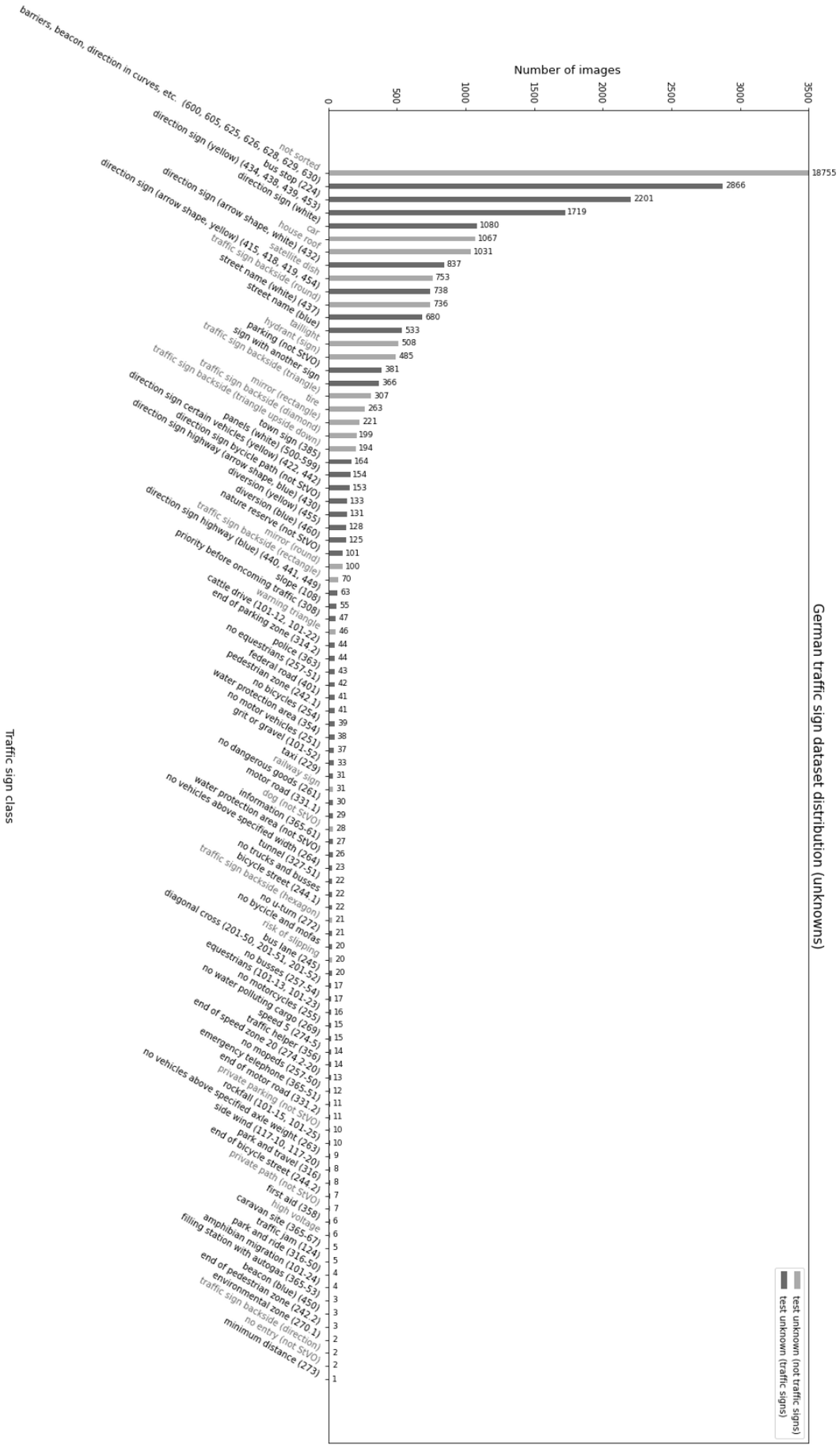

Figure 16: **Class distribution of the unknown classes in the test data set.** Dark gray bars indicate actual traffic signs, and light gray bars indicate objects similar to traffic signs based on color or shape.

