The Mapillary Traffic Sign Dataset for Detection and Classification on a Global Scale

Christian Ertler[0000−0001−6385−7907], Jerneja Mislej[0000−0002−6372−559X], Tobias Ollmann[0000−0002−8224−3692], Lorenzo Porzi[0000−0001−9331−2908], Gerhard Neuhold[0000−0002−1689−7284], and Yubin Kuang[0000−0002−4079−0962]

Facebook

Abstract. Traffic signs are essential map features for smart cities and navigation. To develop accurate and robust algorithms for traffic sign detection and classification, a large-scale and diverse benchmark dataset is required. In this paper, we introduce a new traffic sign dataset of $105K$ street-level images around the world covering 400 manually annotated traffic sign classes in diverse scenes, wide range of geographical locations, and varying weather and lighting conditions. The dataset includes $52K$ fully annotated images. Additionally, we show how to augment the dataset with $53K$ semi-supervised, partially annotated images. This is the largest and the most diverse traffic sign dataset consisting of images from all over the world with fine-grained annotations of traffic sign classes. We run extensive experiments to establish strong baselines for both detection and classification tasks. In addition, we verify that the diversity of this dataset enables effective transfer learning for existing large-scale benchmark datasets on traffic sign detection and classification. The dataset is freely available for academic research[1].

1 Introduction

Robust and accurate object detection and classification in diverse scenes is one of the essential tasks in computer vision. With the development and application of deep learning in computer vision, object detection and recognition has been studied [5,17,24] extensively on general scene understanding datasets [4,11,18]. In terms of fine-grained detection and classification, there are also datasets that focus on general hierarchical object classes [11] or domain-specific datasets, *e.g.* on bird species [32]. In this paper, we will focus on detection and fine-grained classification of traffic signs on a new dataset.

Traffic signs are key map features for navigation, road safety and traffic control. More specifically, traffic signs encode information for driving directions, traffic regulation, and early warning. Accurate and robust perception of traffic signs is also essential for localization and motion planning in different driving scenarios.

As an object class, traffic signs have specific characteristics in their appearance. First of all, traffic signs are in general rigid and planar. Secondly, traffic

[1] www.mapillary.com/dataset/trafficsign

Table 1. Overview of traffic sign datasets. The numbers include only publicly available images and annotations. *Unique* refers to datasets where each traffic sign bounding box corresponds to a unique traffic sign instance (*i.e.* no sequences showing the same physical sign). *70,428/17,666 (train-val/test) signs are within the taxonomy. ** *All* includes train, val, test, and partial (semi) sets. †TT100K provides only 10,000 images containing traffic signs. ‖45 classes have more than 100 examples. ¶MVD contains back *vs.* front classes. ‡video-frames covering only 15,630 unique signs. §signs within the partially annotated set correspond to physical signs within the training set

Dataset	Images	Classes	Signs	Attributes	Region	Boxes	Unique
MTSD (train/val)	**41,909**		***206,386**	occluded, exterior,		✓	✓
MTSD (test)	**10,544**	**400**	***51,155**	out-of-frame, dummy,	**global**	✓	✓
MTSD (all)**	**105,830**		**354,154**	ambiguous, included		✓	§✗
TT100K [35]	†100,000	‖221	26,349	✗	China	✓	✓
MVD [22]	20,000	¶2	174,541	✗	**global**	✓	✓
BDD100K [34]	100,000	1	343,777	✗	USA	✓	✗
GTSDB [10]	900	43	852	✗	Germany	✓	✗
RTSD [26]	‡179,138	156	‡104,358	✗	Russia	✓	✗
STS [13]	3777	20	5582	✗	Sweden	✓	✗
LISA [21]	6610	47	7855	✗	USA	✓	✗
GTSRB [29]	✗	43	39,210	✗	Germany	✗	✗
BelgiumTS [31]	✗	108	8851	✗	Belgium	✗	✗

signs are designed to be distinctive from their surroundings. In addition, there is limited variety in colors and shapes for traffic signs. For instance, regulatory signs in European countries are typically circular with a red border. To some degree, the aforementioned characteristics limit the appearance variation and increase the distinctness of traffic signs. However, traffic sign detection and classification are still very challenging problems due to the following reasons: (1) traffic signs are easily confused with other object classes in street scenes (*e.g.* advertisements, banners, and billboards); (2) reflection, low light condition, damages, and occlusion hinder the classification performance of a sign class; (3) fine-grained classification with small inter-class difference is not trivial; (4) the majority of traffic signs—when appearing in street-level images—are relatively small in size, which requires efficient architecture designs for small objects.

