

A Text Localization Method Based on Weak Supervision

Jiyuan Zhang^{1,2}, Chen Du^{1,2}, Zipeng Feng^{1,2}, Yanna Wang¹, Chunheng Wang¹

¹*Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*

²*University of Chinese Academy of Sciences, Beijing 100049, China*

{zhangjiyuan2017, duchen2016, fengzipeng2017, wangyanna2013, chunheng.wang}@ia.ac.cn

Abstract—Recently, numerous deep learning based scene text detection methods have achieved promising performances in different text detecting tasks. Most of these methods are trained in a supervised way, which requires a large amount of annotated data. In this paper, we explore a weakly supervised method to locate text regions in scene images. We propose a fully convolutional network (FCN) architecture to implement binary classification. The training data we used do not need any text location annotation, we only need to divide the training data into two categories according to whether it contains text or not. We can obtain the text localization map (TLM) directly from the last convolutional layer. By setting a fixed threshold, the TLM is converted to a mask map. Then the connected component analysis and the text proposals method based on Maximally Stable Extremal Regions (MSERs) are used to get the text region bounding boxes. We conduct comprehensive experiments on standard text datasets. The results show that our text localization method achieves comparable recall performance with other methods and has more stable property.

Keywords—weak supervision; fully convolutional network; text localization map;

I. INTRODUCTION

With the development of the Internet and portable mobile devices, more and more scene images are created. The text contains rich semantic information and contributes to the analysis and understanding of scene images.

In the past years, there have been many deep neural network based methods proposed for scene text detection tasks. We require a large amount of annotated image data to train these neural networks. Annotating these images in bounding box level is a time-consuming and laborious task.

In this paper, we propose a text localization method based on weak supervision. We do not need any image annotation of the objects. First, we train a binary classification network, the network is a fully convolutional architecture. By merging multi-layer information together, it can capture more low-level features which are more suitable for text classification and localization. We use the $conv_{1 \times 1}$ to generate the text localization map (TLM). It is a two-dimension feature layer that represents the text location confidence score. Then, we use a fixed threshold and connected components analysis to convert the TLM to a binary mask. It works like the attention mechanism in the image. The MSERs and Single Linkage Criterion proposed in [1] is used to extract text region proposals out of the image.

This work is inspired by the research of Zhou et al. [2, 3] and Wei et al. [4, 5]. Li et al. [6] firstly use the weakly supervisory way to generate class activation maps (CAM) and use MSERs to obtain the bounding box of the text regions. They adapt the network architecture in [3] by substituting the global average pooling (GAP) layer with the spatial pyramid average pooling (SPP) layer. They generate the CAM in two steps: first, they train the text attention neural network to classify the images into text images or neural images. The features output from the last spatial pyramid pooling (SPP) layer are passed to softmax classifier and the corresponding weights are recorded. Then the CAM is calculated by weighting the feature maps with the recorded weights.

Comparing with the previous works, our work makes the following contributions:

- (1) The work is a further exploration of weak supervision methods in scene text detection.
- (2) The network we designed uses a U-Net shape architecture rather than the CAM architecture, it allows images with arbitrary scales as input. By merging low-level and high-level features it can capture more fine-grained features.
- (3) Without extra steps, we can directly obtain the TLM from the last convolutional layer.
- (4) Our proposed method is stable and has comparable recall performance with other supervised methods.

The rest of the paper is organized as follows. Section II introduces the related work in the field of scene text detection and weakly supervised methods. The details of the proposed method are presented in Section III. In Section IV, we compare our method with other methods on standard text datasets. Section V concludes the paper.

II. RELATED WORK

A. Traditional Natural Scene Text Detection Method

The traditional scene text detection methods can be divided into two categories: the methods based on connected components and the methods based on sliding windows.

The methods based on connected components use the bottom-up strategy to detect text. Such as edge detection methods and text-level detection methods. The edge detection methods obtain the text proposal region by detecting the edges or corners of text and then use a classifier to tell whether it belongs to text or not. The text-level detection

methods utilize the characteristics that the scene text pixels usually have similar color and stroke width, by processing specifically, the adjacent pixels show connectivity in their spatial structure. These methods obtain text proposals by detecting connected components in the image. Some of the representative methods are external regions (ERs) [7, 8], maximally stable extremal regions (MSER) [9, 10], stroke width transform (SWT) [11, 12] and binarized normed gradients (BING) [13].

the methods based on sliding windows use the top-down strategy [14–18] to detect texts. They adopt multiscale sliding windows to scan the whole image and extract features from the text candidate regions, the confidence scores of the text regions are obtained by combining a trained classifier.

