

Fully Automated Traffic Sign Substitution in Real-World Images for Large-Scale Data Augmentation

Daniela Horn* and Sebastian Houben*

Abstract—Video-based traffic sign recognition is a key ability of autonomous vehicles but a demanding challenge due to the enormous number of classes and natural conditions in the wild. We address this problem with a fully automatic close-to-life image-to-image translation technique for traffic sign substitution in natural images (cf. Fig. 1). The work is intended as data augmentation technique and allows for training rare or unavailable traffic sign classes, or otherwise uncommon cases in visual traffic sign detection and classification. To this end, we extend our previous data generation model [1] and propose a rendering pipeline to create convincing traffic sign images with realistic background and camera recording artifacts.

Experiments are conducted by exchanging traffic sign classes on different parts of the German Traffic Sign Recognition Benchmark (GTSRB) [2]. We demonstrate that the pipeline is well-suited for generating representative images of unseen traffic sign classes. A baseline image classification setup trained on real data shows an overall performance similar to being trained with a comparable number of artificial data samples. Our code is made publicly available under an open source license.¹

I. INTRODUCTION

Video-based perception and interpretation of traffic signs in a close surrounding are important capabilities for both advanced driver assistance systems (ADAS) and autonomous vehicles. While ADAS may assist in inattentive moments or confusing situations, autonomous vehicles can additionally make use of traffic signs as landmarks and match them with map information to improve localization.

Although they were designed to stand out, the detection and classification of traffic signs are still challenging problems as

- the number of possible classes is extensive,
- the variance in appearance due to weather, background, relative pose to the camera, design choices, and recording conditions is high, and
- the frequency of encountering different classes is extremely unbalanced.

In the past, huge datasets with example images have been compiled [2], [3], [4], [5] to address the main obstacles for traffic sign classification. They are representative in portraying numerous weather, background and recording conditions. Still, in particular the high number of classes and the often nationally decided choices for icon designs or color schemes pose problems which cannot be solved by data collection alone.

* The authors are with the Institute for Neural Computation, Ruhr University Bochum, Universitaetsstrasse 150, 44780 Bochum, Germany firstname.lastname@ini.rub.de

¹github.com/Sirius291/TrafficSignSubstitution

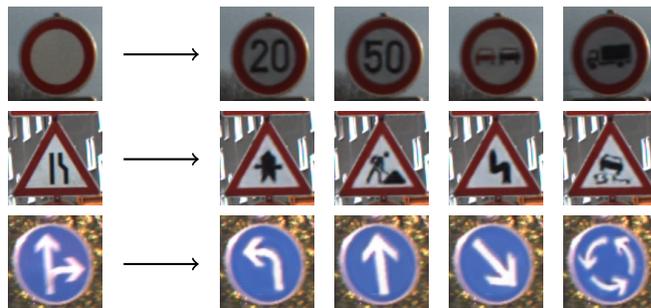


Fig. 1. Overview of contribution. Natural traffic sign images (left column) taken from the GTSRB dataset serve as a basis for the automatic generation of various other realistic traffic sign classes (right columns). Our approach is able to extract information on traffic sign pose, illumination, and motion blur from the original image and use them to generate more natural looking samples of new traffic signs.

This paper extends our work from 2019 [1], in which we proposed a style transfer technique that translated between icon-like depictions of arbitrary traffic signs (then named *prototype images*, now renamed to *cartoon images* for ease of association) and life-like images of traffic signs. To this end we developed a deep neural network architecture based on the CycleGAN training paradigm [6]. Results showed that training an out-of-the-box classifier with a comparable number of generated traffic sign images instead of their real counterparts would decrease the classification rate by 5 to 20 percentage points.

After careful inspection, we attribute the suboptimal performance of the approach to the low-quality backgrounds in the generated images. Fig. 2 illustrates the nature of the problem. It exemplifies the change in background structure when the background color in the corresponding cartoon image is changed. At the same time one can clearly perceive the variance in image artifacts such as illumination and motion blur that the transfer imposes on the life-like images. The generated backgrounds lack consistency for larger structures like buildings and seem to prefer vegetation as it is more straightforward to generate. Further investigation shows that mountings, like the pole to which the sign is attached, are clearly useful for a traffic sign classifier but are oftentimes neglected or even omitted by our previously proposed generation process.

