Context-aware Training Image Synthesis for Traffic Sign Recognition

Akira Sekizawa and Katsuto Nakajima
Department of Information Systems and Multimedia Design, Tokyo Denki University, Tokyo, Japan

Keywords: Traffic Sign Recognition, Object Detection, Synthetic Data, Data Augmentation.

Abstract: In this paper, we propose a method for training traffic sign detectors without using actual images of the traffic signs. The method involves using training images of road scenes that were synthetically generated to train a deep-learning based end-to-end traffic sign detector (which includes detection and classification). Conventional methods for generating training data mostly focus only on producing small images of the traffic sign alone and cannot be used for generating images for training end-to-end traffic sign detectors, which use images of the overall scenes as the training data. In this paper, we propose a method for synthetically generating road scenes to use as the training data for end-to-end traffic sign detectors. We also show that considering the context information of the surroundings of the traffic signs when generating scenes is effective for improving the precision.

1 INTRODUCTION

To implement advanced driver assistance systems and achieve fully autonomous driving, researchers are actively working on object recognition technology for recognizing humans and the objects surrounding vehicles. Sensors that are used for object recognition include RGB cameras, millimeter-wave radar, and LiDAR (laser radar). For example, LiDAR is effective for detecting humans with high accuracy. In contrast, RGB cameras are necessary for recognizing traffic signs because the system must be able to read regulations printed on the surfaces of the signs. In the field of image recognition by RGB cameras, end-to-end object recognition methods based on deep learning approaches have made it possible to recognize objects with high precision and speed. However, it is necessary to collect large amounts of diverse training data when using end-to-end object recognition methods compared with using conventional methods.

Methods for collecting training data can be classified into the following three categories: methods that use publicly available datasets, methods that involve manually collecting new data, and methods that involve synthesizing new data. For traffic sign recognition, it is possible to use published datasets in only a few situations. Because traffic sign standards are different in each country, traffic sign datasets from one country cannot be used for training traffic sign recognition systems in another country. Furthermore, manual collection of traffic sign images is an extremely time-consuming process. For example, even if the consideration is limited to only a set of regulatory signs from all classes of traffic signs in Japan, there are still over 60 classes of signs. Furthermore, because there is bias in the locations where each class of sign is installed, it would be extremely expensive to travel to the locations where each traffic sign is installed to photograph them. Alternatively, it is possible to collect images of the signs from the internet instead of photographing them on-site. However, it is difficult to use the internet to collect traffic sign images that are not installed in many places.

Researchers are studying methods for synthesizing data as a third method of collecting training data. Synthetic generation of training data for traffic sign recognition is advantageous because it reduces the cost of collecting data and makes it possible to generate traffic sign datasets for any given country. A method for generating images of single traffic signs was first proposed by Ishida et al. in 2006 (Ishida et al., 2006). Their method used with realistic degradation involving application of various degradation models such as blur and rotation to the template images of the traffic signs. Subsequently, several other methods have also been proposed (Hoessler et al., 2007; Medici et al., 2008; Møgelmose et al., 2012; Moiseev et al., 2013; Haselhoff et al., 2017). In contrast, research on methods for generating images of the overall scenes that include traffic signs could not be found. In several previous research projects, researchers performed data augmentation to increase

Sekizawa, A. and Nakajima, K.
Context-aware Training Image Synthesis for Traffic Sign Recognition.
DOI: 10.5220/0007387904660473
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 466-473
ISBN: 978-989-758-354-4

the number of images in the datasets by pasting synthesized images of the traffic signs on the existing background images (Zhu et al., 2016; Peng et al., 2017; Uršič et al., 2017). These previous research projects did not evaluate the precision of the traffic sign recognition model trained by only the synthetically generated traffic sign images. They also did not investigate methods for improving the traffic sign image synthesis method itself.

In this paper, we make two contributions. Our first contribution is to propose a novel method for synthetically generating the training data for an end-to-end traffic sign detector. Our proposed method synthesizes the traffic signs at the locations where they are supposed to be when pasting them on the background images by focusing on the context of the surroundings of an object, that end-to-end object detectors would be expected to use. An example of an image of a road scene generated using the proposed method is shown in Figure 1. The proposed method uses only published traffic sign datasets and traffic sign pictograms to generate the training data. Therefore, it is unnecessary to manually collect new data. Our second contribution is the evaluation of the precision of the models that are trained using only the synthesized data. Although there have been some reports on the precision of the systems trained on a mix of synthetically generated and manually collected road scenes do exist, studies regarding the precision when the systems are trained on synthesized data alone are not known.

