Advances in Intelligent Traffic and Transportation Systems
M. Shafik (Ed.)

doi:10.3233/ATDE230010

Scene Text Recognition for Text-Based Traffic Signs

Youssef TAKI[1] and Elmoukhtar ZEMMOURI
ENSAM Meknes, Moulay Ismail University, Morocco

Abstract. Scene Text Recognition (STR) enables the Advanced Driver Assistance System (ADAS) to recognize text in natural context, such as object labels, instructions, and text-based traffic signs. STR helps self-driving cars make informed decisions such as which direction to take, how fast to go, and what to do next. Traffic signs are categorized into three categories: traffic lights based on symbols and texts, and additional traffic. Traffic signs recognition is a very important task in ADAS, although many researchers have had impressive success with symbol-based traffic signs, there are very few researchers working on the other types of signs due to the difficulties they encounter, chief among which is the lack of publicly available datasets. In addition to the many factors that make text-traffic signs difficult to recognize, including complex backgrounds, noise, lightning conditions, different fonts, and geometric distortions in the signs. In this paper, we will survey some modern and effective methods of scene text recognition and discuss some of the problems they face, taking a closer look at the problem of text recognition of traffic signs in the first place.

Keywords. Scene Text Recognition (STR), text traffic signs recognition, Advanced Driver Assistance System (ADAS), Intelligent Transportation System (ITS), intelligent vehicles, deep learning, convolutional neural networks

1. Introduction

With recent advances in self-driving cars, and an advanced driver assistance system, Vehicle-mounted systems are expected to have a thorough understanding of their surroundings and to provide reliable information to drivers via road context indicators or autonomous navigation. Traffic panels/signs play a crucial role in this system. Drivers monitor traffic signs and act on the information provided by the signs. Drivers, intentionally or unintentionally, ignore traffic signs in various circumstances such as when the vehicle is at high speed or the driver is distracted by something, which can cause horrific accidents. Traffic panels/signs can be divided into 2 or 3 categories (Fig. 1): symbol-based traffic signs such as warning or mandatory signs, text-based traffic signs that contain many textual information and Supplementary traffic signs. Although there is a lot of research on symbol-based traffic sign recognition [1, 2, 3, 43], there is little research that focuses specifically on text recognition on traffic signs [4, 5, 7]. This could be due in part to the task's difficulty as a result of issues such as complex backgrounds, noise, lightning, different fonts, and geometrical distortions in the image,

[1] Corresponding Author, Youssef TAKI, ENSAM Meknes, Moulay Ismail University, Morocco; E-mail: youssefTaki1301@gmail.com.

but in my view, the most difficult aspect of this problem is the lack of a publicly available dataset such as The German Traffic Signs Detection Benchmark (GTSDB) [8] and The German Traffic Signs Recognition Benchmark (GTSRB) [9] in symbol-based traffic signs recognition problem.

Text reading in natural scenes, referred to as scene text recognition (STR), is often used for various purposes: translation, various card recognition to input personal information, text traffic sign recognition [Real Datasets], etc. Unlike Optical Character Recognition (OCR), which focuses on reading structured text: text in a typed document with a consistent background, row, font, and density. STR deals with the unstructured texts: text at random places in a landscape, sparse text, no proper row structure, complex background, random places in the image, no standard font. In this research, we focus on the problem of recognizing textual traffic signs, which suffers from all these problems such as curved texts, closed texts, etc., and which suffers from other critical problem, which is the lack of a suitable dataset.

To address the first part of the challenges, previous works [10, 11, 12] developed multiple models of STR, using deep neural networks, for example to process the variable number of characters in each input text, the recurrent neural network (RNN) was proposed to solve this problem by [10, 13], while [3] use Vision Transformer and attention mechanism to overcome the same problem. Whereas, for curvilinear text processing, spatial transformer network [14] (STN) has been proposed to normalize them.

For the second part of the challenges related to the dataset, this problem does not exist in the STR in general, because there are a lot of public suitable dataset, the synthetic datasets like MJSynth (MJ) [15], SynthText (ST)[16], and the real datasets like Street View Text (SVT) [17], ICDAR2003 (IC03) [39], SVT Perspective (SVTP) [42].

Figure 1. Instances of the three types of traffic signs (a) symbol-based traffic signs, (b) Supplementary traffic signs, (c) text-based traffic signs.

