Hierarchical Feature Fusion With Text Attention For Multi-scale Text Detection

Chao Liu[1], Yuexian Zou[1,2*] and Wenjie Guan[1]

[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
*Corresponding author: zouyx@pkusz.edu.cn

***Abstract*—Scene text detection for practical application remains challenging since it is a typical multi-scale object detection task with complex and varying condition. It is noted that the single shot detector (SSD) has shown predominance in object detection among deep learning-based methods but has limited capability in handling multi-scale text detection (MTD) task. In this paper, we strive to improve the performance of MTD task under SSD framework. Specifically, a novel single-shot word-level text detector is proposed. First, to extract the features keeping text-related information for multi-scale text objects, a hierarchical feature fusion module is designed to capture multi-scale inception features and multi-level features with semantic information. Second, to suppress background disturbance in the feature map, a text attention module is developed to learn the location information of the texts existing in the image, which helps to accurately inferring the words, especially when they are of extremely small size. Experimental results on three public word-level texts datasets demonstrate the effectiveness of our proposed method which achieves F-measure 0.89 and 0.79 on the ICADR2013 and ICDAR2015 respectively, and achieve highest accuracy of 0.87 on the SVT.**

Keywords- Multi-scale scene text detection, single-shot detector, hierarchical feature fusion, text attention.

I. INTRODUCTION

Recently, scene text detection has received plenty of attention and made tremendous progress under the development of deep convolutional neural networks (DCNNs). It is noted that multi-scale text detection (MTD) task still remains challenging. As shown in Fig.1, we can see that billboards on the commercial street and slogans in the stores are common scenes in reality, which contain different scales of texts and have a lot of small-scale texts in the distant view. The differences between general objects and texts mainly lie on the diversity, size and aspect ratio. Based on the above reasons, from experimental results, we can see that the performance of the state-of-art text detection methods is still unsatisfactory for MTD task.

The main stream DCNNs based text detection methods could be grouped into two main streams: region-based methods and region-free methods. Firstly, the region-based methods [1-3], generally modified from Faster R-CNN[4], divide the text detection task into two steps: Region Proposal Network (RPN) and RoI-wise Classification Network (RCN). One of the concerns for region-based methods is their computational cost. Besides, they employs the top-most layer of ConvNet to detect objects at different scales, it imposes the burden of this single layer while ignoring other multi-scale layers. Secondly, region-free based methods[5-8], universally modified from SSD[9], employ a single feed-forward convolutional network to directly predict bounding boxes and their classes. Therefore, compared to region-based methods, most of them run faster and are able to meet the requirement of real-time text detection. Meanwhile, multiple features within a ConvNet are used by the region-free based methods to predict texts of different scales which maintain higher detection accuracy as well.

Fig.1. Examples of multi-scale text detection. Billboards on the commercial street (left) and slogans in the store (right).

However, preliminary experiments evidence that the performance of the region-free based methods for multi-scale text detection is unsatisfied since they cannot accurately detect most of the small-scale text objects. Analyzing shows that the small-scale texts usually generate small feature maps on higher layers which may lose relevant location information but maintain more semantic information. It is noted that the lower layer features are of high resolution, where the location information of the small-scale text object may retain but contain less semantics.

According to above observations, in this work, we employ SSD as our backbone network to keep its advantages in terms of higher object detection accuracy and better computational efficiency. Moreover, for natural scene text detection tasks which have large proportion of small-scale texts, it is observed that SSD-based methods are inefficient or inaccurate in detecting small-scale objects. To solve this, we proposed a novel word-level text single-shot detector. First, we inspired by recent fusion fashion idea to design a hierarchical feature fusion module aims to capture multi-scale inception features and multi-level features with semantics. Second, to suppress background disturbance in the feature map, a text attention module is developed which coarsely identify text regions via a learned attention map. It is essentially to generate accurate inference of words, especially for extremely small size.

The remaining of the paper is organized as follows. Sec.2 presents pipelines and algorithms of our proposed MTD methods. Sec.3 shows the experimental results and performance analysis. Finally, Sec.4 gives the conclusion.