Traffic sign detection and classification have been studied extensively in computer vision [14,20,25,35]. However, these studies were done in relatively constrained settings in terms of the benchmark dataset: the images and traffic signs are collected in a specific country; the number of traffic sign classes is relatively small; the images lack diversity in weather conditions, camera sensors, and seasonal changes. Extensive research is still needed for detecting and classifying traffic signs at a global scale and under varying capture conditions and devices.

The contributions of this paper are manifold:

– We present the **most diverse traffic sign dataset** with $105K$ images from all over the world. The dataset contains over $52K$ fully annotated images,

Fig. 1. *Top*: Taxonomy overview. The sizes are relative to the number of samples within MTSD. *Bottom*: Example images in MTSD with bounding box and class annotations (green boxes without template indicate *other-sign*).

covering 400 known traffic sign classes and other unknown classes, resulting in over $255K$ signs in total.
- Without introducing any additional annotation cost, we show how to augment the dataset with **real semi-supervised samples by propagating labels** to nearby images which helps to get more samples, especially in the long-tail of the class distribution. The dataset includes about $53K$ extra images collected in this way.
- We establish **extensive baselines** for detection and classification on the dataset, shedding light on future research directions.
- We study the impact of **transfer learning** using our traffic sign dataset and other datasets released in the past. We show that pre-training on our dataset boosts average precision (AP) of the binary detection task by 4–6 points, thanks to the completeness and diversity of our dataset.

Related Work. Traffic sign detection and recognition has been studied extensively in the previous literature. The German Traffic Sign Benchmark Dataset (GTSBD) [30] is one of the first datasets that was created to evaluate the classification branch of the problem. Following that, there have also been other traffic sign datasets focusing on regional traffic signs, *e.g.* Swedish Traffic Sign Dataset [12], Belgium Traffic Sign Dataset [20], Russian Traffic Sign Dataset [26], and Tsinghua-Tencent Dataset (TT100K) in China [35]. For generic traffic sign detection (where no class information of the traffic signs is available), there has been work done in the Mapillary Vistas Dataset (MVD) [22] (global) and BDD100K [34] (US only). A detailed overview and comparison of publicly available traffic sign datasets can be found in Table 1.

For general object detection, there has been substantial work on CNN-based methods with two main directions, *i.e.* one-stage detectors [17,19,23] and two-stage detectors [3,5,6,24]. One-stage detectors are generally much faster, trading off accuracy compared to two-stage detectors. One exception is the one-stage RetinaNet [17] architecture that outperforms the two-stage Faster-RCNN [24] thanks to a weighting scheme during training to suppress trivial negative su-

pervision. For simultaneous detection and classification, recent work [2] shows that decoupling the classification from detection head boosts the accuracy significantly. Our work is related to [2] as we also decouple the detector from the traffic sign classifier.

To handle the scale variation of objects in the scene, many efficient multi-scale training and inference algorithms have been proposed and evaluated on existing datasets. For multi-scale training, in [15,27,28], a few schemes have been proposed to distill supervision from different scales efficiently by selective gradient propagation and crop generation. To enable efficient multi-scale inference, feature pyramid networks (FPN) [16] were proposed to utilize lateral connections in a top-down architecture to construct an effective multi-scale feature pyramid from a single image.

To develop the baselines presented in this paper, we have chosen Faster-RCNN [24] with FPN [16] as the backbone. Given the aforementioned characteristics of traffic sign imagery, we have also trained a separate classifier for fine-grained classification as in [2]. We elaborate on the details of our baseline methods in Section 4 and Section 5.

2 Mapillary Traffic Sign Dataset

In this section, we present a large-scale traffic sign dataset called Mapillary[2] Traffic Sign Dataset (MTSD) including $52K$ images with $257K$ fully annotated traffic sign bounding boxes and corresponding class labels. Additionally, it includes a set of over $53K$ nearby images with more than $84K$ semi-supervised class labels, making it more than $105K$ images. In the following we describe how the dataset was created and present our traffic sign class taxonomy consisting of 400 classes. Examples can be found in Figure 1 (bottom).

2.1 Image Selection

There are various conventions for traffic signs in different parts of the world leading to strong appearance differences. Even within a single country, the distribution of signs is not uniform: some signs occur only in urban areas, some only on highways, and others only in rural areas. With MTSD, we present a dataset that covers this diversity uniformly. In order to do so, a proper preselection of images for annotation is crucial. The requirements for this selection step are: (1) to have a uniform geographical distribution of images around the world, (2) to cover images of different quality, captured under varying conditions, (3) to include as many signs as possible per image, and (4) to compensate for the long-tailed distribution of potential traffic sign classes.