This paper addresses these shortcomings and features an extended generation pipeline that aims to replace traffic signs in real images by ones that were created artificially. That is, in this line of work we exchange suboptimal parts of the generation process by traditional rendering approaches but

maintain realistic image artifacts (e.g., illumination, motion blur) that are already captured well in our earlier approach.

This line of work makes use of the previously proposed CycleGAN for both unsupervised pose estimation and background segmentation as well as for style transfer in the ultimately generated life-like images. In detail, our contributions are:

- a fully unsupervised pose estimation scheme for traffic sign images
- a fully unsupervised background segmentation scheme for traffic sign images
- an image generation pipeline substituting traffic signs in real traffic sign images
- an extensive evaluation of our approach and a full comparison to our earlier work.

II. RELATED WORK

The basis of our previous approach and the core of the data augmentation pipeline presented in this work is a CycleGAN model [6] that trains a bidirectional style transfer (cf. [7] for a recent review of the topic) between two image domains without the need for matching inter-domain image pairs. Two adversarial losses [8] impose a resemblance of the transferred results with the respective target domains. Furthermore, a cycle-consistency loss is introduced that enforces the two transfer mappings to be inverse mappings of one another.

Traffic sign substitution can be likened to other keypoint guided image-to-image translation techniques such as swapping faces [9], facial expressions [10], [11], hand gestures [12], or entire persons [13], [14], [15]. These methods heavily rely on pre-annotated keypoints, such as facial landmarks or body joints, for training. Decreasing this effort is therefore extensively covered in research. We cite the work by Reed et al. [16] aiming to identify points of interest by evaluating verbal image descriptions and let a user manually define the arrangement of these landmarks.

Our method avoids the need for additional labeling, which would be infeasible for a data augmentation technique. Instead, due to the rather simple geometry in traffic sign images, we propose a hand-designed extraction of pose and background segmentation from the cartoon representation.

Traffic sign image generation is also covered in the work of Luo et al. [17], in which the authors paste a traffic sign icon into a random image section of a traffic scene and likewise conduct a style transfer by a model trained with an adversarial loss. While their approach allows for high variance in background appearance, there are several aspects that lead to mediocre results.

Their generative network requires the input of affine transformation parameters, allowing for unnatural poses given a certain location in the background image. Illumination information is calculated from the given background and imposed on the foreground, hereby ignoring natural differences in exposure of fore- and background. The authors add Gaussian blur with random kernel size for more variation at the cost of a plausible focal length and a clear distinction of fore- and

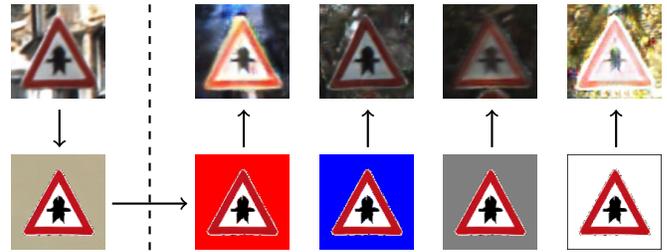


Fig. 2. Some examples that were generated by our previous style transfer approach [1]. Corresponding life-like and cartoon images are aligned column-wisely: The original top left image was transferred to its bottom left cartoon counterpart. Then, the background color of the cartoon image was altered for the generation of life-like images with different backgrounds. The resulting images exhibit variance in recording artifacts like illumination and motion blur but provide no realistic background.

background with regard to blurriness. Furthermore, inconsistencies which we identified in our own earlier approach (cf. Sec. I) and which we are dealing with in this paper, e.g., missing mountings, are not taken into account.

Our approach is to provide highly realistic images for training and only use backgrounds and surroundings that have been recorded before and are therefore less arbitrary. We automatically extract necessary information, i.e., a valid pose estimation for the substitute, as well as illumination, and motion blur information, from the original image foreground, in order to generate consistent samples in all three respects.

III. METHOD

We begin with a brief recapitulation of the most important aspects of our previous approach [1] in Sec. III-A and map out the extension in detail in Sec. III-B. The entire procedure is based on some pre-annotated basis of natural traffic sign images for both the training of the CycleGAN transfer model and as a source for the exchange of traffic sign classes. Please refer to Fig. 3 to follow the explanations in this paragraph. The approach gives rise to an unsupervised pose estimation and background segmentation of traffic signs in Sec. III-C. We use the GTSRB to this end, but similar datasets would be feasible as well. Finally, Sec. III-D states current restrictions to the approach.