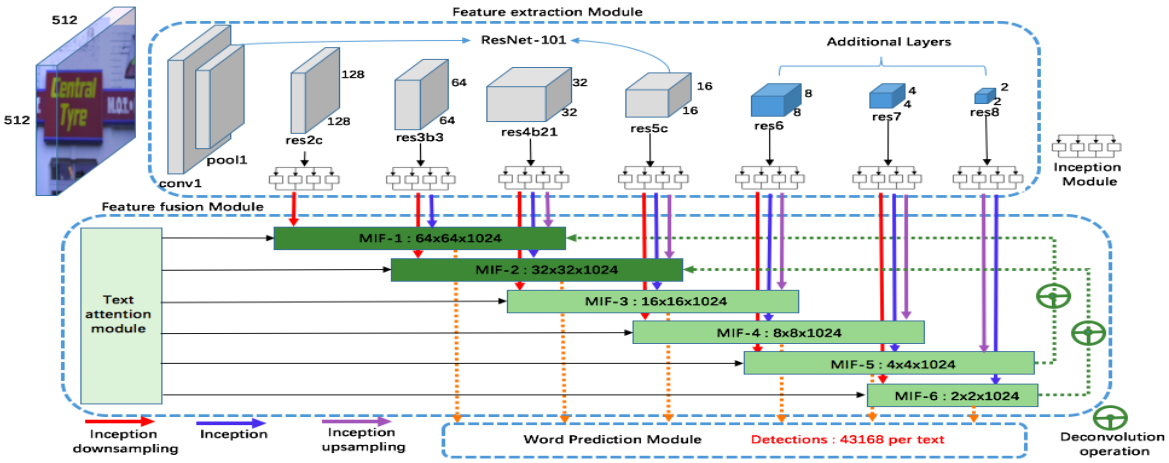

Fig.2. The architecture of our proposed multi-scale text detection method

II. PROPOSED MTD METHOD

In this section, the pipeline and algorithms of our proposed MTD method are described in details. The pipeline of our method is depicted in Fig.2. Our method mainly contains three modules: the feature extraction module, the feature fusion module and the word prediction module. In our design, for the feature extraction module, ResNet-101[10] is taken as the basic network instead of VGG-16 which is used in original SSD. Meanwhile, the inception module is designed for getting features in wide scope of scales. For the feature fusion module, a hierarchical feature fusion module cascaded with a text attention module are implemented. For the word prediction module, it incorporates output of multiple layers to obtain the final detection results by a non-maximum suppression (NMS)[9] process. The details will be addressed in the following subsections respectively.

2.1 Feature extraction module

Aiming to extract highly abstracted features and obtain more contextual information of different levels to improve the performance of detection, we use deeper convolutional CNN structure, like ResNet-101. As shown in Fig. 2, we inherit the ResNet-101, keeping the layers from conv1 to res5c unchanged, and appending three additional convolutional layers, named as res6, res7 and res8, respectively.

In our design, we take the last layer of each stage in ResNet-101 as the detection layers. As shown in Fig.2, these detection layers are denoted as res2c, res3b3, res4b21, res5c, res6, res7, and res8, respectively, which are associated with different scales of default boxes to predict word-level bounding boxes. Besides, inspired by the inception architecture of GoogLeNet[11], we design an inception module for each detection layer to capture multi-scale feature by using different sizes of convolutional kernels. Details of the inception module are shown in Fig.3. By doing so, the final inception features will have multi-scale receptive fields and are able to focus on image content in a wide scope of scales. From Fig. 3, we can see that the inputs of the inception module are processed through four different operations: 1×1 *conv,* 3×3 *conv* with 1x1 *conv*, 3×3 *maxpooling* with 1x1 conv, and 5×5 conv where the 5×5 *conv* is divided into 5×1 and 1×5 dilated convolution layers[12] and set the dilation rate as 2.0 for computational efficiency. The aim of using dilated convolutional layers is to increase the area of receptive field exponentially without decreasing its spatial dimensions. The final output of the inception module is the 1024-channels inception features which are generated by concatenating four 256-channels inception features.