In order to get a pool of pre-selected images satisfying the aforementioned requirements, we sample images in a per-country manner. The fraction of target

[2] www.mapillary.com/app is a street-level imagery platform hosting images collected by members of their community.

images for each country is derived from the number of images available in that country and its population count weighted by a global target distribution over all continents (*i.e.* 20 % North America, 20 % Europe, 20 % Asia, 15 % South America, 15 % Oceania, 10 % Africa). We further make sure to cover both rural and urban areas within each country by binning the sampled images uniformly in terms of their geographical locations and sample random images from each of the resulting bins. In the last step of our image sampling scheme, we prioritize images containing at least one traffic sign instance according to the traffic sign detections given by the Mapillary API and make sure to cover various image resolutions, camera manufacturers, and scene properties[3]. Additionally, we add a distance constraint so that selected images are far away from each other in order to avoid highly correlated images and traffic sign instances. Statistics of the dataset can be found in Section 3.

2.2 Traffic Sign Class Taxonomy

Traffic signs vary across different countries. For many countries, there exists no publicly available and complete catalogue of signs. The lack of a known set of traffic sign classes leads to challenges in assigning class labels to traffic signs annotated in MTSD. The potential magnitude of this unknown set of traffic signs is in the thousands as indicated by the set of template images described in Section 2.3.

For MTSD, we did a manual inspection of the templates that have been chosen by the annotators and selected a subset of them to form the final set of 400 classes included in the dataset as visualized in Figure 1 (top). This subset was chosen and grouped such that there are no overlaps or confusions (visual or semantic) among the classes. All these classes defined by disjoint sets of templates build up our traffic sign class taxonomy. We map all annotated traffic signs in MTSD that have a template selected within this taxonomy to a class label. We would like to emphasize that our flexible taxonomy allows us to incrementally extend MTSD. It enables to add more classes by grouping templates to new classes and mapping traffic sign instances to these new classes based on already assigned templates.

2.3 Annotation Process

The process of annotating an image including image selection approval, traffic sign localization by drawing bounding boxes, and class label assignment for each box is a complex and demanding task. To improve efficiency and quality, we split it into 3 consecutive tasks, with each having its own quality assurance process. All tasks were done by 15 experts in image annotation after being trained with explicit specifications for each task.

[3]Details on how scene properties are defined and derived are included in the supplementary materials.

Image Approval. Since initial image selection was done automatically based on the heuristics described in Section 2.1, the annotators needed to reject images that did not fulfill our criteria. In particular, we do not include non-street-level images or images that have been taken from unusual places or viewpoints. Further we discarded images of very low quality that could not be used for training (*i.e.* extremely blurry or overexposed). However, we still sample images of low quality in the dataset which include recognizable traffic signs as these are good examples to evaluate recognition of traffic signs in real-world scenarios.

Sign Localization. In this task, the annotators were instructed to localize all traffic signs in the images and annotate them with bounding boxes. In contrast to previous traffic sign datasets where only specific types of traffic signs have been annotated (*e.g.* TT100K [35] includes only standard circular and triangular shaped signs), MTSD contains bounding boxes for all types of traffic related signs including direction, information, highway signs, *etc.*

To speed up the annotation process, each image was initialized with bounding boxes of traffic signs extracted from the *Mapillary* API. The annotators were asked to correct all existing bounding boxes to tightly contain the signs (or reject them in cases of false positives) and to annotate all missing traffic signs if their shorter sides were larger than 10 pixels. We provide a statistical analysis of the manual interactions of the annotators in supplemental material.

Sign Classification. This task was done independently for each annotated traffic sign. Each traffic sign (together with some image context) was shown to the annotators who were asked to provide the correct class label. This is not trivial, since the number of traffic sign classes is large. To the best of our knowledge, there is no globally valid traffic sign taxonomy available; even then, it would be impossible for the annotators to keep track of the different traffic sign classes.

To overcome this issue, we used a set of previously harvested template images of traffic signs from Wikimedia Commons [33] and grouped them by similarity in appearance and semantics. This set of templates (together with their grouping) defines the possible set of traffic sign classes that can be selected by the annotators. In fact, we store an identifier of the actual selected template which allows us to link the traffic sign instances to our flexible traffic sign taxonomy without even knowing the final set of classes beforehand (see Section 2.2).

Since it would still be too time-consuming to scroll through the entire list of templates to choose the correct one out of thousands, we trained a neural network to learn an embedding space (with the grouped template images) which is predicting the similarities between an arbitrary image of a traffic sign instance and the templates. We used this proposal network to assist the annotators in choosing the correct template by pre-sorting the template list for each individual traffic sign.