Confusion Matrices

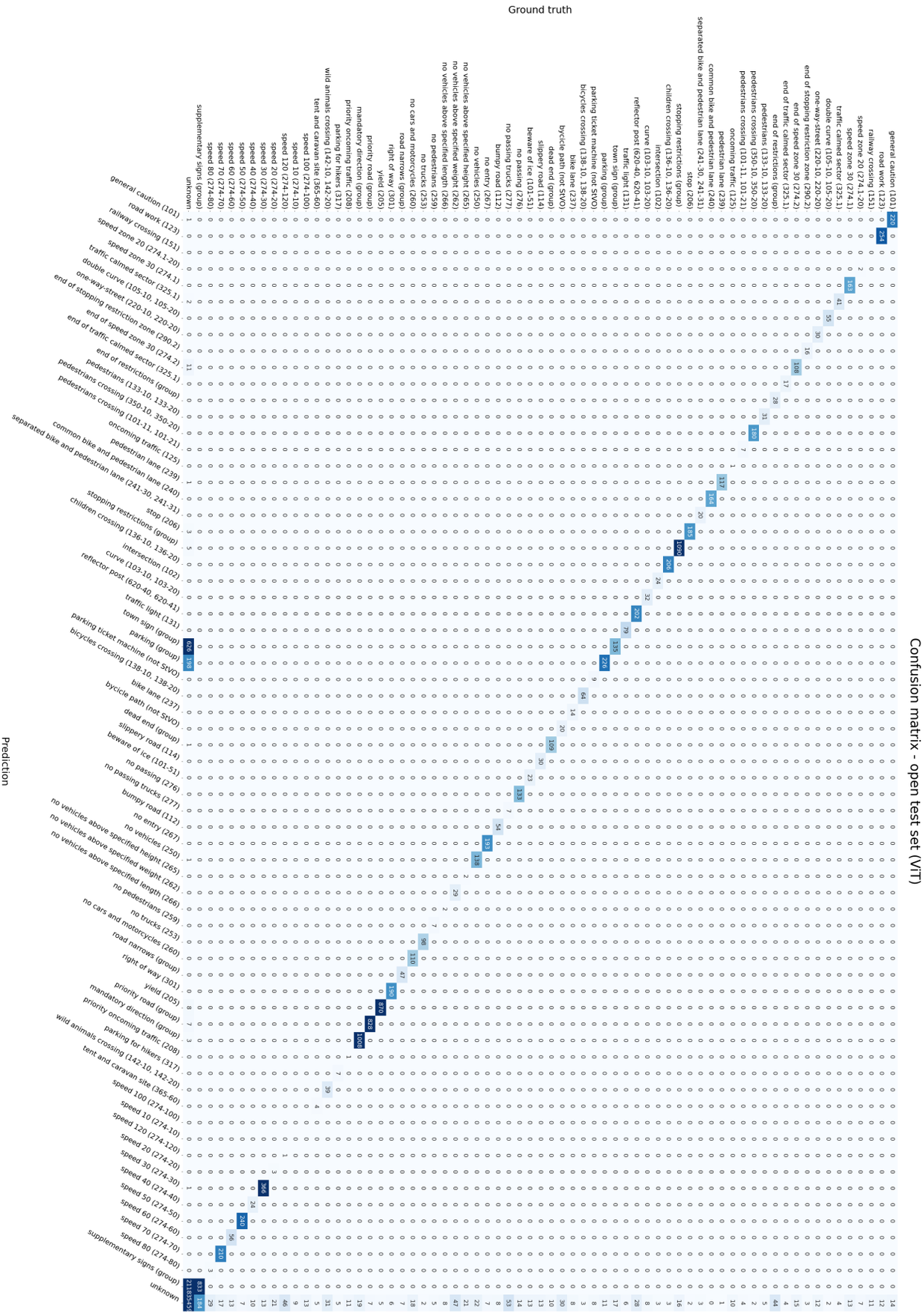

Figure 17: **Confusion matrix - open test set** (ViT). All unknown images are grouped in class *'unknown'*.