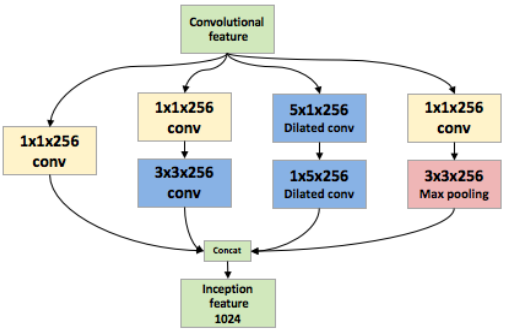

Fig.3. Our proposed inception Module.

2.2 Feature fusion module

For the feature fusion module, a hierarchical feature fusion module cascaded with a text attention module are implemented.

2.2.1. Hierarchical feature fusion module

We proposed a hierarchical feature fusion module for incorporating semantic and contextual information in feature maps. Specifically the hierarchical feature fusion method consists of two modules, the inception feature fusion module and the multi-level feature fusion module. The design details are given as follows.

The multi-layer inception feature fusion module: which is inspired by HyperNet[13]. To further enhance the convolutional features by aggravating Multi-layer Inception Features of the detection layers named as MIF in this paper, and the inception features are generated from the inception module described in Sec2.1. In our design, the MIFs are computed by three adjacent layers: the preceding inception feature layer, the current inception feature layer and the behind inception feature layer, which is shown in Fig.4. It is noted that down-sampling and up-sampling are utilized for the lower inception feature and higher inception feature respectively, which ensure the same resolution, and element-wise addition is implemented for generating the fused features. As shown in Fig. 2, at the last processing of this module, there are two inception features left which are res7 inception feature and res8 inception feature. We down-sample the res7 inception feature to have the same feature resolution with that of res8 inception feature.

Then an extra MIF is generated. In total, six MIFs are generated which are combined to yield a 1024-channels feature vector.

The multi-level feature fusion module: It is a fact that the lower layers feature maps are of higher resolution containing abundant details but less semantic information. To improve the small-scale texts detection capability, the abstract semantic features from higher layers are fused into the lower layers. To implement the fusion, we up-sampled the feature maps of the higher layers via a deconvolutional operation to keep the higher layer features and lower layer features with the same resolution. And then the element-wise addition can be operated. It is clear that the fused features contain more semantic and contextual information which is good for detecting small-scale text objects. In our design, as shown in Fig. 2, the MIF-5 and MIF-6 are up-sampled which are fused with MIF-2 and MIF-1 respectively.

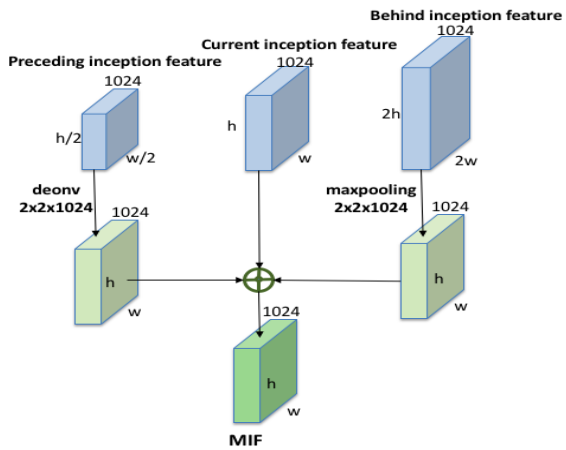

Fig.4. Our proposed MIF fusion module. (Features fusion is operated by element-wise addition)

2.2.2. Text attention module

The existence of ambiguous and small-scale texts affects the performance of text detection. We employ the text attention module trying to learn the probability heat map of texts. The learned probability heat map fed back into the feature maps of each detection layer to enhance the text-related features which aims to suppress the background disturbance in the feature map.