A. Traffic Sign Generation

The CycleGAN-based architecture from our previous work [1] trains a bidirectional mapping between a set of cartoon images and a set of real traffic sign images. The cartoon images consist of an icon-like depiction of a traffic sign in front of a homogeneously but arbitrarily colored background (cf. Fig. 2, bottom row). Due to the cycle-consistency enforced during training (concatenating both mappings should result in the original input), the style transfer will respect image structures as closely as possible. In fact, since the cartoon image background is chosen randomly and does not correlate to any traffic sign class, it is used to encode real background texture but also recording artifacts like illumination and motion blur. One should point out that the CycleGAN tends to create small, barely visible variations in

the background color to aim for cycle-consistency but also encode more spatial information.

B. Overview of New Generation Pipeline

The proposed generation pipeline in this paper can be divided into two main parts: extraction and composition. During extraction, all necessary information is taken directly or indirectly from a given real traffic sign image. The substitute traffic sign is created afterwards in the composition part of the algorithm. Fig. 3 depicts the different intermediate results of both directions.

Our pipeline heavily relies on the simplified representation of the cartoon domain. Starting with a real-world sample from the GTSRB, we first deploy the CycleGAN (cf. Sec. III-A) to transfer it into its cartoon pendant. The abstract nature of the cartoon space facilitates the extraction of the binary background segmentation and consequently the calculation of the traffic sign pose (cf. Sec. III-C).

A substitute is generated from previously gathered information and a simple icon of the target traffic sign class. We apply the inverse pose to the icon of the traffic sign class we aim to embed in order to receive a tilted version. Replacing the background with the one from the earlier generated cartoon representation yields a new cartoon image with the same pose and encoded background but a different class. The new cartoon is transferred into the life-like domain by use of the CycleGAN. Once again the segmentation mask helps to combine the newly generated traffic sign with the original real background. Borders are crossfaded during this process to avoid artifacts. The resulting composition is the substituted traffic sign image.

C. Binary Segmentation and Pose Estimation

The background segmentation on the generated cartoon is performed by estimating the mean of the background color distribution via sampling pixels in vicinity to the image border. Computing and thresholding the pixel-wise Euclidean distance from the mean provides a simple way to discriminate pixels of the background and the traffic sign.

Traffic sign pose estimation on the cartoon image is slightly more involved as a certain degree of deformation on the generated traffic sign representation is common during style transfer. We found that the following robust feature matching scheme based on ORB features [18] yields fast and reliable results: We estimate the homography between the cartoon and a straight icon of the same traffic sign class assuming little to no rotation around the traffic sign plane normal. That is, two ORB keypoints from the two images are only matched if no significant rotation between them is present. We then deploy a RANSAC approach in which only those random samples of keypoints are considered that lead to a homography which maps the cartoon’s background mask to a mask extracted from the icon image by separating transparent background pixels from the non-transparent foreground. The mapping is achieved by computing the Jaccard coefficient for each sampled homography and choosing the one with maximum value. In order to generate more varied

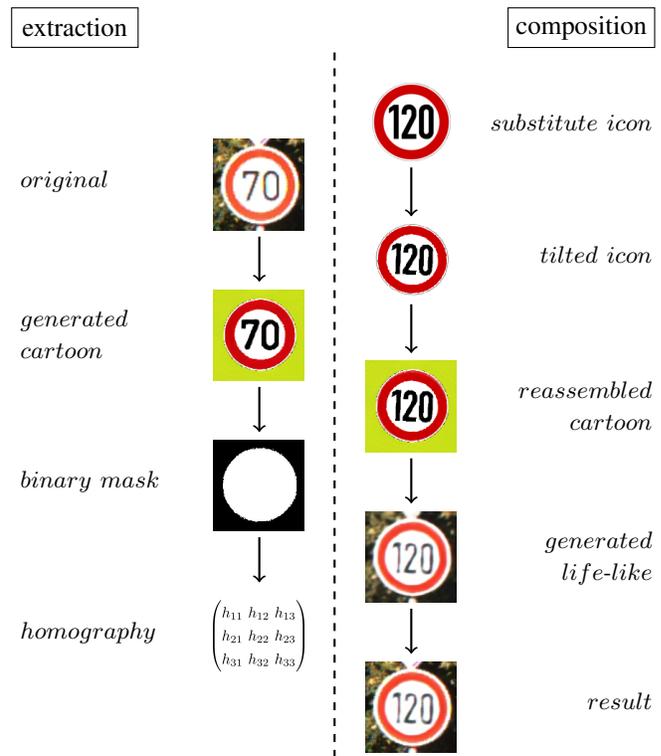


Fig. 3. Scheme of the generation process. The pipeline can be split in two. The extraction part collects all necessary information from the original image, while the composition part generates a realistic substitute image with the help of the previously gathered information. Please note that the generated traffic sign image has adapted the recording characteristics from its source image: In this example it exhibits a slight overexposure.

traffic sign substitutes, for round traffic signs a random roll rotation of up to 5° is applied to the resulting homography.