But for text-based traffic signs recognition, there is no suitable general dataset, to overcome this problem, many researchers generate and collect their data to train and evaluate their models [5, 7], But this task is very difficult, expensive and time consuming, especially for young researchers who want to work in this field.

The main contribution of this paper is to give a helping hand to all the beginners in this field and to give a general idea of the problems they are facing because as far as we know, this is the first paper that addresses the problem of text traffic sign recognition in relation to the problem of scene text recognition in general.

The rest of our paper is organized as follows: In Section 2, Several publicly available scene text recognition datasets are presented with some text-based traffic sign datasets discussed. In Section 3, a step-by-step explanation of the STR model framework is given. In Section 4, some STR and textual traffic signs methods are discussed. Ultimately, conclusions and perspectives on future studies are suggested in Section 5.

2. STR Datasets

In the scene text recognition problem, the data set used in the training and evaluation phase plays an important role in the performance of any model. In this section, we study the different training and evaluation datasets used in state of art, and then discuss the differences between them and which is more suitable.

2.1. Train Dataset

Since there is no large data set of real data, the practice in training a STR model is to use synthetic data. There are two main sets of synthetic data:

- MJSynth (MJ) [15] is a STR-specific synthetic dataset containing 8.9-million-word box images. As shown in Fig. 2a, each word is generated from 90,000 English words and over 1,400 different fonts.
- SynthText (ST) [16] Is just another synthetically generated dataset that was created with the intention of detecting scene text. Although SynthText was designed for scene text detection tasks, it was also used for STR by cropping word boxes. SynthText data contains 5.5 million cropped training word boxes, as shown in Fig. 2b.

In the STR training literature, each data set contributes 50% of the total training data set. Combining the two data sets at 100% resulted in performance degradation [2].

Figure 2. (a) MJSynth (MJ), (b) SynthText (ST).

2.2. Test Dataset

The test dataset consists of several small, real, and publicly available STR datasets for text in real images.

These data sets are generally grouped into two groups: first, regular datasets containing text images of horizontally placed characters with spaces in between, this type is relatively easy to identify by STR models, among which are:

- Street View Text (SVT) [17]: Outdoor Street images from Google Street View are included. Some of these images contain noise, are blurry, or have low resolution. SVT is made up of 257 training images and 647 evaluation images.
- ICDAR2003 (IC03) [39] was developed for the ICDAR 2003 Robust Reading competition for reading camera-captured scene texts. It includes 1,156 training images and 1,110 evaluation images.
- IIIT5K-Words (IIIT) [40] is a dataset compiled from Google image searches, with 2,000 images for training and 3,000 images for evaluation.

Second, irregular datasets contain text with difficult appearances such as curved, vertical, perspective, low resolution, or distorted, among which we mention:

- IC15 [41] contains images of text for the ICDAR2015 Robust Reading Competition. Many images are blurry, noisy, rotated, and sometimes low-resolution; the literature uses two versions: 1) 1,811 images and 2) 2,077 images.
- SVTP [42] contains 645 test images from Google Street View.
- CT [44] has 288 images that focus on curved text images captured from shirts and product logos.

Figure 3. (a) regular text, (b) irregular text.

2.3. Dataset for Text Signs Recognition

It can be seen by looking at papers published in text traffic signs problem over the past decade that in addition to their very small numbers compared to other fields, there is a long period of time between paper and paper.

As mentioned in the introduction to the paper, there is a shortage of text traffic signs dataset, unlike the symbol traffic signs, there are many benchmark datasets such as The German Traffic Signs Detection Benchmark (GTSDB) [8] and The German Traffic Signs Recognition Benchmark (GTSRB) [9], which explains the weak scientific research in this area. Among the first works on the recognition of textual traffic signs, there are [4], presented by Jack Greenhalgh, in this work the dataset used was obtained from Jaguar Land Rover Research captured by camera. there is also the work of X. Rong et al. [5], They've gathered a novel, difficult dataset of traffic and road guide panels. This dataset contains a wide range of highway guide panels, totaling 3841 high-resolution individual images, 2315 of which have traffic guide panel level annotations (1911 for training and 404 for testing, with all testing images manually labeled with ground truth tight text region bounding boxes), and 1526 of which have no traffic signs. All of the photos were taken from the AAroads website [6] and captured through the eye of a dash camera mounted on a car and include a variety of traffic advisory panels

such as direction, toll plaza, destination distance, and exit indicator. [5]. The most recent work in this field is the work of Sana Khairinejad et al. [7] in 2022, in this work new data collecting and labeling (Persian text-based traffic panels in Iran), this dataset contains 4000 images with 12 Gigabyte sizes. Since different persons and cameras have taken the images, they are in both vertical and horizontal shapes, and their sizes are different. (Fig. 4)