As shown in Fig. 2 , the text attention module is formulated in our end-to-end architecture as a unified framework. This module intends to learn the spatial regions of text from MIFs. In our design, we choose MIF-1 as the input of this module since MIF-1 has high feature resolution providing more localization information for small-scale or low-level texts. It is noted that the learned probability heatmap of texts will indicate the location information of the texts existing in the image. Then, the probability heatmap of text is taken as the text attention map which is resized into the same resolution of MIFs. This operation encodes the text attention information into MIFs to enhance the text-related features. The flowchart of this module is depicted in Fig.5. It is worthy to address that, for learning the text attention map in a supervise manner, an auxiliary loss via a binary mask indicating the text or non-text at each pixel location as well as a softmax are adopted. The input of this module is the feature of MIF-1, it resizes into the same resolution with original image by deconvolutional operation, and filter by 1×1 conv to obtain 2-channels features which is $F \in R^{512\times512\times2}$. Then, we aim to obtain text attention by judging the positive or negative of the 2-channels feature through a softmax function which optimizes text attention toward the provided text binary mask :

$$\alpha = softmax(Conv_{1\times1}(deconv_{3\times3}(F_{MIF-1}))) \qquad (1)$$

Then the positive of the softmax function, $\alpha^{+} \epsilon\, R^{512\times512}$ is defined as the text attention map, which manifests the pixel-wise possibility of the existing of the texts. In the end, the text attention map is encoded into six MIFs, by resizing text attention maps into corresponding spatial size of MIFs.

$$\widehat{F}_{MIF} = resize(\alpha^{+}) \odot F_{MIF} \qquad (2)$$

where $\widehat{F}_{MIF}$ is the resultant multi-layer inception features with the text attention. As discussed above, it is expected that the learned text attention map is able to provide the text location related information which helps to detect ambiguous texts and reduce the number of false detection.

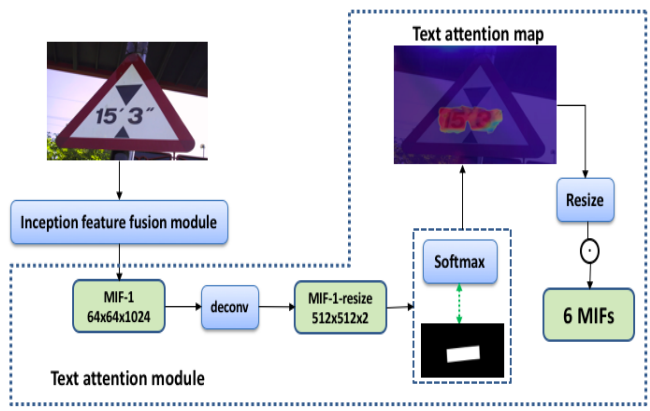

Fig.5. The flowchart of the proposed text attention module.

2.3 Word prediction module

In principle, we follow the conventional approaches to generate the text bounding boxes, where NMS and the same loss function are used. The loss function is defined as follows[9]:

$$L(x,c,l,g) = \frac{1}{N}(L_{conf}(x,c) + \alpha L_{loc}(x,l,g)) \qquad (3)$$

where N is the number of default boxes. In our algorithm, the smooth L_1 loss is used for L_{loc} and a 2-class softmax loss is adopted for L_{conf}. To obtain more accurate bounding boxes, we further take the following approaches. First, we consider the differences between general objects and texts which are diversity, size and aspect ratio. Thus, we are inspired by [6] to modify the original aspect ratios for general object detection, we set the eight aspect ratios as: (0.5, 2, 3, 4, 5, 7, 9, 11), and a vertical offset which is half of height of the cell is set to solve default boxes densely on the horizontal direction. This generates 72 default boxes in total for each MIF, which can handle abundant shape variances of texts. Second, it is noted that the text objects usually have rectangular shape rather than square. To fit this property, the 1×5 convolutional filter is used to replace the original filter 3×3 filter. The aim of these improvements is to adapt the practical situation with quadrilateral shape of texts.

III. EXPERIMENTS

To evaluate our proposed method, we conduct extensive experiments on three mainstream benchmarks: ICDAR2013, ICDAR2015 and Street View Text.

Table 1. Comparison of state-of-art results on the ICDAR 2013, ICDAR 2015 and SVT-50.