D. Restrictions

Due to the fixed-size input and output structure of the CycleGAN model, it has only been trained with images of 128×128 pixels resolution. Larger images were downsampled accordingly during training, smaller images were discarded. In order to also substitute classes in low-resolution images, we upscale them before the CycleGAN transfer. In doing so we are able to significantly extend the number of possible surroundings in which classes may be substituted.

Although the CycleGAN itself is capable of synthesizing all visually similar traffic sign classes with which it was trained, substituting traffic signs in images with our new approach is currently only possible for traffic signs of the same category. For an overview of the traffic sign categories in the GTSRB, we refer to Tab. I.

While the most important factor for a successful substitution is a shared outer shape, we also found that mixing different categories of round signs (i.e., restriction, derestriction, and direction signs) has its challenges. Apart from the fact that most direction signs appear in much lower heights than all other traffic signs and therefore show backgrounds different to those of other categories, the overall color scheme of an image is used by the CycleGAN to extract illumination

TABLE I

CATEGORIZATION OF TRAFFIC SIGNS GIVEN IN THE GTSRB DATASET. EACH OF THE 43 TRAFFIC SIGN CLASSES IS SORTED INTO ONE OF THE FIVE CATEGORIES BASED ON THEIR SHARED VISUAL FEATURES. (FIGURE AS SEEN IN [1])

Category	Characterization	Examples
Warning Signs	upright triangular shape, red border, white background, black content	
Restriction Signs	circular shape, red border, white background, mostly black content	
Derestriction Signs	circular shape, white background, diagonal bars, gray content	
Direction Signs	circular shape, blue background, white arrows	
Miscellaneous Signs	no common features	

information. As the major part of the image samples is taken by the respective traffic sign, its ground color influences the substitution process. A rather dark-colored direction sign of a certain illumination will result in a far darker derestriction sign substitute which no longer fits the chosen background and is thus a potentially problematic image w.r.t. further usage. As a matter of fact, images with unique features or geometry (*Yield Way*, *STOP*, *Priority Street*, *One-Way Street*) can only be substituted within the same class. This is a restriction of our approach, however, we want to point out that the majority of signs under the *Vienna Convention on Road Traffic* share a common geometry and only differ in their choice of pictograms and designs like font types and thickness of lines and borders.

IV. EXPERIMENTS

We see our method in two fields of application: As a technique, first, to extend the number of training examples for underrepresented traffic sign classes, and, second, to allow for training classes that are not given in the dataset at all. In order to objectively assess the aptitude of the generated images for training, we use a multi-class SVM classifier [19] on HOG features [20]. This line of experiments is also in accordance with our previous work [1]. Although state-of-the-art approaches regularly deploy Convolutional Neural Networks for multi-class image classification, these methods might adapt the feature extraction to artifacts of our data generation pipeline and, hence, skew the results. We have shown the general success of our method for both SVMs and CNNs in our previous paper [1] and thus refer to said publication for more details regarding the evaluation with CNNs.

As a baseline experiment we trained an SVM on the GTSRB, i.e., entirely on real-world images, and refer to it as SVM_{Base} . A number of experiments with full or partial use of our generated training samples were conducted in order to assess the quality of our approach. The hyperparameters for all SVMs were chosen to be $C = 100$ and $\gamma = 0.1$, without further cross validation on each experiment as we opted for a proper comparison to previous results with the

TABLE II
COMPOSITION DETAILS OF TRAINING DATASETS FOR THE VARIOUS SVM CLASSIFIERS USED IN THE EXPERIMENTS

Key	Training Dataset Composition
SVM_{Base}	real images, GTSRB training set [2], serves as baseline <i>high variation in class size, avg. 449 samples per class</i>
SVM_{Prev}	generated images as in [1] <i>equal distribution of samples (449 per class)</i>
SVM_{Gen}	generated images (ours) <i>equal distribution of samples (449 per class)</i>
SVM_{Imb}	generated images (ours) <i>unbalanced class distribution equal to SVM_{Base}</i>
SVM_{5000}	generated images (ours) <i>equal distribution of samples (5000 per class)</i>
SVM_{Lev}	mixture of real and generated images (ours) <i>SVM_{Base} dataset plus augmentation of underrepresented classes with generated samples for leveled classes; equal distribution of samples (1110 per class)</i>

same hyperparameters rather than peak performance. The experiments differ only in the training datasets. An overview of the variations in composition and number of training samples is given in Tab. II.