B. Deep Learning Based Scene Text Detection Method

The performance of traditional text detection methods is limited to the handcrafted features. With the development of deep learning technology, the deep neural network can combine low-level features to obtain a more abstract high-level presentation of text regions. The deep learning based scene text detection methods can be divided into anchor-based methods and pixel-based methods.

The anchor-based methods are inspired by faster-RCNN [19] and the SSD [20] series framework. Zhong et al. [21] propose a unified framework for text proposal generation based on faster-RCNN. Tian et al. [22] use fixed-length anchors to get the text regions and send them to a Bi-LSTM network. By merging space features and sequence features, the network can obtain the text proposals. To handle the large variation in aspect ratios of words, Liao et al. [23] design several inception-style output layers that utilize both irregular convolutional kernels and default boxes. In the following work, they improved the previous work and proposed the Textboxes++ method [24], its main contribution is expanding the horizontal text detector into the arbitrary orientation text detector.

The pixel-based methods take the text detection task as a general segmentation task. they use the fully convolutional network structure to determine whether the pixel in the image belongs to the foreground(text) or background. The most representative method is EAST [25]. Considering the efficiency, the semantic segmentation methods usually predict the text/non-text score map on small feature maps. These methods can avoid the direction of text alignment and the effect of the aspect ratio variation. Li et al. [27] propose a novel end-to-end framework by combining semantic segmentation and anchor-based methods in one network to deal with the large variances in size and aspect ratio.

C. Weakly Supervised Detection Method

While using the deep learning based method to detect text in the scene images, the scale of the training set will have an important impact on the results. The small scale training set

will lead to the overfitting problem, the large scale training set will consume too much labor to annotate the data. Some researchers provide a valid solution [28–30] to handle the problem by synthesizing images.

There are some methods try to detect the object in images without data location annotation. Zhou et al. [2, 3] find that the convolutional neural network can localize the objects by training a classification network. They use the class activation maps (CAM) to indicate the object regions. But this method can only identify the most discriminative part of the object. Wei et al. [4] propose an adversarial erasing approach to obtain the whole object regions by training several classification networks and merge the results. Singh et al. [31] use a strategy to hide the patches of input images randomly so that it can find other discriminative parts of the object. Wei et al. [5] enhance their previous work by proposing an Adversarial Complementary Learning approach for discovering the entire object via weakly supervised training. Li et al. [6] adapt the CAM to generate localization maps in the text detection task, it allows multiscale inputs by substituting the global average pooling layer (GAP) with spatial pyramid average pooling (SPP) layer.

Unlike [3] and [6], our proposed method is very easy to implement and has good generalization ability and stability. The network architecture is more suitable for text features extraction and the TLM can be obtained in a more concise way.

III. THE PROPOSED METHOD

A. The Feature Extraction and Merging Branch

We design the classification network with the idea of semantic segmentation. As is shown in Figure 1. Considering the low-level feature is also important for text classification and detection tasks, we adopt the U-Net architecture. It is a kind of fully convolutional network architecture that can merges multiscale features together. Similar to the pipeline in EAST [25] algorithm, the whole model architecture can be decomposed into three parts: feature extraction branch, feature merging branch, and the TLM output branch.

In the feature extraction branch, we use VGG16 as the base feature extraction backbone network. The backbone network was pre-trained on ImageNet dataset. Begin from the second convolutional and pooling block, we denote the output feature maps as f_i , $i = 4, 3, 2, 1$, and the size of each feature map is $1/16$, $1/64$, $1/256$, $1/1024$ of the input image, respectively.

In the feature merging branch, we gradually merge them according to the following formulas:

$$g_i = \begin{cases} \text{upsample}(h_i) & \text{if } i \leq 3 \\ \text{conv}_{3 \times 3}(h_i) & \text{if } i = 4 \end{cases} \quad (1)$$

$$h_i = \begin{cases} f_i & \text{if } i = 1 \\ \text{conv}_{3 \times 3}(\text{conv}_{1 \times 1}([g_{i-1}; f_i])) & \text{otherwise} \end{cases} \quad (2)$$

where g_i is the upsampled results of input h_i in the first three blocks, after the upsampling, the size was doubled. Then we concatenate g_i and the corresponding f_i to make the full use of the multi-layer information. The $conv_{1 \times 1}$ is applied in the next layer to reduce computation and then followed by a $conv_{3 \times 3}$ that the features of the text can be extracted furthermore. In the final layer of the feature merging branch, we use $conv_{3 \times 3}$ to generate the output feature map. The output feature map size is 1/16 of the input image, and the channel is expanded to 32. The Batch Normalization is used after each concatenation and convolution layer, and the activation function in each convolution layer is ReLU.