Figure 4. Examples of the main dataset collected by Sana Khairinejad et al. [7]

3. STR Model Framework

Text reading in natural scenes is generally divided into two tasks (Fig. 5): the detection of text regions in scene images, referred to as scene text detection (STD), and text recognition of regions, referred to as scene text recognition (STR) [1]. Once you have detected the bounding boxes that contain the text, the next step is to recognize the text.

Figure 5. Text reading in natural scenes pipeline.

In this work, our focus is on a text recognition (STR) task, in which a STR identifies each text character in an image in the correct sequence. Unlike object recognition, in which there is generally only one class of object, a text image may contain zero or more characters. Thus, STR models are more complex. [3]

The objective of this section is to introduce the four-stage text recognition (STR) framework that was presented by Baek et al. [2] and describe the module options in each stage that have been used by previous works like [2, 3, 10, 11, 12].

According to [2], STR is implemented in four stages: 1) Transformation or Rectification, 2) Feature Extraction (Backbone), 3) Sequence Modelling, 4) Prediction (Fig. 6).

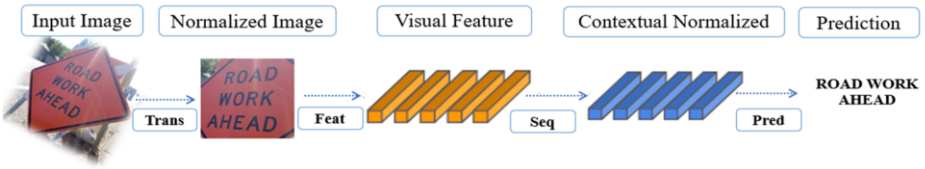

Figure 6. Visualization of an example flow of scene text recognition [2]. The model decomposes into four stages.

1) Transformation (Trans):

As mentioned above, whether we are talking about text recognition in general or about textual traffic signs, Text images in natural scenes come in diverse shapes, curved, tilted, and distorted texts. If these input images are fed unchanged, the next feature extraction stage will have difficulty extracting patterns and this affects accuracy or needs to know a consistent representation with respect to this geometry which needs more data sets and increases time consumption. To reduce this burden, the transformation stage removes distortion from images and normalizes perspective or curved text to horizontal or normal text. [2] This makes it easier for Feature Extraction (Backbone) module to determine invariant features. This is generally done by the Spatial Transformer Network (STN) [14]. Spatial Transformer modules, introduced by Jaderberg et al. [14], are a popular method for increasing a model's spatial invariance against spatial transformations such as translation, scaling, rotation, cropping, and non-rigid deformations. They can be added to existing convolutional architectures in two ways: immediately after the input or in deeper layers. Figure 7 presents the spatial transformer mechanism in three parts.

Thin-Plate-Spline transform (TPS) [18], a type of spatial transformations network (STN) [14], and most famously, has been applied in many papers, such as RARE (Robust-text recognizer with Automatic Rectification) [19], STAR-Net (SpaTial Attention Residue Network) [20], and TRBA (TPS-ResNet-BiLSTM-Attention) [2], A spatial attention residue network for scene text recognition [21]. Some papers use some other type of transformation, while in some cases no transformation is used as in CRNN (Convolutional Recurrent Neural Network) [22], and Rosetta [23].

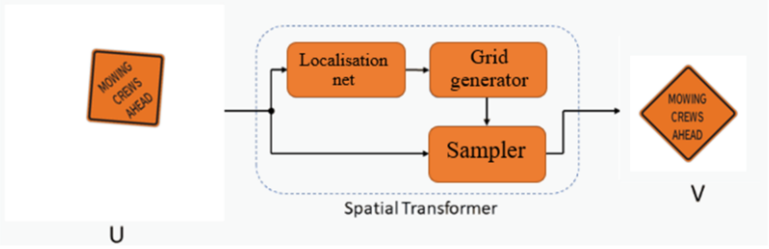

Figure 7. Spatial transformer module transforms inputs to a canonical pose by 3 steps: Localization net, Grid generator and sampler.