All classifiers were tested on real-world samples of the GTSRB dataset, which is provided with a predetermined split into training and test set. In accordance with [1] we further subdivided the training set into two halves for training the CycleGAN model and for conducting the following experiments. Likewise the test set was split into two halves for finetuning the generation pipeline and testing the trained SVM classifiers.

A. Training on Generated Images for All Classes

In order to evaluate the overall quality of our generated images, we train SVM_{Gen} , SVM_{Imb} , and SVM_{5000} on a purely synthetic traffic sign dataset featuring all 43 GTSRB sign classes. SVM_{Gen} uses 449 generated samples per class as a direct comparison to SVM_{Prev} , our previous approach.

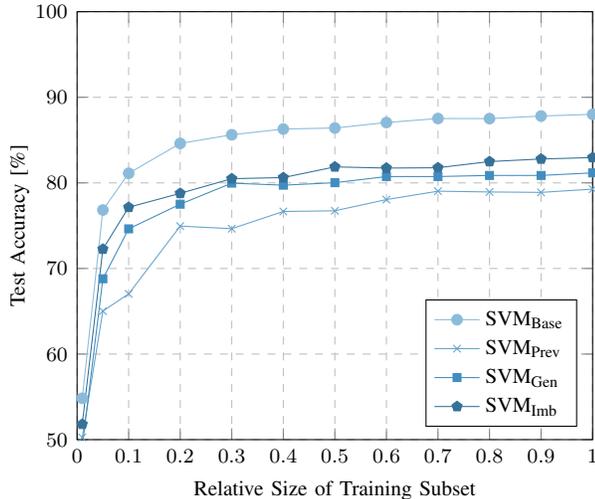


Fig. 4. Performance of different training datasets on the GTSRB test set with reduced sample sizes. Partitions were generated by random choice of samples from the respective original dataset.

SVM_{Imb} was trained on a highly unbalanced dataset of generated samples. The number of samples per class corresponds to the distribution of the original GTSRB dataset ranging from 90 to 1110 samples. This experiment thus is a direct comparison to SVM_{Base}. The fact that we are able to generate great amounts of data was used in SVM₅₀₀₀, in which we generated a dataset with 5000 samples for each class. Finally, SVM_{Lev} is the only experiment in this set, which combines real-world and artificial samples. In order to show that our generated images can improve existing datasets by complementing underrepresented classes, we have taken the complete training dataset from the baseline classifier SVM_{Base} and evened out the number of samples per traffic sign class to fit the most prominent one with 1110 samples. The results are juxtaposed with SVM_{Base} and SVM_{Prev} in Tab. III.

For a better understanding on how the number of training samples influences our results for the different approaches, we have conducted a number of experiments with subsets of the SVM_{Base}, SVM_{Prev}, SVM_{Gen}, and SVM_{Imb} datasets. Fig. 4 depicts the results.

B. Training on One Unseen Generated Image Class

To gauge the aptitude of our technique for generating images from new and fully unseen traffic sign classes, we conducted three further experiment series in which either of the classes *No Entry (Trucks)*, *Slippery Road*, and *Pass Right* is removed from the training datasets of the CycleGAN.

Subsequently, a set of generated images of the previously removed class was added to the different dataset compositions in order to mimic generating a completely unknown traffic sign class. Besides our baseline classifier SVM_{Base} and SVM_{Prev}, resulting from a former publication, we have conducted these experiments with the dataset compositions from SVM_{Gen}, SVM_{Imb}, and SVM₅₀₀₀, i.e., the number of added samples for only the replaced traffic sign class varies

TABLE III
CATEGORY-WISE CLASSIFICATION ACCURACY FOR DIFFERENT TRAINING DATASETS. REFER TO TABLE II FOR DATASET COMPOSITION DETAILS. ALL CLASSIFIERS WERE TESTED ON THE SAME PARTITION OF THE GTSRB DATASET. BEST RESULTS ARE HIGHLIGHTED.