Figure 1: Road scene generated using the proposed method.

2 RELATED WORKS

2.1 Synthetic Image Generation of Single Traffic Signs

Images of single traffic signs are used as the training data for pure traffic sign classifiers. The main method for synthetically generating images of single traffic signs is a degradation model. In this method, degraded images are generated by applying functions (degradation models) that express the degradation such as a rotation, a blur, or a translation to an ideal template image of a traffic sign. A diverse set of synthetic images is generated by varying the parameters of the degradation model. Generation of training data using degradation models was first proposed by Baird for application to the task of optical character recognition (OCR) for optically scanned documents (Baird, 1992). Ishida et al. proposed a degradation model for traffic sign recognition with three parameters, namely, rotation, blur, and translation (Ishida et al., 2006). Ishida et al. also proposed a method for adaptively determining the values of the degradation parameters using a genetic algorithm (Ishida et al., 2007). Which were manually determined based on experience. Møgelmose et al. suggested a degradation model for traffic sign recognition that included six parameters, namely, hue change, luminosity change, rotation, Gaussian blur, Gaussian noise, and occlusion (Møgelmose et al., 2012). Moiseev et al. proposed a degradation model that included hue change, saturation change, three-dimensional rotation, projective transformation, Gaussian blur, translation, scaling, and Gaussian noise (Moiseev et al., 2013). Classifiers that were trained using the degradation model proposed by Moiseev et al. achieved precision that were higher than those of the classifiers trained using real images in the traffic sign recognition benchmark for German traffic signs (GTSRB) (Stallkamp et al., 2012).

Another method for generating realistic synthetic images transfers the degradation of an actual image to a synthesized image. Haselhoff et al. proposed a method for generating synthetic traffic signs by transferring the features of the appearance of the real traffic signs to other synthetic traffic signs using a Markov random field (Haselhoff et al., 2017). In Haselhoff's method, the traffic sign classes of the transfer source can be different from the classes of the destination. Therefore, it is possible to use this method to synthetically generate images of the traffic signs from one country to another country.

2.2 Synthetic Image Generation of Overall Scenes that Include Traffic Signs

Images of the road scenes that include traffic signs are used as the training data for the end-to-end object detectors. To the best of our knowledge, research that directly investigates the synthetic generation of road

scenes for traffic sign detection is not yet reported. There are several instances of previous work in which the researchers have generated road scenes that include traffic signs as a part of a data augmentation process.

Zhu et al. generated road scenes by pasting synthetic traffic signs that were generated using a degradation model on background images at random locations (Zhu et al., 2016). Zhu's degradation model included rotation within the range from -20° to 20°, scaling within the range of 20 pixels to 200 pixels, projective transformation within the range that was appropriate for traffic signs, and random noise. Zhu et al. combined the synthetically generated scene images with manually collected scene images and used them for training.

Peng et al. generated scene images by extracting the images of single traffic signs from existing traffic sign datasets and pasting them on background images at random locations (Peng et al., 2017). In Peng's method, actual images were used for both the traffic sign images and background images. Peng et al. extracted the images of single traffic signs from the GTSRB dataset (Stallkamp et al., 2012) and used the KITTI Road/Lane Detection Evaluation 2013 dataset (Fritsch et al., 2013) for the background images. Note that the researchers did not apply any additional degradation or deformation to the traffic sign images when pasting them on background images.

Uršic et al. pasted both the synthetically generated traffic signs and actual traffic signs on background images at random locations (Uršič et al., 2017). Uršic et al. applied deformations due to affine transformations and scaling transformations to both the synthetically generated images and actual images of the traffic signs, changed their luminosity and contrast, and applied motion blur and shadow. In addition, Uršic et al. proposed to avoid pasting traffic sign images in the center of the background image. The center of the background image is normally occupied by the road, and it is very rare for a traffic sign to be located in this part of the image.

The goal of these research projects was data augmentation. The researchers trained their models by using both collected scene images and synthetically generated scene images simultaneously. There have been no reports of studies in which training was performed using synthetically generated scene images alone. Furthermore, these studies did not focus on improving the methods for synthetically generating scene images either.

3 PROPOSED METHOD

In this section, we describe a method for synthetically generating images of scenes that include traffic signs to serve as the training data for end-to-end traffic sign detectors. Our preliminary experiments showed that the precision of the traffic sign detector that were trained using Zhu et al's method, in which synthetically generated traffic signs were pasted at random locations (Zhu et al., 2016) (hereinafter referred to as the random method), had a mAP that was approximately 10% lower than that of the detectors that were trained using actual training images. An example of a road scene that was generated using the random method is shown in Figure 2. The procedure flow is shown in Figure 3.