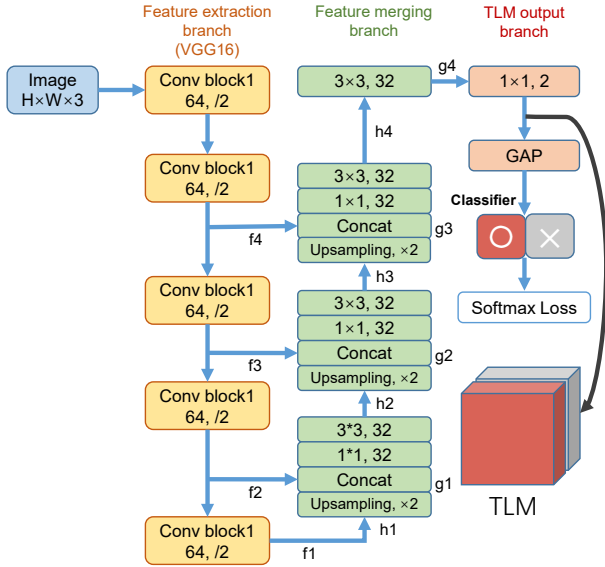


Figure 1. Architecture of classification network

B. The Classification Part and TLM Generation

Previous class activation map generation methods need an extra step to generate object localization maps after the forward pass. Inspired by the work [5], we use a $conv_{1 \times 1}$ connected to the output feature map in the feature merging branch. The convolutional kernel is set to 2 that corresponding to the foreground(text) and the background activation map respectively. So the text localization maps (TLM) can be directly obtained in the forward pass.

$$TLM = conv_{1 \times 1}(g_4) \quad (3)$$

The TLM is a two-channel feature map, and each channel represents the activation of each class (text/background). In the classification stage, we set the first category to image-contain-text and the second category to image-without-text, so the first channel in the TLM represents the text class activation. We extract the first channel and resize it to the input image size with the bilinear interpolation method.

Then the TLM is followed by a global average pooling layer (GAP) and a softmax activation function is used for classification. We use the categorical cross-entropy as the loss function.

Because of the fully convolutional architecture, the network allows images with arbitrary scales and aspect ratios as input. In shallow convolutional layers, the receptive field is small, it learns some local features such as the edges or the corners of text. In deep convolutional layers, the receptive field is large, it learns more high-level, abstract features.

The upsampling operation in Feature merging branch increases the resolution of the output feature map, so it can be concatenated with high-resolution features from the Feature extraction branch. The successive convolutional layer can then learn to assemble a more precise output based on this information.

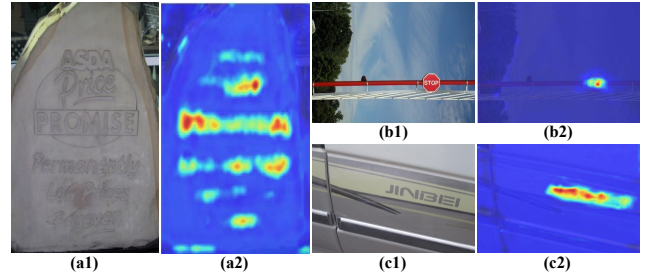


Figure 2. TLM heatmap in different situations: (a) The carved text (b) The tiny scale text (c) The tilted text

In Figure 2. We demonstrate some TLM results on IC-DAR2013 dataset in the form of the heatmap. We notice that the method can handle different situations and whatever the text scale changes, the TLM is always sensitive to the text region and ignore the background noise automatically.

C. Text Proposals Generation

According to the observation of the generated TLM, we find that the activation of text in the TLM is a distributed representation. Some parts of the text corners or edges have high activation value and the TLM is sensitive to the most discriminative part of the text. The threshold method is not enough to cover the whole text region. What is more, the text in scene images has large scale variance. We can hardly split the text line but to take them as a whole part in some dense-text images.

Combining all the above considerations, we firstly segment the foreground and background from TLM with a fixed threshold. Then, we use the connected components analysis to merge each text tiny patch to form a text region mask. The mask works like the attention mechanism to the image. Furthermore, the text extraction method presented in [1] was applied to generate bounding-box level text proposals, this method is based on MSERs and uses Single Linkage Criterion (SLC) to aggregate tiny text unit into text line

proposals. At last, we use the quicksort algorithm to rank the proposals according to the confidence value. The pipeline is shown in Figure 3.