ICDAR 2013					ICDAR 2015				SVT-50	
Method	Recall	Precision	F-measure	Time(s)	Method	Recall	Precision	F-measure	Method	Precision
SSD[9]	0.60	0.80	0.69	0.10	Seglink[7]	**0.77**	0.73	0.75	Jaderberg[14]	0.68
FCRN[15]	0.76	**0.94**	0.84	1.27	HUST_MCLAB	0.44	0.38	0.41		
Pan He[8]	**0.86**	0.88	0.87	0.13	Yao *et al.*[16]	0.57	0.72	0.64	TextBoxes	0.84
CTPN[17]	0.73	0.93	0.82	0.14	CTPN	0.52	0.74	0.61		
TextBoxes[6]	0.74	0.88	0.81	1.29	Pan He	0.73	0.80	0.77	FCRN all[15]	0.76
Our Method	0.85	0.91	**0.89**	0.57	**Our Method**	0.75	**0.84**	**0.78**	**Our Method**	**0.87**

Our experiments are conducted on two NVIDIA Titan X GPUs with 12GB memory each. In our experiments, we use SynthText[15] as our model pretrainig data which 250k images images have been used. The training image resolution are of 512x512 and the well-known training manner Stochastic Gradient Descent (SGD) is used. Momentum and weight decay are set to 0.9 and 0.0005 respectively. Base learning rate is set as 0.0001. A multistep learning rate policy is adopted. The NMS threshold and confidence threshold are ste 0.3 and 0.7, respectively. We first train our model on SynthText for 120K iterations, then finetune it on ICDAR2013, ICDAR2015 and SVT, respectively. The other settings followed the scheme by [9]. For illustration purpose, some text detection examples of our method compared with TextBoxes[6] are shown in Fig.6. It is clear to see that our proposed method gives more accurate text detection results, especially for more difficult text objects (small-scale and ambiguous texts), which is even difficult task for human.

The experimental results on ICDAR2013, ICDAR2015, and SVT-50 are given in Table.1, where five state-of-art MTD methods are compared for ICDAR2013 and ICDAR2015, respectively. Meanwhile, there are three comparison methods for SVT. To give over all evaluation, recall, precision and F-measure are taken as performance measures. It is noted that F-measure can be considered as a good metric for accurately evaluating the detection performance. As shown in Table.1, for ICDAR2013, our method gives the highest score of F-measure, third best score of Precision and seconds best of Recall. For ICDAR2015, it is more encouraged that our method obtains best results in Precision and F-measure and the second best in Recall. For SVT-50, our method also gets the best score.

Analyzing the experimental results of our method shown in Table 1, we have the following discussions: 1) **On the ICDAR 2013[18],** this dataset is designed for near-horizontal text detection which focuses on world-level evaluation. From the results, we can conclude that our method is comparable for normal word-level text detection task and even provides more accurate detection results since we get highest F-measure (0.89). 2) **On the ICDAR 2015[19],** the characteristic of this dataset are multi-scale and multi-oriented, which fits part of our research focus that is multi-scale problem. From the results shown in Table 1, we can see that our method achieves the highest score of precision (0.84) and F-measure (0.78) respectively which demonstrates our method is robust for MTD task. 3) **On the SVT-50[20],** the images in this dataset have the lower image resolution than ICDAR2013 and ICDAR2015 which is difficult for detection since contains less contextual information in the convolutional features, but we get the highest accuracy of 0.87, and the second best is 0.84 achieved the highest accuracy of 0.87, and the second best is 0.84 achieved by TextBoxes. This result clearly indicates that our method has better ability to handle ambiguous and blurring texts since we employ the text attention in our module. By analyzing the results of three datasets above, our method can handle MTD task more powerful than recent state-of-art methods

We take an experimental method to rough evaluate the computational cost of our method. It is a common practice to compute the GPU running time for CNN-based methods [8]. We compare running time of various methods on ICADR 2013 are given in Table 1. It is clear to see that our method takes 0.57s for detecting one image using a single GPU, which is about at the middle in terms of computational cost. There are some rooms to improve the computational efficiency.