Specifically, we use a metric learning approach [1] to train a 3-layer network (similar to but shallower than the baseline classification network mentioned in Section 5) to learn a function $f(x) : \mathbb{R}^d \rightarrow \mathbb{R}^k$ that maps a d-dimensional input vector to a k-dimensional embedding space. In our case, x are input images encoded as vectors of size $d = 40 \times 40 \times 3$ and $k = 128$. We train the network

with a contrastive loss [7] such that the cosine similarity

$$\mathrm{sim}(x_1, x_2) = \frac{x_1^T x_2}{\|x_1\|_2 \, \|x_2\|_2} \tag{1}$$

between two embedding vectors x_1 and x_2 with group labels $\hat{y}_1$, $\hat{y}_2$ should be high if the samples are within the same template group, whereas the similarity should be lower than a margin m if the samples are from different groups:

$$\mathcal{L} = \begin{cases} 1 - \mathrm{sim}(x_1, x_2), & \text{if } \hat{y}_1 = \hat{y}_2 \\ \max\left[0, \mathrm{sim}(x_1, x_2) - m\right] & \text{else} \end{cases} . \tag{2}$$

We choose $m = 0.2$ and train the network using a generated training set by blending our traffic sign templates to random background images after scaling, rotating and sheering it by a reasonable amount.

For cases in which this strategy fails to provide a matching template, we provided a text-based search for templates. For details about the annotation UI, we refer to the supplemental material.

Additional Attributes. In addition to bounding boxes and the matching traffic sign templates, the annotators were asked to provide additional attributes: *occluded* if the sign is partly occluded; *ambiguous* if the sign is not classifiable at all (*e.g.* too small, of bad quality, heavily occluded *etc.*); *dummy* if it looks like a sign but is not (*e.g.* car stickers, reflections, *etc.*); *out-of-frame* if the sign is cut off by the image border; *included* if the sign is part of another bigger sign; and *exterior* if the sign includes other signs. Some of these attributes were assigned during localization (if context information is needed). The rest was assigned during classification.

Annotation Quality. All annotations in MTSD were done by expert annotators going through a thorough training process. Their work was monitored by a continuous quality control (QC) process to quickly identify problems during annotation. Moreover, our step-wise annotation process (*i.e.* approval followed by localization followed by classification) ensures that each traffic sign was seen by at least two annotators. The second annotator operating in the classification step was able to reject false positive signs or to report issues with the bounding box in which case the containing image was sent back to the localization step.

In additional quality assurance (QA) experiments done by a 2^{nd} annotator on $5K$ images including $26K$ traffic signs, we found that (1) only 0.5 % of bounding boxes needed correction; (2) the false negative rate was 0.89 % (corresponding to a total number of only 212 missing signs, most of them being very small); (3) the false positive rate was at 2.45 %. Note that this is in the localization step before classification, where a second annotator has been asked to classify the sign and could potentially fix false positives.

2.4 Partial Annotations

In addition to the fully-annotated images, we provide another set of images with partially annotated bounding boxes and semi-supervised class labels. Given

Fig. 2. Example from the partially annotated set. The leftmost image is from the fully annotated set. The 3 other images show the same sign from different perspectives in the partial set with propagated class labels. Best viewed zoomed in and in color.

the fully-annotated images, the annotations of this set of images are generated automatically in a semi-supervised way.

We achieve this by finding correspondences between the manual annotations in the fully-annotated images and automatic detections in geographically neighboring images from the Mapillary API. To find these correspondences, we first use Structure from Motion (SfM) [8] to recover the relative camera poses between the fully-annotated images and the partially annotated images. With these estimated relative poses, we find correspondences between annotated signs and automatically detected signs by triangulating and verifying the re-projection errors for the centers of the bounding boxes between multiple images. Having these correspondences, we propagate the manually annotated class labels to the automatic detections in the partially annotated images. Since there is no guarantee that all traffic signs are detected through Mapillary's platform, this results in a set of images with partially annotated bounding box annotations. Note that, for unbiased evaluation, we ensure that the extension is done only in the geographical neighborhood of images in the training set (based on the split discussed in Section 2.5). Example images can be found in Figure 2 and the effect on the class distribution in Figure 3 (top/right). A more detailed description of how this set was created can be found in supplemental material.

2.5 Dataset Splits

As common practice with other datasets such as COCO [18], MVD [22] and PASCAL VOC [4], we split MTSD into training, validation and test sets, consisting of 36,589, 5320, and 10,544 images, respectively. We provide the image data for all sets as well as the annotations for the training and validation set; the annotations for the test set will not be released in order to ensure a fair evaluation. Additionally, we provide a set of 53,377 images with partial annotations as discussed in Section 2.4 for training as well.

Each split is created in a way to match the distributions described in Section 2.1. Especially, we ensure that the distribution of class instances is similar for each split, to avoid rare classes being under-represented in the smaller sets (*i.e.* validation/test sets). The same holds true for the additional sign attributes (*e.g.* *ambiguous*, *etc.*).