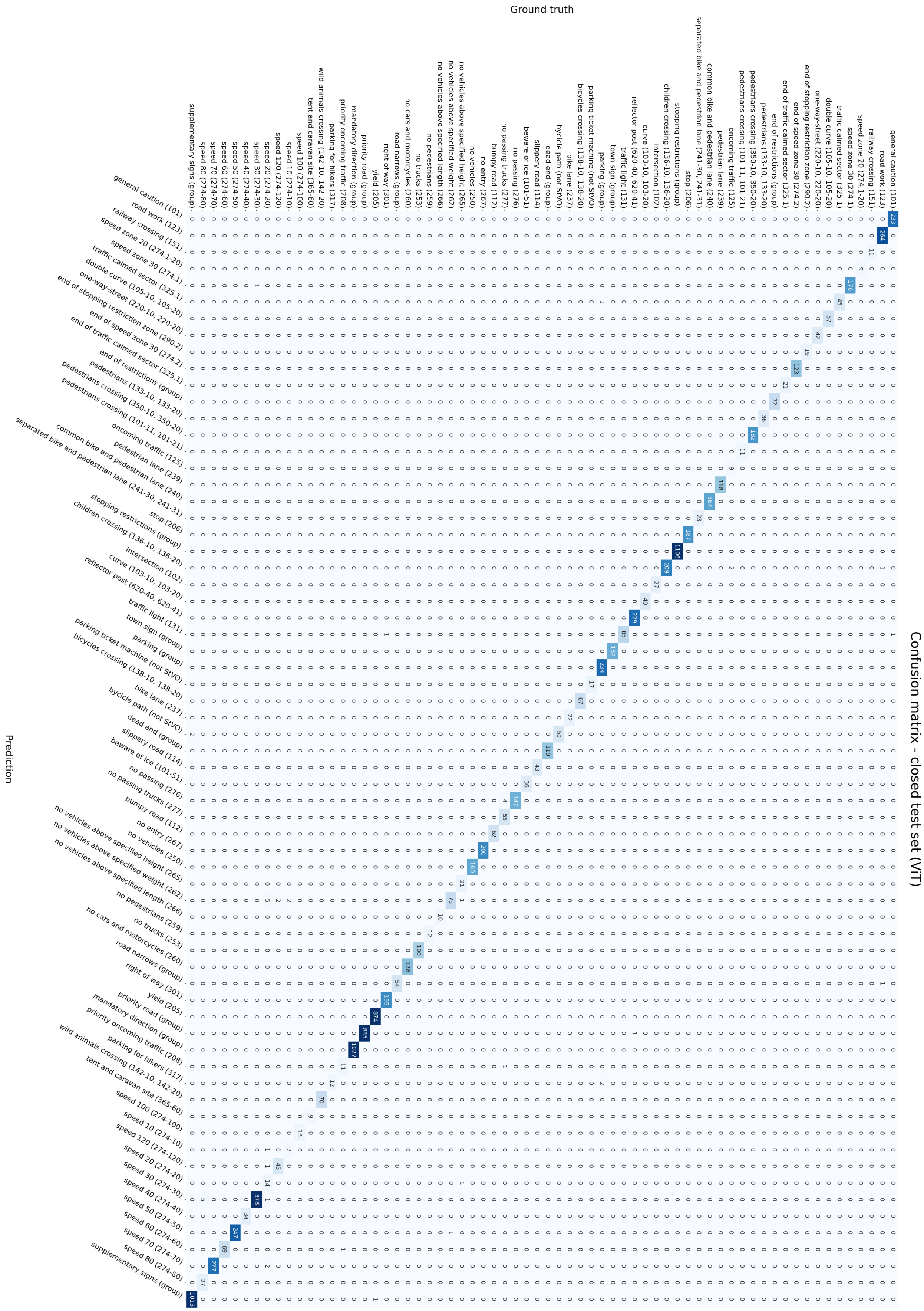

Figure 18: **Confusion matrix - closed test set** (ViT).

Figure 19: **Confusion matrix - open test set** (ARPL+). All unknown images are grouped in class 'unknown'.

Figure 20: **Confusion matrix - closed test set** (ARPL+).

Figure 21: **Confusion matrix - open test set** (Baseline+ MSP). All unknown images are grouped in class 'unknown'.

Figure 22: **Confusion matrix - closed test set** (Baseline+ MSP).