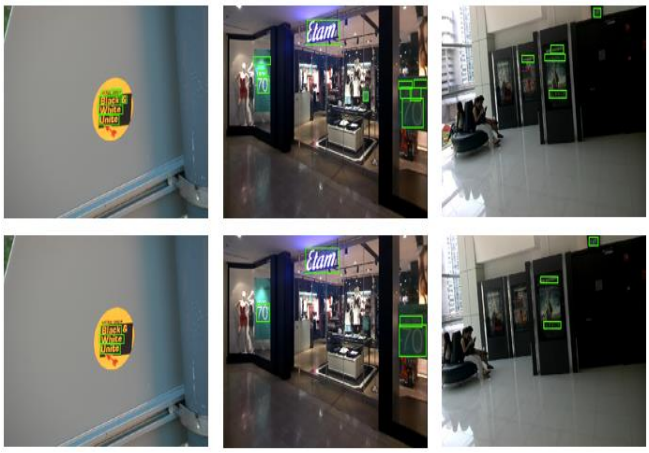

Fig.6. Illustration of Detection results. Top: our proposed method; Bottom:TextBoxes[6].

IV. Conclusion

In this study, aiming to promote multi-scale texts detection performance, we proposed a novel end-to-end MTD system under SSD framework. Making use of the properties of the nature scene text detection tasks, the feature extraction model, feature fusion module, text attention module are carefully designed. Experimental results validate the effectiveness of our proposed MTD method which outperforms previous state-of-art text detection approaches on word-level annotated benchmarks. Our method achieves F-measure 0.89 and 0.79 on the ICADR2013 and ICDAR2015 respectively, and achieve highest accuracy of 0.87 on the SVT.

V. ACKNOWLEDGMENT

This paper was partially supported by the Shenzhen Science & Technology Fundamental Research Program (No: JCYJ20160330095814461) & Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (ZDSYS201703031405467). Special Acknowledgements are given to Aoto-PKUSZ Joint Research Center of Artificial Intelligence on Scene Cognition & Technology Innovation for its support.

VI. REFERENCE

[1] W. He, X. Y. Zhang, F. Yin, and C. L. Liu, "Deep Direct Regression for Multi-Oriented Scene Text Detection," pp. 745-753, 2017.

[2] Y. Dai *et al.*, "Fused Text Segmentation Networks for Multi-oriented Scene Text Detection," 2017.

[3] Y. Nagaoka, T. Miyazaki, Y. Sugaya, and S. Omachi, "Text Detection by Faster R-CNN with Multiple Region Proposal Networks," in *Iapr International Conference on Document Analysis and Recognition*, 2018, pp. 15-20.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *International Conference on Neural Information Processing Systems*, 2015, pp. 91-99.

[5] X. Zhou *et al.*, "EAST: An Efficient and Accurate Scene Text Detector," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2642-2651.

[6] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A Fast Text Detector with a Single Deep Neural Network," 2016.

[7] B. Shi, X. Bai, and S. Belongie, "Detecting Oriented Text in Natural Images by Linking Segments," pp. 3482-3490, 2017.

[8] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single Shot Text Detector with Regional Attention," pp. 3066-3074, 2017.

[9] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in *European Conference on Computer Vision*, 2016, pp. 21-37.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," pp. 770-778, 2015.

[11] C. Szegedy *et al.*, "Going deeper with convolutions," pp. 1-9, 2014.

[12] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : Deconvolutional Single Shot Detector," 2017.

[13] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 447-456.

[14] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading Text in the Wild with Convolutional Neural Networks," *International Journal of Computer Vision,* vol. 116, no. 1, pp. 1-20, 2014.

[15] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic Data for Text Localisation in Natural Images," pp. 2315-2324, 2016.

[16] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao, "Scene Text Detection via Holistic, Multi-Channel Prediction," 2016.

[17] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting Text in Natural Image with Connectionist Text Proposal Network," pp. 56-72, 2016.

[18] D. Karatzas *et al.*, "ICDAR 2013 Robust Reading Competition," in *International Conference on Document Analysis and Recognition*, 2013, pp. 1484-1493.

[19] D. Karatzas *et al.*, "ICDAR 2015 competition on Robust Reading," in *International Conference on Document Analysis and Recognition*, 2015, pp. 1156-1160.

[20] K. Wang and S. Belongie, "Word Spotting in the Wild," in *European Conference on Computer Vision*, 2010, pp. 591-604.