Figure 2: Example of a road scene generated using the random method.

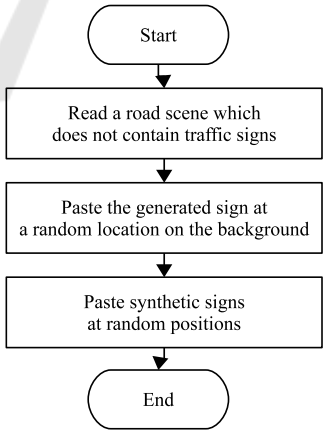

Figure 3: Flowchart of the random method.

Although the random method exhibits a high rendering quality for generating single traffic signs, the locations where the signs are pasted and sizes of the signs are random. Therefore, the traffic signs are pasted at locations where the traffic signs are not sup-

posed to exist, such as floating in the air or on center of the road. Our hypothesis for the reason why images generated using random method yield a low performance is that the context information of the surroundings of the traffic sign is lost in the generated scene images. End-to-end object detectors use the context information from the surroundings of an object to detect and classify objects. Therefore, using scene images in which the context information is destroyed will have a negative effect on the training.

In our proposed method, we generate training data by pasting traffic signs at locations at which the traffic signs were originally located in the scene. An example of a generated scene using our proposed method is displayed in Figure 4. The procedure flow is depicted in Figure 5.

Figure 4: Example of a road scene generated using the proposed method.

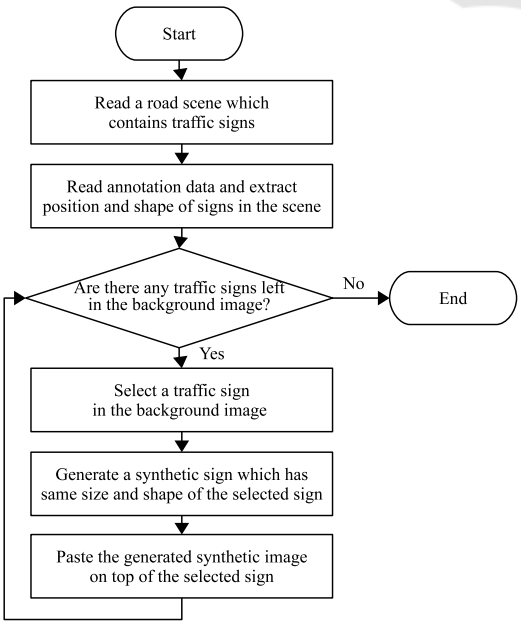

Figure 5: Flowchart of the proposed method.

To obtain the locations at which traffic signs exist within the scenes, we used published traffic sign datasets. Because these datasets included numerous scenes that include traffic signs, we can use these images as the background images and paste synthetically generated traffic signs at overlapping locations where the traffic signs is originally located. This made it possible to synthetically generate scene images that preserved the context information surrounding the traffic signs.

3.1 Collection of Background Images

Our proposed method uses images of scenes that include traffic signs as background images. It is necessary for the locations and shapes of the traffic signs to be annotated earlier. It is possible to use a traffic sign dataset from a country or region that is different from the target country or region. This is because although different countries may drive on different sides of the road, the locations at which the traffic signs are installed and the methods for installing them are practically the same. In addition, because the traffic signs are overwritten by the synthetically generated traffic signs, the method is not affected by the differences in the traffic signs between countries.

3.2 Generation of Single Traffic Signs

When generating images of single traffic signs, it is necessary to consider the following four factors:

- Shape
- Size
- Rotation
- Degradation of appearance

In terms of the shape, it is necessary to choose a traffic sign of a shape that can completely covers the original traffic sign. For example, if a triangular traffic sign is pasted on top of a traffic sign that is originally circular, part of the original traffic sign will still be visible, as shown in Figure 6. This would result in an inappropriate context surrounding the sign.

Therefore, it is desirable for the original sign and pasted sign to have the exact same shape. However, some shapes, such as octagons, do not appear very frequently, and the number of existing scene images that include these signs is small. Therefore, it is desirable to allow exceptions, such as allowing octagonal signs to be pasted on top of circular signs.