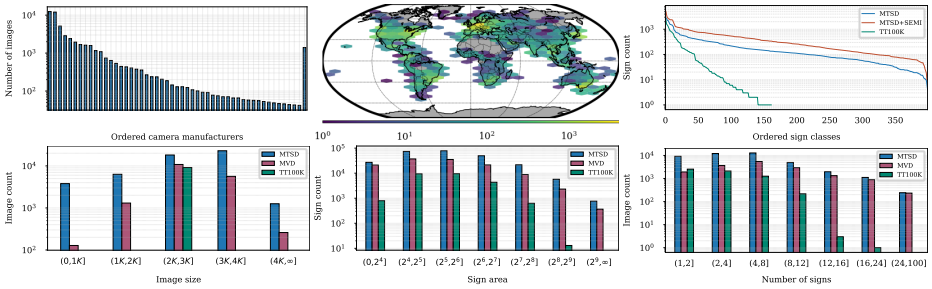

Fig. 3. *Top:* Distribution of camera devices; Geographical distribution of images; Distribution of traffic sign classes. *Bottom:* Images binned by size; Signs binned by sizes; Images binned by #signs. Size bins in $\sqrt{\text{pixel area}}$.

3 Statistics

In this section, we provide image and traffic sign statistics of MTSD and compare with previous datasets (TT100K [35] and MVD [22]). Unless stated otherwise, all numbers refer to the fully-supervised set of MTSD only.

3.1 Image Properties

For a dataset to reflect a real-world image capturing setting with diverse geographical distribution, the image selection strategy described in Section 2.1 used for MTSD ensures a good distribution over different capturing settings.

Camera Sensors. In Figure 3 (top/left), we show the distribution of camera manufacturers used for capturing the images of MTSD. In total, the dataset covers over 200 different sensor manufacturers (we group the tail of the distribution for displaying purposes) which results in a large variety of image properties similar to the properties described in [22]. This is in contrast to the setup used for TT100K [35] which contains only images taken by a single sensor setup, making MTSD more challenging in comparison.

Image Sizes. The diversity in camera sensors further results in a diverse distribution over image resolutions as shown in Figure 3 (bottom/left). MTSD covers a broad range of image sizes starting from low-resolution images with 1 MPixels going up to images of more than 16 MPixels. Additionally, we include 1138 360-degree panoramas stored as standard images with equi-rectangular projection. Besides the overall larger image volume compared to other datasets, MTSD also covers a larger fraction of low-resolution images, which is especially interesting for pre-training and validating detectors applied on similar sensors, *e.g.* built-in automotive cameras. For comparison, TT100K only contains images of 2048^2 px and even for this resolution the volume of images is smaller than in MTSD.

Geographical Distribution. The heat map in the middle of Figure 3 (top) shows the resulting geographical distribution of the images, covering almost all habitable areas of the world with higher density in populous areas.

Fig. 4. Results from our detection and classification baseline on the validation set (green colored: true positive, red: missing detections).

3.2 Traffic Sign Properties

The fully-annotated set of MTSD includes a total number of 257,541 traffic sign bounding boxes out of which more than $88K$ have a class label within our taxonomy covering 400 different traffic sign classes. The remaining traffic signs sum up as ambiguous signs, directional signs, information signs, highway shields, exterior signs, barrier signs, and other signs that do not fall into our taxonomy.

Class Distribution. The right plot in Figure 3 (top) shows a comparison of the traffic sign class distribution between MTSD and TT100K. Note that MVD is not included here since it does not have labels of traffic sign classes. MTSD has approximately twice as many traffic sign classes as TT100K; if we use the definition of a trainable class in [35] (which are classes with at least 100 traffic sign instances within the dataset) this factor increases to approximately 3 between TT100K and MTSD. This difference gets even higher if we consider the instances from the partially annotated set of MTSD.
Sign Sizes. The plot in the middle of Figure 3 (bottom) compares the areas of signs in terms of pixels in the original resolution of the containing image. MTSD covers a broad range of traffic sign sizes with an almost uniform distribution up to 256^2 px. MVD has a similar distribution with a lower overall volume. In comparison to TT100K, MTSD provides a higher fraction of extreme sizes which poses another challenge for traffic sign detection.
Signs per Image. Finally, the plot on the right of Figure 3 (bottom) shows the distribution of images over the number of signs within the image. Besides the higher volume of images, MTSD contains a larger fraction of images with a large number of traffic sign instances (*i.e.* > 12). One reason for this is that the annotations in MTSD cover all types of traffic signs, whereas TT100K only contains annotations for very specific types of traffic signs in China.

4 Traffic Sign Detection

One task defined on MTSD is binary detection of traffic signs, *i.e.* localization without inferring specific class labels. The goal is to predict a set of axis-aligned bounding boxes with corresponding confidence scores for each image.