2) Feature extraction (Feat):

The role of Feature Extraction (Backbone) stage is extract visual feature representation of each character symbol from the input image. This is generally performed by a module composed of convolutional neural networks (CNNs) [1], which are the same feature extractors used in object recognition tasks [3].

The 4 most popular architectures used are VGG [24], RCNN [25], ResNet [26], and SqueezeNet [27], for example Rosetta, STAR-Net and TRBA use ResNet. RARE and CRNN extract features using VGG. It is now common to use transformer decoder-based models to replace CNN to extract features in new papers such as [3].

3) Sequence modeling (Seq):

Since STR is dependent on sequence prediction, we are talking about a long-term dependency. This is where the need for the sequence modeling stage arises .The sequence modeling stage's role is to establish a consistent context between recent and historical character features [3]. Therefore, some previous work such as CRNN, GRCNN, RARE, STAR-Net and TRBA use BiLSTM after feature extraction phase. But this stage takes a lot of computing time and memory. To reduce computational complexity and memory consumption, some papers such as Rosetta [23] remove this stage completely and pass directly to the Prediction stage.

4) Prediction (Pred):

The role of the prediction stage is to predict character sequences from contextual features coming from the feature extraction stage or sequence modeling stage. By summarizing the previous work, we generally have three prediction options: (1) Connectionist temporal classification (CTC) [28], (2) attention-based sequence prediction (Attn) [11, 29], and (3) transformers [3].

4. STR Methods

There are several techniques for scene text recognition. We will be discussing some of the best techniques in this section.

4.1. CRNN

Convolutional Recurrent Neural Network (CRNN) is a combination of CNN, RNN, and CTC (Connectionist Temporal Classification) for scene text recognition. The network architecture has been published in 2015 by Baoguang Sh et al [22]. This neural network architecture consists of 3 parts (Fig. 8): 1) Convolution layers which extract features sequence from the input image, 2) a deep bidirectional recurrent neural network predicts label sequence with some relation between the characters, 3) transcription layer converts the output that comes from RNN into a label sequence.

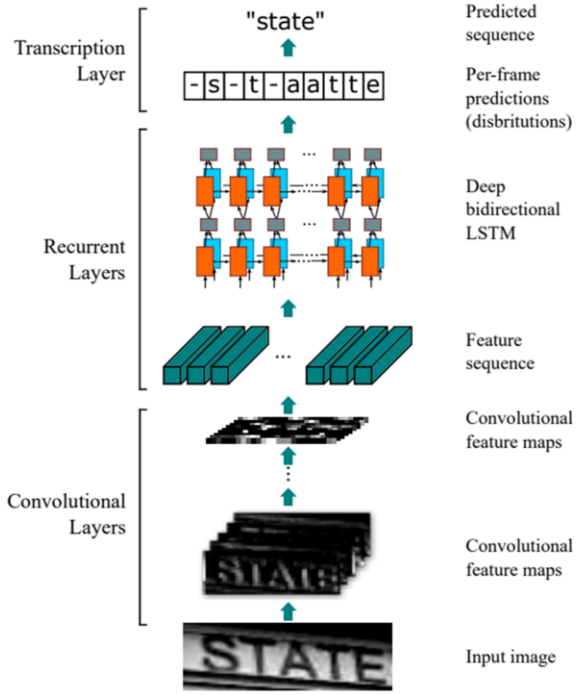

Figure 8. The CRNN network architecture [22].

4.2. TRBA

From its name, the architecture of this method consists of 4 parts: the first is a thin-plate spline (TPS) in the transformation stage, the second is ResNet as feature extractors, the third is Bi-Directional LSTM (BiLSTM) in the Sequence modeling stage, and finally, the attention-based sequence prediction model (Attn) in the prediction stage, this network architecture was published in 2019 by Jeonghun Baek et al [2] and since that time a lot of research papers based on it have been published like [1, 3].