In our proposed method, we do not require the classes of the original sign and pasted sign to match. If we were to require the classes to match, then this

Figure 6: Example of an image in which signs of different shapes were pasted on an image.

would increase the constraints, and it would not be possible to generate a sufficient number of diverse scene images. In addition, because there is not much difference between the context surrounding the signs for the different classes of signs, there is not much advantage in requiring the types to match. As for the size, it is necessary to make the width and height of the synthetically generated sign to match those of the bounding rectangle of the original sign obtained based on the annotations.

Regarding the rotation, it is desirable to calculate the projective transformation matrix of the original sign and apply the same projective transformation to the synthetically generated sign. However, because it is difficult to calculate the projective transformation matrix of an original sign from an image, we use the following method. First, because traffic signs are positioned perpendicular to the ground, we assume that there is no rotation in the direction of the x-axis or z-axis. To determine the rotation in the y-axis direction, we assume that it is possible to approximate the deformation due to rotation in the y-axis direction by aligning the widths and heights of the bounding boxes of the original sign and pasted sign. An example of a synthetic sign aligning only the widths and heights is shown in Figure 7. The projective deformation of the traffic signs that are far away and small can be ignored. Therefore, it is possible to sufficiently deal with the signs that have undergone projective transformation simply by aligning the widths and heights of the bounding rectangles, as we will demonstrate in our evaluation later.

We processed the appearance of the signs to make them resemble the actual images by using a degradation model. Because our purpose is to paste the synthetically generated signs within the scene, it is desirable to consider whether the image matches the brightness and sharpness of the background and whether it looks realistic when determining the degradation to apply to the sign image. However, this topic is left for future research. In this paper, random degradation is applied.

(a) Actual sign (b) Synthesized sign

Figure 7: Example of a synthetic image created by pasting on top of a sign that has undergone a projective transformation.

4 EVALUATION

To evaluate the effect of road scene generation by considering the context, we trained an end-to-end object detector, Faster R-CNN, using the following three sets of training data:

1. Actual images
2. Synthesized images (proposed method)
3. Synthesized images (Zhu's random method (Zhu et al., 2016))

In this experiment, we used the Tsinghua–Tencent 100K traffic sign dataset (TT100K) (Zhu et al., 2016). TT100K includes 16,811 road scene images of 2048×2048 pixels with annotations, each captured in China. The dataset includes 6,103 images in the training set; 3,067 images in the test set; and 7,641 other images (of which 6,544 are background images that do not include traffic signs). In this experiment, we use all images in the training set and test set. TT100K includes total 182 classes of traffic signs. In addition, TT100K also includes template images for the 128 classes. In this experiment, we use 79 of classes. A list of the traffic signs used in this experiment is shown in Figure 8. This set of 79 classes is the intersection of the following two sets of classes.

- 151 traffic sign classes included in the training set
- 128 traffic sign classes for which a template image has been provided

4.1 Generation of Training Data

In this experiment, we prepared three types of training data, including actual images (Set 1), synthetic images that were generated using the proposed method

Figure 8: List of the traffic signs that were used in this experiment.

(Set 2), and synthetic images that were generated using Zhu's random method (Set 3).

For Set 1, we used 6,103 images as is from the TT100K training set.

For Set 2, we used 6,103 synthetic images that were generated using the proposed method. We used 6,103 images from the TT100K training set as the background images that included the traffic signs which were needed by the proposed method. We used Moiseev's degradation model for generating the traffic signs. In this experiment, we used only the brightness change, saturation change, Gaussian blur, Gaussian noise, and scaling from Moiseev's degradation model. We did not apply any geometric deformation to the sign, such as a three-dimensional rotation, translation, or projective transformation. The parameters that were given to the degradation model were tuned by hand.

For Set 3, we used 6,103 synthesized scene images that were generated by pasting the synthesized signs at random locations. To validate the impact of context information on model precision alone, we set all the conditions to be the same between Set 2 and Set 3, except for the locations at which the signs were pasted. Thus, we used same background images for Set 2 and Set 3 and made the number, size, and shapes of the synthesized signs included in the scene images to be the same.

There is one issue that must be resolved. The scene images used for Set 2 originally contain traffic signs. If these images are used as the background images for Set 3, then the scene images would simultaneously include both the traffic signs that were originally included in the scene images and synthetic traffic signs that were pasted at random locations. Therefore, we used the inpainting method of Telea et al. (Telea, 2004) to remove the traffic signs that were originally contained in the background images. We set the radius parameter for the inpainting to a value of 3. An example of traffic sign removal using inpainting is shown in Figure 9.