Metrics. Given a set of detections with estimated scores for each image, we first compute the matching between the detections and annotated ground truth

within each image separately. A detection can be successfully matched to a ground truth if their Jaccard overlap (IoU) [4] is > 0.5; if multiple detections match the same ground truth, only the detection with the highest score is a match while the rest is not (*double detections*); each detection will only be matched to one ground truth bounding box with the highest overlap.

Having this matching indicator (TP vs. FP) for every detection, we define average precision (AP) similar to COCO [18] (*i.e.* $\text{AP}^{\text{IoU}=0.5}$ which resembles AP definition of PASCAL VOC [4]). Specifically, we compute precision as a function of recall by sorting the matching indicators by their corresponding detection confidence scores in descending order and accumulate the number of TPs and FPs. AP is defined as the area under the curve of this step function. Additionally, we follow [18] and compute AP in different scales: AP_s, AP_m, and AP_l refer to AP computed for boxes with area $a < 32^2$, $32^2 < a < 96^2$, and $a > 96^2$.

Baseline and Results. In Table 2, we show experimental results using a Faster R-CNN based detector [24] with FPN [16] and residual networks [9] as the backbone. During training we randomly sample crops of size 1000×1000 at full resolution instead of down-scaling the image to avoid vanishing of small traffic signs, as traffic signs can be very small in terms of pixels and MTSD covers traffic signs from a broad range of scales in different image resolutions. We use a batch size of 16, distributed over 4 GPUs for FPN50 models; for FPN101 models, we use batches of size 8. Unless stated otherwise, we train using stochastic gradient descent (SGD) with an initial learning rate of 10^{-2} and lower the learning rate when the validation error plateaus. For inference, we down-scale the input images such that their larger side does not exceed a certain number of pixels (either 2048 px or 4000 px) or operate on full resolution if the original image is smaller.

Besides training on MTSD, we conduct transfer-learning experiments on TT100K and MVD[4] to test the generalization properties of the proposed dataset. We use the same baseline as for the MTSD experiments and train it on both datasets, one with ImageNet initialization and one with MTSD initialization. The models trained with ImageNet initialization are trained to convergence. To ensure a fair comparison, we fine-tune only for half the number of epochs when initializing with MTSD weights. The results in Table 2 show that MTSD pre-training boosts detection performance by a large margin on both datasets, regardless of the input resolution. This is a clear indication for the generalization qualities of MTSD.

5 Simultaneous Detection and Classification

The second task on MTSD is simultaneous detection and classification of traffic signs, *i.e.* multi-class detection. It extends the detection task to demand a class

[4] We convert the segmentations of *traffic-sign–front* instances to bounding boxes by taking the minimum and maximum in the x, y axes. Note that this conversion can be inaccurate if signs are occluded.

Table 2. Detection baseline results on MTSD, TT100K and MVD.Numbers in brackets refer to absolute improvements when pre-training on MTSD in comparison to ImageNet. *[35] using multi-scale inference with scales 0.5, 1, 2, and 4

	Max 4000px				Max 2048px			
	AP	AP_s	AP_m	AP_l	AP	AP_s	AP_m	AP_l
MTSD								
FPN50 ours	87.84	72.91	91.88	93.54	80.08	52.12	88.81	94.72
FPN101 ours	88.38	73.89	92.10	93.69	81.65	56.32	89.18	94.80
TT100K								
multi-scale*	91.79	84.56	96.40	92.60	-	-	-	-
FPN50 ours	-	-	-	-	91.27	84.01	95.87	90.13
+ *MTSD*	-	-	-	-	**97.60** (+6.33)	93.13	99.03	98.44
MVD (traffic signs)								
FPN50 ours	72.90	46.60	79.93	85.42	64.00	30.70	75.28	86.50
+ *MTSD*	**76.31** (+3.41)	51.00	83.49	88.33	**68.29**	33.60	79.45	89.53

label for each traffic sign instance within our taxonomy. For instances that do not have a label within our taxonomy, we introduce a general class *other-sign*.

Metric. The metric for this task is mean average precision (mAP) over all 400 classes; per-class AP is calculated as described in Section 4. The matching between predicted and ground truth boxes is done in a binary way by ignoring the class label. After that, we filter out all *other-sign* ground truth instances and detections since we do not want to evaluate on this general class.

Baseline. A trivial baseline for this task would be to extend the binary detection baseline from Section 4 to the multi-class setting by adding a 401-way classification head. However, preliminary experiments showed that a straightforward training of such a model does not yield acceptable performance. We hypothesize that this is due to (1) scale issues for small signs before RoI pooling and, (2) under-represented class variation within the training batches given that the majority of traffic sign instances are *other-sign*.