In this section, we tried to analyze some previous methods, in Table 1, we summarize a comparison of the previous methods for STR problems:

Table 1. Comparison of different methods for scene text recognition

Model	Train data	IIIT 3000	SVT 647	IC03 860--867	IC13 857--1015	IC15 1811--2077	SP 645	CT 288	Year
CRNN [22]	MJ+ST	81.8	80.1	91.7--91.5	89.4--88.4	65.3--60.4	65.9	61.5	2015
R2AM [25]	MJ+ST	83.1	80.9	91.6--91.2	90.1--88.1	68.5--63.3	70.4	64.6	2016
RARE [19]	MJ+ST	86.0	85.4	93.5--93.4	92.3--91.0	73.9--68.3	75.4	71.0	2016
STAR-Net [20]	MJ+ST	85.2	84.7	93.4--93.0	91.2--90.5	74.5--68.7	74.7	69.2	2016
GCRNN [30]	MJ+ST	82.9	81.1	92.7--92.3	90.0--88.4	68.1--62.9	68.5	65.5	2017
Rosetta [23]	MJ+ST	82.5	82.8	92.6--91.8	90.3--88.7	68.1--62.9	70.3	65.5	2018
TRBA [2]	MJ+ST	87.8	87.6	94.5--94.2	93.4--92.1	77.4--71.7	78.1	75.2	2019
ViTSTR [3]	MJ+ST	88.4	87.7	94.7--94.3	93.2--92.4	78.5--72.6	81.8	81.3	2021

But for the problem of textual traffic signs, we cannot compare the existing methods because the comparison would not be fair without a publicly available

reference dataset. But in general, the STR framework is the same for the recognition of text signals, we first need to use standard trap detection techniques (YOLO, SSD,..) to detect the text in the image and create a bounding box around the part of the image that contains text, once we detect the bounding boxes that contain text, then the next step is to recognize the text by following the four stages of the STR framework.

5. Conclusion and Discussion

The STR problem is the root of all sub-problems related to text recognition such as text traffic lights, which we want to focus on more, but a deeper understanding of the source, makes us better understand our problem. In the STR problem There is general knowledge that training STR models on real data is almost impossible because real data is insufficient. But in 2021, Jeonghun Baek et al [1] refutes this hypothesis by only showing sufficient Performance using real data with fewer labels. In this research, due to the small size of the real data set compared to the synthetic data (about 1.7% of the synthetic data), they used some data augmentation techniques, along with the use of semi and self-supervised methods (Pseudo-Label (PL) [35], Mean Teacher (MT) [36], RotNet [37], Momentum Contrast (MoCo) [38]) in their model to overcome this problem.

Coming back to our problems, the data is also the biggest obstacle here, according to a benchmark study [2], it is difficult to obtain sufficient real data due to the high cost of labeling. Thus, we need to apply new methods in this problem to overcome this problem as they did in [1], such as unsupervised and semi-supervised learning.

Through this survey of comparative studies of different ways to identify the text of the scene. The STR problem obviously has great utility and efficacy, but text-based traffic recognition is still far from what ADAS is expected to do, in this paper, we have tried to build a base for us and lay the foundation for all researchers who want to work and research this problem.

References

[1] Jeonghun Baek, Yusuke Matsui, Kiyoharu Aizawa, What If We Only Use Real Datasets for Scene Text Recognition? Toward Scene Text RecognitionWith Fewer Labels, arXiv:2103.04400v2 [cs.CV] 5 Jun 2021

[2] Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J., Lee, H.: What is wrong with scene text recognition model comparisons? dataset and model analysis. In: ICCV. pp. 4715{4723 (2019)

[3] Rowel Atienza, Vision Transformer for Fast and Efficient Scene Text Recognition, arXiv:2105.08582v1 [cs.CV] 18 May 2021

[4] Jack Greenhalgh and Majid Mirmehdi, Recognizing Text-Based Traffic Signs, in IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (2015)

[5] Xuejian Rong, Chucai Yi, and Yingli Tian, Recognizing Text-Based Traffic Guide Panels with Cascaded Localization Network, ECCV 2016 Workshops, Part I, LNCS 9913, pp. 109–121, 2016.

[6] http://www.aaroads.com.