Classification Accuracy (%)						
Category	SVM _{Base}	SVM _{Prev}	SVM _{Gen}	SVM _{Imb}	SVM ₅₀₀₀	SVM _{Lev}
Warning	78.16	75.76	74.75	76.20	79.54	83.16
Restriction	87.44	72.21	75.23	78.53	80.07	88.35
Derestriction	80.87	86.34	84.15	83.06	90.71	86.34
Direction	94.37	85.86	89.77	90.46	91.03	94.37
Miscellaneous	98.65	96.62	98.36	97.87	97.97	99.13
Total	88.01	79.27	81.17	82.96	84.70	89.75

TABLE IV
CLASS-WISE CLASSIFICATION ACCURACY FOR SUBSTITUTION OF TRAFFIC SIGN “No Entry (Trucks)” (1st row). UNMENTIONED TRAFFIC SIGN CLASSES SHOW NO ALTERATION. HIGHLIGHTED ENTRIES DIFFER FROM THE BASELINE CLASSIFIER. FOR SVM_{BASE} AND SVM_{IMB} THE NUMBER OF SAMPLES FOR CLASS “No Entry (Trucks)” IS 210, FOR SVM_{PREV} AND SVM_{GEN} 449, AND FOR SVM₅₀₀₀ 5000.

Class	Replacement of Class “No Entry (Trucks)” Classification Accuracy (%)				
	SVM _{Base}	SVM _{Prev}	SVM _{Gen}	SVM _{Imb}	SVM ₅₀₀₀
No Entry (Trucks)	97.18	88.73	100.00	100.00	100.00
Speed Limit 20	72.73	72.73	72.73	78.79	72.73
Speed Limit 30	86.97	86.70	86.97	84.57	86.97
Speed Limit 50	89.45	89.45	89.45	88.65	89.45
Speed Limit 60	81.66	81.66	81.66	80.79	81.66
Speed Limit 70	94.48	94.48	94.48	93.90	94.48
Speed Limit 80	77.29	77.29	77.29	76.03	77.60
Derestrict 80	87.01	85.71	87.01	87.01	87.01
Speed Limit 100	74.44	73.99	74.44	74.44	74.44
Speed Limit 120	76.04	76.04	76.04	73.73	76.50
Proh. Overtaking	99.58	99.58	96.67	98.75	93.33
Proh. Overtaking (Trk)	96.82	96.82	96.82	97.13	97.13
Right of Way	78.89	78.89	78.89	79.90	78.89
One-Way Street	98.86	98.86	98.86	98.86	97.71
Danger	86.55	86.55	86.55	85.96	86.55
Att. S Curve	56.60	54.72	56.60	54.72	56.60
Att. Slippery Road	67.11	67.11	67.11	65.79	67.11
Att. Road Will Narrow	78.05	78.05	78.05	75.61	78.05
Att. Construction Site	87.70	87.70	87.70	88.52	87.70
Att. Traffic Lights	74.23	74.23	74.23	75.26	74.23
Att. Pedestrians	28.12	28.12	28.12	25.00	28.12
Att. Playing Children	80.00	80.00	80.00	81.25	80.00
Att. Bicycle	85.11	85.11	85.11	87.23	85.11
Att. Snowfall	52.86	52.86	52.86	51.43	52.86
Roundabout	70.45	72.73	70.45	70.45	72.73
Derestrict Overtaking	84.85	84.85	69.70	72.73	63.64
Total	88.01	87.87	87.85	87.62	87.73

TABLE V

CLASS-WISE CLASSIFICATION ACCURACY FOR SUBSTITUTION OF TRAFFIC SIGN “ATTENTION SLIPPERY ROAD” (1st ROW).

UNMENTIONED TRAFFIC SIGN CLASSES SHOW NO ALTERATION.

HIGHLIGHTED ENTRIES DIFFER FROM THE BASELINE CLASSIFIER. FOR SVM_{BASE} AND SVM_{IMB} THE NUMBER OF SAMPLES FOR CLASS “NO ENTRY (TRUCKS)” IS 240, FOR SVM_{PREV} AND SVM_{GEN} 449, AND FOR SVM₅₀₀₀ 5000.