To overcome the scale issue and to have better control over batch statistics during training, we opted for a two-stage architecture that uses our binary detectors in the first stage and a decoupled shallow classification network in the second stage. Such decoupling has been shown to improve detection and recognition accuracy [2]. The classification network consists of seven 3×3 convolutions (each followed by batch normalization) with 2×2 max-pooling layers after the 2^{nd} and 6^{th} convolution layer. We start with 32 features in the first layer and double this number after each pooling layer. The last convolution is followed by spatial average pooling and a fully-connected layer with 256 features resulting in a 401-way classification head with softmax activation (400 and *other-sign*) and a single sigmoid activation for foreground/background classification.

We use image crops predicted by the detector (both foreground and background) together with crops from the ground truth scaled to 40×40 px as input and optimize the network using cross-entropy loss. To balance the distribution

Table 3. Simultaneous detection and classification results. *+ det/cls MTSD* refer to MTSD pre-training of detection/classification models. The numbers in brackets are absolute improvements over [35]

	mAP		mAP_s	mAP_m	mAP_l
MTSD					
FPN50 ours	81.7		73.0	84.1	84.2
FPN101 ours	81.8		74.4	84.4	84.9
TT100K					
multi-scale	81.6		68.3	86.5	85.7
FPN50 ours	**89.9**	(+8.3)	83.9	93.0	84.3
+ *det MTSD*	**93.4**	(+11.8)	88.2	94.8	93.6
+ *cls MTSD*	**95.7**	(+14.1)	91.3	96.9	96.7

of traffic sign classes in a batch, we uniformly sample 100 classes with 4 samples each and add another 100 background crops per batch. We train the network with SGD for 50 epochs starting with a learning rate of 10^{-3} lowered by a factor of 0.1 after 30 and 40 epochs.

Results. We show results of our baseline in Table 3. Our classifier with FPN101 binary detector reaches 81.8 mAP over all 400 classes. Figure 4 shows visual examples of our baseline's predictions and Figure 5 (left) shows typical failure cases of the classification network.

To verify our baseline, we train with the same setup on TT100K and compare the results with the baseline in [35][5]. Our two-stage approach outperforms their baseline by 8.3 points, even though the performances of the binary detectors are similar (see Table 2). This validates that the decoupled classifier, even with a shallow network, is able to yield good results. The accuracy is further improved when we pre-train the classifier and the detector on MTSD and then fine-tune them on TT100K, which further validates the generalization effectiveness of MTSD.

6 Classification with Partial Annotations

To evaluate the quality and existence of complementary information in our partially annotated, semi-supervised training set, we conduct classification experiments with and without the additional traffic sign samples and evaluate the performance on $53.3K$ traffic sign crops ($17.6K$ of the MTSD test set + $35.7K$ additional crops). After 14 independent trainings for each data configuration, we found a consistent improvement of 0.6 points in terms of mean class accuracy (standard deviation of multiple trainings is shown in Figure 5 (right). However, major improvements can be found in the long-tail of the class distribution where we have limited numbers of fully-supervised annotations. Figure 5 (right) shows

[5] We convert their results to the format used by MTSD and evaluate using our metrics.

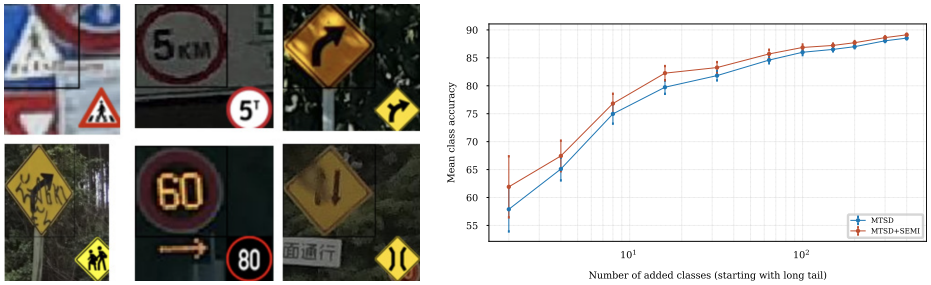

Fig. 5. Left: Failure cases of the classification network on MTSD. Right: Evaluation of the semi-supervised training set with varying number of classes. Reporting mean/std of multiple trainings.

mean class accuracy over varying number of classes where classes are added starting from the long-tail of the original class distribution. We can see a consistent gain of 1–4 points for the long-tail up to the first 100 classes. The gain decreases but is still noticeable when we add more classes that are well represented in the fully-supervised training set.

This shows the value of our partially annotated set as a straightforward way to augment existing datasets to better represent the long-tail of classes without introducing additional labeling costs. We want to point out that this method is a continuous data source for additional training data as it can be repeated as often as new nearby images become available.