(a) Original image

(b) Inpainted image

Figure 9: Example of the removal of traffic signs using inpainting.

4.2 Training Method

Here, we explain the training method for faster R-CNN. The training parameters below were selected based on the research of Cheng et al. (Cheng et al., 2018) who performed training using a combination of TT100K and faster R-CNN. We used the momentum SGD for the network optimization method. The learning rate was initialized as 0.001 and was set as 0.0001 after the seventh epoch. The input resolution for the faster R-CNN was set as 1280×1280 pixels. Although it would have been ideal to perform the evaluation using the original resolution of TT100K, which was 2048×2048 pixels, we reduced the input resolution in this evaluation to improve the training speed. The training was conducted until the 15th epoch. GTX 1080 Ti was used for the training. 15 epochs of the training required approximately half a day.

4.3 Evaluation Results and Discussion

We used actual images (Set 1), synthesized images generated using the proposed image (Set 2), and synthesized images generated using Zhu's random method (Set 3) to train a total of three models. The progress of mean Average Precision (mAP) of the models is shown in Figure 10. The baseline shown in the graph refers to the random method. The comparison of mAP in the 15th epoch after the training has completed shows that the model trained using the proposed method has mAP that is approximately 8%

higher than that of the model trained using the random method. The difference between the proposed method and random method lies in only the locations at which the synthesized signs are pasted. Therefore, the results demonstrate that preserving the context information is important for improving the precision when generating the training data for an end-to-end traffic sign detector and classifier.

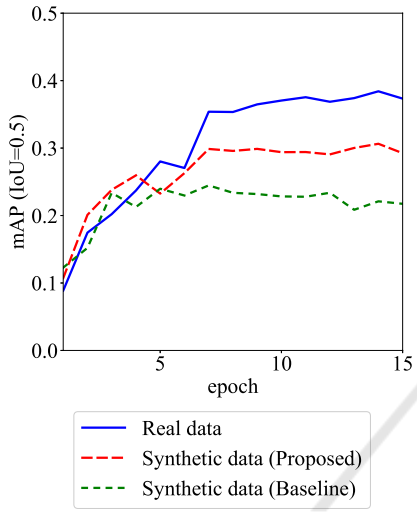

Figure 10: Relationship between epoch and mAP of faster R-CNN.

In contrast, mAP of the model that was trained using the proposed method was approximately 7% lower than that of the model trained using the actual images. It is thought that there are two reasons for this difference. The first reason is that in this study, we did not consider the context information when determining the values of the parameters given to the degradation model when generating the sign images. For example, we did not set the brightness parameter to match the brightness of the area surrounding the signs. Therefore, there are cases in which a dark synthesized sign is pasted on a bright area.

The second reason is that the edge between the synthesized signs and background image is obvious. It is necessary to use techniques such as alpha blending when pasting the traffic signs on the background images.

To analyze the difference in the performances of the proposed model and the model trained using the actual images, we calculated the accuracy-recall curves for each traffic sign size (Figure 11). The graphs demonstrate two points. For large signs that have a size in the range from 64 to 192 pixels, the difference between the proposed method and random method is significant. The graphs demonstrate the effectiveness of the proposed method. However, for the signs that have a size in the range from 32 to 64 pixels, the precisions of the proposed method and random method are both lower than that of the models trained using actual images.