	Replacement of Class “Attention Slippery Road”				
	Classification Accuracy (%)				
Class	SVM _{Base}	SVM _{Prev}	SVM _{Gen}	SVM _{Imb}	SVM ₅₀₀₀
Att. Slippery Road	67.11	60.53	94.74	94.74	100.00
Speed Limit 30	86.97	86.70	86.97	86.97	86.97
Derestrict 80	87.01	85.71	87.01	87.01	87.01
Right of Way	78.89	78.89	78.89	78.89	78.39
Att. Rd Curves Left	37.93	37.93	41.38	41.38	37.93
Att. Rd Curves Right	81.58	81.58	78.95	78.95	76.32
Att. S Curve	56.60	54.72	56.60	56.60	56.60
Att. Construction Site	87.70	87.70	88.11	88.11	87.70
Att. Traffic Lights	74.23	74.23	74.23	74.23	73.20
Att. Bicycle	85.11	85.11	85.11	85.11	87.23
Att. Snowfall	52.86	52.86	52.86	52.86	51.43
Att. Deer Crossing	98.56	98.56	97.84	97.84	93.53
Total	88.01	87.89	88.35	88.35	88.23

with each dataset version according to the chosen guideline while images for all other classes are taken from the GTSRB dataset. Refer to Tab. II for details on the composition of training datasets.

The test dataset is the same as described in Sec. IV-A, consisting entirely of natural image data from the GTSRB. Test performance for the three generated classes is given in Tab. IV, V, and VI, respectively.

V. RESULTS

In Tab. III we compare the performance of a general classification approach on one of six traffic sign datasets featuring real data, data generated by our previous method [1], and data generated by the newly proposed pipeline in four different compositions. While all our fully artificial datasets reveal a minor drop in performance of less than 7 percentage points compared to the real counterpart, all datasets created by our newly proposed algorithm outperform SVM_{PREV}.

A category-wise comparison shows that all generative approaches surpass the real dataset in underrepresented categories such as *Derestriction*, scaling with the amount of samples given by the respective dataset composition. However, they fall short in other categories that contain most of the remaining classes and samples. While SVM_{IMB} shows a performance drop in all categories w.r.t. to SVM_{BASE}, which uses the same amount of training samples per class,

TABLE VI

CLASS-WISE CLASSIFICATION ACCURACY FOR SUBSTITUTION OF TRAFFIC SIGN “PASS RIGHT” (1st ROW). UNMENTIONED TRAFFIC SIGN CLASSES SHOW NO ALTERATION. HIGHLIGHTED ENTRIES DIFFER FROM THE BASELINE CLASSIFIER. FOR SVM_{BASE} AND SVM_{IMB} THE NUMBER OF SAMPLES FOR CLASS “NO ENTRY (TRUCKS)” IS 1020, FOR SVM_{PREV} AND SVM_{GEN} 449, AND FOR SVM₅₀₀₀ 5000.

	Replacement of Class “Pass Right”				
	Classification Accuracy (%)				
Class	SVM _{Base}	SVM _{Prev}	SVM _{Gen}	SVM _{Imb}	SVM ₅₀₀₀
Pass Right	95.87	78.47	88.50	88.50	93.22
Speed Limit 30	86.97	86.70	86.97	86.97	86.97
Derestrict 80	87.01	85.71	87.01	87.01	87.01
Stop	92.36	93.06	93.06	93.06	93.06
Att. S Curve	56.60	54.72	56.60	56.60	56.60
Turn Left	100.00	100.00	100.00	100.00	98.31
Forward or Right	96.92	98.46	98.46	98.46	98.46
Total	88.01	87.06	87.65	87.65	87.89

SVM_{IMB} still clearly outperforms SVM_{PREV} and SVM_{GEN} in strong categories, e.g., *Restriction*. This is due to the fact, that SVM_{PREV} and SVM_{GEN} use a fix number of 449 samples per class, while SVM_{IMB} copies the class distribution of SVM_{BASE}, resulting in an average of 712.5 samples per class for the category *Restriction*. The importance of overall class sizes can also be seen in Fig. 4, in which all datasets show similar performance for the same amount of size reduction, regardless of the dataset composition.

All fully generated datasets seemingly carry over some of the variance in image composition from which small categories can profit, but fall short in achieving en-par figures for the main categories. Our only dataset consisting of both real and synthetic images, however, matches or outperforms even our baseline classifier. This emphasizes the fact that our approach can improve existing datasets by augmenting underrepresented categories.