[7] Saba Kheirinejad, Noushin Riahi, Reza Azmi, Text-Based Traffic Panels Detection using the Tiny YOLOv3 Algorithm (2022), https://dx.doi.org/10.55708/js0103008

[8] Houben S, Stallkamp J, Salmen J, Schlipsing M, Igel C (2013) Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In: The 2013 international joint conference on neural networks (IJCNN). IEEE, pp 1–8

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

[10] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. In TPAMI, volume 39, pages 2298–2304. IEEE, 2017. 1, 2, 4, 5, 10, 11, 12
[11] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In CVPR, pages 4168–4176, 2016. 1, 2, 3, 4, 5, 10, 11, 12
[12] Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Wassell. Synthetically supervised feature learning for scene text recognition. In ECCV, 2018. 1, 2
[13] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In ICCV, pages 5086–5094, 2017. 1, 2, 3, 4, 5, 10, 11, 12, 13
[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In NIPS, pages 2017–2025, 2015. 4
[15] Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and arti-_cial neural networks for natural scene text recognition. NIPS Workshop on Deep Learning (2014)
[16] Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in nat- ural images. In: CVPR. pp. 2315{2324 (2016)
[17] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In ICCV, pages 1457–1464, 2011. 3
[18] Bookstein, F.L.: Principal warps: Thin-plate splines and the decomposition of de- formations. Trans on Pattern Analysis and Machine Intelligence 11(6), 567{585 (1989)
[19] Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X.: Robust scene text recognition with automatic recti_cation. In: CVPR. pp. 4168{4176 (2016)
[20] Liu, W., Chen, C., Wong, K.Y.K., Su, Z., Han, J.: Star-net: a spatial attention residue network for scene text recognition. In: BMVC. vol. 2, p. 7 (2016)
[21] Wei Liu, Chaofeng Chen, Kwan-Yee K Wong, Zhizhong Su, and Junyu Han. Star-net: A spatial attention residue network for scene text recognition. In BMVC, volume 2, 2016. 1, 2, 4, 11
[22] Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. Trans on Pattern Analysis and Machine Intelligence 39(11), 2298{2304 (2016)
[23] Borisyuk, F., Gordo, A., Sivakumar, V.: Rosetta: Large scale system for text de- tection and recognition in images. In: Intl Conf on Knowledge Discovery & Data Mining. pp. 71{79 (2018)
[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015. 4, 12
[25] Chen-Yu Lee and Simon Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. In CVPR, pages 2231–2239, 2016. 1, 2, 4
[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016. 4, 12
[27] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, Kurt Keutzer, SQUEEZENET: ALEXNET-LEVEL ACCURACY WITH 50X FEWER PARAMETERS AND <0.5 MB MODEL SIZE, arXiv:1602.07360v4 [cs.CV] 4 Nov 2016
[28] Alex Graves, Santiago Fern´andez, Faustino Gomez, and J¨urgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In ICML, pages 369–376, 2006. 1, 5, 13
[29] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In ICCV, pages 5086–5094, 2017. 1, 2, 3, 4, 5, 10, 11, 12, 13
[30] Wang, J., Hu, X.: Gated recurrent convolution neural network for ocr. In: NeuRIPS. pp. 334{343 (2017)
[31] X. Rong, C. Yi and Y. Tian, "Recognizing text-based traffic guide panels with cascaded localization network," in European Conference on Computer Vision, 2016
[32] J. Greenhalgh and M. Mirmehdi, "Recognizing Text-Based Traffic Signs," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 13, pp. 1360 -- 1369, 2015.
[33] Y. Zhu, M. Liao, M. Yang and W. Liu, "Cascaded Segmentation-Detection Networks for Text-Based Traffic Sign Detection," IEEE Transaction Intelligent Transportation Systems, vol. 19, no. 1, pp. 209 -- 219, 2018.
[34] X. Peng, X. Chen and C. Liu, "Real-time Traffic Sign Text Detection Based on Deep Learning," in Materials Science and Engineering, 2020.
[35] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Workshop on challenges in representation learning, ICML, 2013. 4, 5, 18
[36] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In NeurIPS, 2017. 4, 5, 18

[37] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In ICLR, 2018. 5, 7, 18
[38] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In CVPR, 2020. 5, 7, 18.
[39] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, and Robert Young. Icdar 2003 robust reading competitions. In ICDAR, pages 682–687, 2003. 3
[40] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In BMVC, 2012. 3
[41] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In ICDAR, pages 1156–1160, 2015. 3
[42] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In ICCV, pages 569–576, 2013. 3
[43] Y. Taki, E. Zemmouri. (2020) An Overview of Real-time Traffic Sign Detection and Classification. In procedings of the 5th International Conference on Smart City Applications, October 7-9, 2020, Safranbolu, Türkiye.
[44] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. In ESWA, volume 41, pages 8027–8048. Elsevier, 2014.