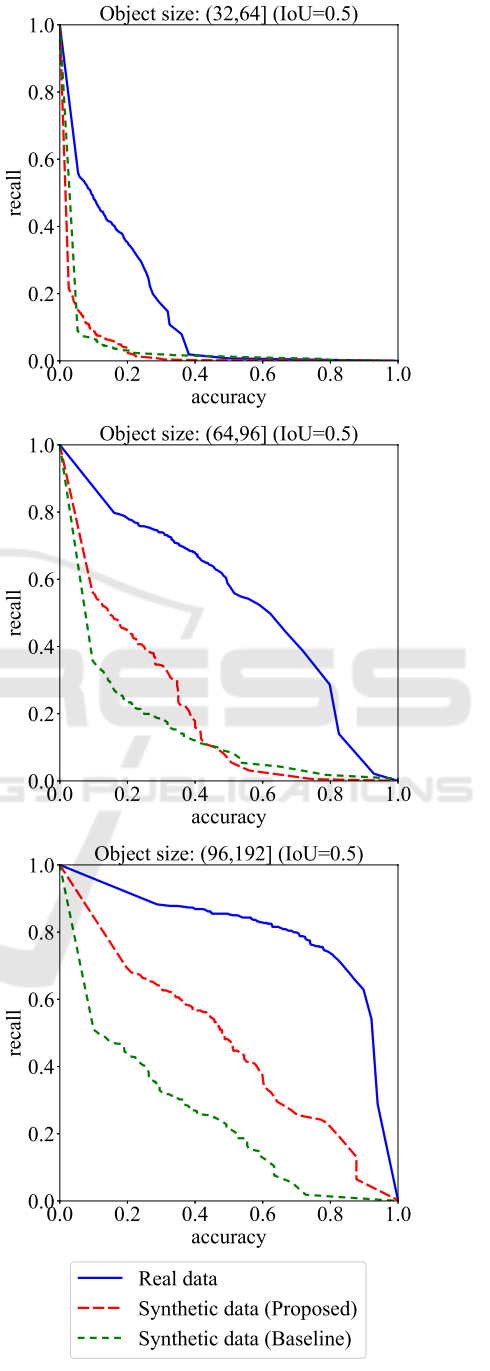

Figure 11: Accuracy-recall curves based on traffic sign size.

This suggests that detecting small target objects are relatively more influenced by their surrounding scenes. In this study, the context information was con-

sidered only in relation to selecting the locations for pasting the road signs. In the future, we will also consider the context in relation to setting the parameters to assign to the degradation model when generating the signs.

5 CONCLUSION

In this paper, we propose a method for training end-to-end traffic sign detectors without using actual images of the traffic signs. Our proposed method enables generating scene images that preserve the context information surrounding the traffic signs. The proposed method achieves mAP that is approximately 8% higher than that of the conventional method, in which the signs are pasted at random locations. This result demonstrates that training using scene images that preserve the context information is effective for improving the precision. However, mAP of the proposed method is approximately 7% lower than that of the sign detectors that are trained using actual images. The difference in the precision is high for signs that are relatively small in the scenes compared with the models trained using actual images. It would be possible to improve the precision by considering the context information when determining the values of the degradation parameters for generating synthetic traffic signs.

REFERENCES

Baird, H. S. (1992). *Document Image Defect Models*, pages 546–556.

Cheng, P., Liu, W., Zhang, Y., and Ma, H. (2018). Loco: Local context based faster r-cnn for small traffic sign detection. In *MultiMedia Modeling*, pages 329–341.

Fritsch, J., Kuehnl, T., and Geiger, A. (2013). A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems (ITSC)*.

Haselhoff, A., Nunn, C., Müller, D., Meuter, M., and Roese-Koerner, L. (2017). Markov random field for image synthesis with an application to traffic sign recognition. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1407–1412.

Hoessler, H., Wöhler, C., Lindner, F., and Kreßel, U. (2007). Classifier training based on synthetically generated samples. In *5th International Conference on Computer Vision Systems (ICVS)*.

Ishida, H., Takahashi, T., Ide, I., Mekada, Y., and Murase, H. (2006). Identification of degraded traffic sign symbols by a generative learning method. In *18th International Conference on Pattern Recognition (ICPR)*, volume 1, pages 531–534.

Ishida, H., Takahashi, T., Ide, I., Mekada, Y., and Murase, H. (2007). Generation of training data by degradation models for traffic sign symbol recognition. *IEICE TRANSACTIONS on Information and Systems*, E90-D(8):1134–1141.

Medici, P., Caraffi, C., Cardarelli, E., Porta, P. P., and Ghisio, G. (2008). Real time road signs classification. In *2008 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, pages 253–258.

Møgelmose, A., Trivedi, M. M., and Moeslund, T. B. (2012). Learning to detect traffic signs: Comparative evaluation of synthetic and real-world datasets. In *21st International Conference on Pattern Recognition (ICPR)*, pages 3452–3455.

Moiseev, B., Konev, A., Chigorin, A., and Konushin, A. (2013). Evaluation of traffic sign recognition methods trained on synthetically generated data. In *Advanced Concepts for Intelligent Vision Systems (ACIVS)*, pages 576–583.

Peng, E., Chen, F., and Song, X. (2017). Traffic sign detection with convolutional neural networks. In *Cognitive Systems and Signal Processing (ICCSIP)*, pages 214–224.

Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. (2012). Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323 – 332.

Telea, A. (2004). An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, 9(1):23–34.

Uršič, P., Tabernik, D., Mandeljc, R., and Skočaj, D. (2017). Towards large-scale traffic sign detection and recognition. In *22nd Computer Vision Winter Workshop (CVWW)*.

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