Replacing a single traffic sign class within a dataset consisting of otherwise natural images is covered in Tab. IV, V, and VI. Again, performance on the original GTSRB dataset is juxtaposed to our previous and new data generation approaches. The classifiers exhibit similar overall performance in all five experiments. Our new approaches outperform the previous one for each artificial class, even clearly outperforming the use of real data samples for the classes *No Entry (Trucks)* and *Slippery Road*. Apart from minor deviations, the discrepancy in performance of the underrepresented classes, e.g., *Derestrict Overtaking* when replacing *No Entry (Trucks)* is prominent. This effect scales with the number of generated samples. We attribute this to the skewed distribution after adding a significant number of images to the dataset replacing smaller classes with 210 (*No Entry (Trucks)*) and 240 (*Slippery Road*) samples.



Fig. 5. The depicted traffic signs show the ability of our method to generate realistic looking images from arbitrary (including fictitious) traffic sign icons. Resulting images show adaptation in illumination, pose and motion blur as would be expected from natural traffic sign images.

On the contrary, replacing one of the most frequent classes, *Pass Right*, with originally 1020 training images and effectively cutting the number of samples down to less than one half for SVM_{Prev} and SVM_{Gen} as well as scaling them to 5000, and thus even extending the overall status as highly frequent class, does not show this effect and only leads to minor deviations in the performance of all other classes.

VI. CONCLUSIONS

In this paper we have presented a fully automated image substitution technique for traffic signs. It is able to generate images of traffic signs with arbitrary pictograms (cf. Fig. 5) that exhibit the variance of real-life backgrounds and recording artifacts. In doing so, this case study can represent a number of recognition problems found in natural image data with simple geometric constraints but highly varying appearance due to recording conditions, design, or natural occurrences.

For future work we will look into some of the remaining shortcomings of the newly proposed generation process: Images of low quality or minor size (smallest image samples for substitution were only 25×25 pixels of size), lead to poorly generated cartoons by the CycleGAN. As a consequence, also the pose estimation and background segmentation might be deficient and the overall generation procedure cannot lead to satisfying results anymore. Opening the currently possible substitutions to inter-class substitutions might lead to more variance in background / traffic sign combinations and ultimately produce a more stable training dataset. If and how these caveats and restrictions can be remedied has to be examined closely.

REFERENCES

- [1] D. Spata, D. Horn, and S. Houben, "Generation of Natural Traffic Sign Images Using Domain Translation with Cycle-Consistent Generative Adversarial Networks," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 622–628.
- [2] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German Traffic Sign Recognition Benchmark: A Multi-Class Classification Competition," in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, 2011, pp. 1453–1460.
- [3] F. Larsson and M. Felsberg, "Using Fourier Descriptors and Spatial Models for Traffic Sign Recognition," in *Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, 2011, pp. 238–249.
- [4] 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.
- [5] A. Chigorin and A. Konushin, "A System for Large-Scale Automatic Traffic Sign Recognition and Mapping," *Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, City Models, Roads and Traffic*, vol. II-3/W3, pp. 13–17, 2013.
- [6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.
- [7] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural Style Transfer: A Review," *Transactions on Visualization and Computer Graphics*, 2019.
- [8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 2672–2680.
- [9] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast Face-Swap Using Convolutional Neural Networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3677–3685.
- [10] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry Guided Adversarial Facial Expression Synthesis," in *Proceedings of the ACM International Conference on Multimedia*, 2018, pp. 627–635.
- [11] W. Wang, X. Alameda-Pineda, D. Xu, P. Fua, E. Ricci, and N. Sebe, "Every Smile is Unique: Landmark-Guided Diverse Smile Generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7083–7092.
- [12] H. Tang, W. Wang, D. Xu, Y. Yan, and N. Sebe, "GestureGAN for Hand Gesture-to-Gesture Translation in the Wild," in *Proceedings of the ACM International Conference on Multimedia*, 2018, pp. 774–782.
- [13] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled Person Image Generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 99–108.
- [14] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose Guided Person Image Generation," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 406–416.
- [15] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable GANs for Pose-Based Human Image Generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3408–3416.
- [16] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning What and Where to Draw," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 217–225.
- [17] H. Luo, Q. Kong, and F. Wu, "Traffic Sign Image Synthesis with Generative Adversarial Networks," in *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, 2018, pp. 2540–2545.
- [18] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2564–2571.
- [19] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, p. 273–297, 1995.
- [20] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.