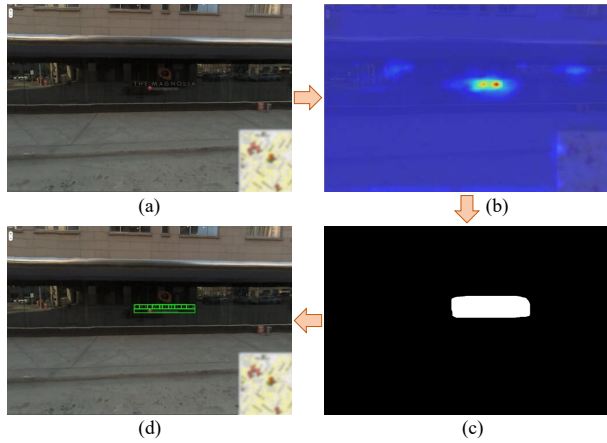


Figure 3. The TLM method pipeline: (a) The original scene image (b) The heatmap of TLM (c) The text mask generated from the TLM (d) Text proposals on text mask regions.

IV. EXPERIMENTS AND RESULTS

A. Datasets and Evaluation Metrics

The scene text dataset we collected are based on the following datasets: (1) The Street View Text (SVT) dataset [17], it contains 350 total street images from 20 different cities, 100 for training and 250 for testing. (2) ICDAR2013 focused scene text dataset [32], it contains 229 training sets and 223 test sets which were extracted from web pages and email messages. (3) MSRA Text Detection 500 Database (MSRA-TD500) [34], it contains 500 natural images, which are taken from indoor and outdoor scenes using a pocket camera.

We use a total of 900 images as the positive dataset. Some of them are extracted from the above datasets, except the 223 test sets in ICDAR2013 dataset and the 250 test sets in SVT datasets. Besides, we manually select images that contain text from PASCAL VOC2007 dataset as supplementary. On the other hand, we randomly selected 2500 images as negative datasets from the following datasets: 1) The PASCAL VOC2007 dataset [35]. 2) The Scene UNderstanding (SUN) database [36]. In particular, we manually clean the negative dataset so that there was no image contains text in it.

B. Training and evaluation Details

We implement our method with Python and Keras (Tensorflow backend). The VGG16 backbone network weights are initialized with the weights trained on ImageNet dataset. We train the network on NVIDIA GeForce 1080Ti with 11GB memory. Considering the memory size and training speed, we resize the input images to 512×512 pixels and set batch size to 8. We also use horizontal/vertical flip operation

to augment the data scale. The Adam optimizer is used for training and the base learning rate is set to $1e-3$.

We use the evaluation framework provided in ICDAR2015 robust reading competition. We compare the recall rate under a certain intersection over union (IoU) threshold with different supervised text detection methods.

C. Evaluation on ICDAR2013 dataset

Table I shows the comparison with other text detection methods on ICDAR2013 dataset. The WSTAN method is a weakly supervised method and the others are supervised methods. We can see that our method outperforms the BING, EdgeBoxes, and GOP in terms of recall rate. Our method has comparable performance with the INCEPTION-RPN and WSTAN methods but more stable. The performance of other methods is dropped quickly with the IoU threshold increasing.

Figure 4 shows the recall rate versus IoU threshold from 0.1 to 0.9 of our method on ICDAR2013 dataset. From the figure, we can see that our method has stable performance, which means the region proposals we extracted are close to the real text region. It is beneficial for the next recognition process.

In Figure 5, we demonstrate the heatmap of TLM versus the WSTAN method on ICDAR2013 dataset. It is obvious that our TLM has more fine-grained presentations in text regions.

The TLM has the pixel-level attention mechanism rather than the region-level attention mechanism [6], Considering from the perspective of traditional methods, it can be seen as a bottom-up text method, it obtains the text region proposal by the corners or the edges activation.

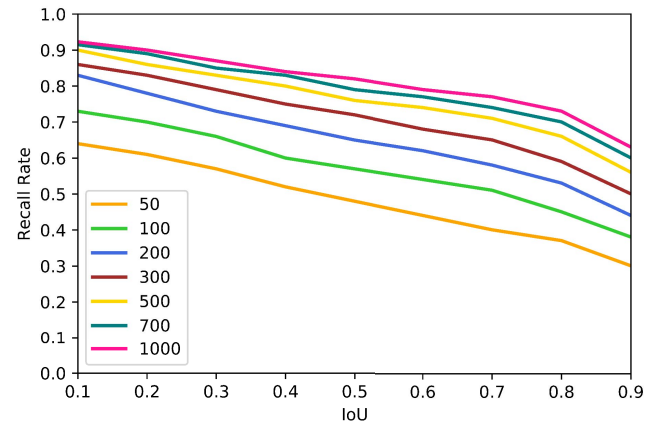


Figure 4. Recall versus IoU threshold of TLM method on ICDAR2013.