7 Conclusion

In this work, we introduce MTSD, a large-scale traffic sign benchmark dataset that includes $105K$ images with full and partial bounding-box annotations, covering 400 traffic sign classes from all over the world. MTSD is the most diverse traffic sign benchmark dataset in terms of geographical locations, scene characteristics, and traffic sign classes. We show in baseline experiments that decoupling detection and fine-grained classification yields superior results on previous traffic sign datasets. Additionally, in transfer-learning experiments, we show that MTSD facilitates fine-tuning and improves accuracy substantially for traffic sign datasets in a narrow domain.

We see MTSD as the first step to drive the research efforts towards solving fine-grained traffic sign detection and classification at a global scale. With the partial annotated set, we show a new scalable way to collect additional training images without the need of extra manual annotation work. Moreover, we also see it paving the way for further research in semi-supervised learning in both classification and detection. In the future, we would like to extend MTSD towards a complete traffic sign taxonomy globally. To achieve this, we see the potential of applying zero-shot learning to efficiently model the semantic and appearance attributes of traffic sign classes.

References

1. Bellet, A., Habrard, A., Sebban, M.: A survey on metric learning for feature vectors and structured data. arXiv preprint arXiv:1306.6709 (2013) 6
2. Cheng, B., Wei, Y., Shi, H., Feris, R., Xiong, J., Huang, T.: Revisiting rcnn: On awakening the classification power of faster rcnn. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 453–468 (2018) 4, 12
3. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. In: Proceedings of the Conference on Neural Information Processing Systems (NIPS). pp. 379–387 (2016) 3
4. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision (IJCV) **111**(1), 98–136 (2015) 1, 8, 11
5. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1440–1448 (2015) 1, 3
6. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 580–587 (2014) 3
7. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1735–1742 (2006) 7
8. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge university press (2003) 8
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016) 11
10. Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., Igel, C.: Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN) (2013) 2
11. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:1811.00982 (2018) 1
12. Larsson, F., Felsberg, M.: Using fourier descriptors and spatial models for traffic sign recognition. In: Proceedings of Scandinavian Conference on Image Analysis (SCIA) (2011) 3
13. Larsson, F., Felsberg, M., Forssen, P.E.: Correlating Fourier descriptors of local patches for road sign recognition. IET Computer Vision **5**(4), 244–254 (2011) 2
14. Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., Yan, S.: Perceptual generative adversarial networks for small object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1222–1230 (2017) 2
15. Li, Y., Chen, Y., Wang, N., Zhang, Z.: Scale-aware trident networks for object detection. arXiv preprint arXiv:1901.01892 (2019) 4
16. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2117–2125 (2017) 4, 11
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2980–2988 (2017) 1, 3

18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Proceedings of the European Conference on Computer Vision (ECCV) (2014) 1, 8, 11
19. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 21–37 (2016) 3
20. Mathias, M., Timofte, R., Benenson, R., Van Gool, L.: Traffic Sign Recognition—How far are we from the Solution? In: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN) (2013) 2, 3
21. Mogelmose, A., Trivedi, M.M., Moeslund, T.B.: Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. IEEE Transactions on Intelligent Transportation Systems (ITS) **13**(4), 1484–1497 (2012) 2
22. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 2, 3, 8, 9
23. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7263–7271 (2017) 3
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proceedings of the Conference on Neural Information Processing Systems (NIPS). pp. 91–99 (2015) 1, 3, 4, 11
25. Sermanet, P., LeCun, Y.: Traffic sign recognition with multi-scale convolutional networks. In: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN). pp. 2809–2813 (2011) 2
26. Shakhuro, V., Konushin, A.: Russian traffic sign images dataset. Computer Optics **40**(2), 294–300 (2016) 2, 3
27. Singh, B., Davis, L.S.: An analysis of scale invariance in object detection snip. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3578–3587 (2018) 4
28. Singh, B., Najibi, M., Davis, L.S.: Sniper: Efficient multi-scale training. In: Proceedings of the Conference on Neural Information Processing Systems (NIPS). pp. 9333–9343 (2018) 4
29. Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural Networks (2012) 2
30. Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: The german traffic sign recognition benchmark: A multi-class classification competition. In: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN) (2011) 3
31. Timofte, R., Zimmermann, K., Van Gool, L.: Multi-view traffic sign detection, recognition, and 3d localisation. Machine Vision and Applications **25**(3), 633–647 (2014) 2
32. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011) 1
33. Wikimedia commons. https://commons.wikimedia.org, accessed: 2019-11-11 6
34. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv preprint arXiv:1805.04687 (2018) 2, 3

35. Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., Hu, S.: Traffic-sign detection and classification in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2, 3, 6, 9, 10, 12, 13