D. Evaluation on SVT dataset

We evaluate the performance of our TLM method on the SVT dataset. Table II shows the results compared with other methods. our TLM pipeline is better than most methods at

Table I
RECALL RATE AT DIFFERENT IOU THRESHOLDS ON ICDAR2013.

Method	Proposals	0.5 IoU	0.7 IoU	0.9 IoU
Ours (TLM)	1000	0.82	0.77	0.63
	500	0.76	0.71	0.56
	300	0.72	0.65	0.50
	100	0.57	0.51	0.38
WSTAN[6]	500	0.87	0.81	0.38
	300	0.80	0.74	0.35
	100	0.57	0.50	0.24
DeepText[21]	500	0.86	0.67	0.04
	300	0.88	0.66	0.04
	100	0.88	0.62	0.036
BING[13]	2716	0.63	0.08	0.00
EdgeBoxes[18]	9554	0.85	0.53	0.08
GOP[37]	855	0.45	0.18	0.08
MSERs group[1]	8164	0.98	0.96	0.79

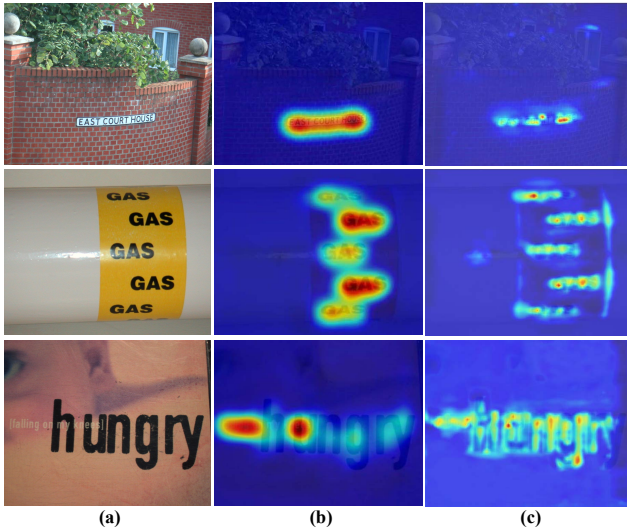


Figure 5. Column (a) shows the original test images on ICDAR2013. Column (b) shows the WSTAN heatmaps. Column (c) are our TLM heatmaps.

0.5 and 0.7 IoU. Due to using the MSERs group[1] in the post-process stage, the method is limit to the performance of MSERs-group, so that our stability properties are not reflected.

In Figure 6, we show some of the heatmaps of TLM on SVT dataset. Comparing the ICDAR2013 focused scene text dataset, SVT contains more challenging text, with high variability and low resolution. Our TLM method still works well on the dataset.

V. CONCLUSION

In this paper, we propose a fully convolutional network to locate the text region in natural scene images in a weakly supervised manner. The training data we used do not need any annotation about the text location. The text localization map can be directly obtained from the last convolutional layer. Extensive experiments show our method

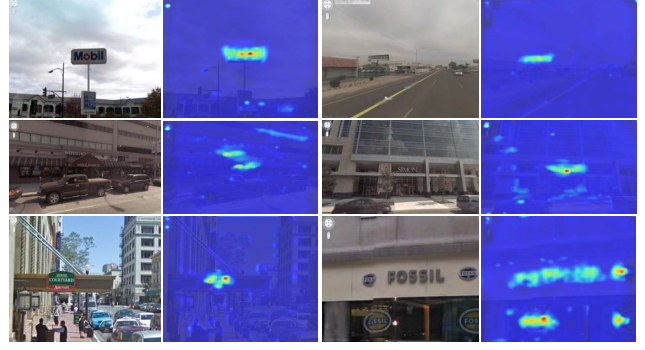


Figure 6. Our TLM heatmaps on the SVT dataset.

Table II
RECALL RATE AT DIFFERENT IOU THRESHOLDS ON SVT DATASET.

Method	Proposals	0.5 IoU	0.7 IoU	0.9 IoU
Ours (TLM)	1000	0.80	0.48	0.04
	500	0.75	0.41	0.02
	300	0.69	0.37	0.02
	100	0.58	0.26	0.01
BING[13]	2987	0.64	0.09	0.00
EdgeBoxes[18]	15319	0.94	0.63	0.04
GOP[37]	778	0.53	0.19	0.03
MSERs group[1]	10365	0.95	0.61	0.06

can mine fine-grained text information. It has comparable recall performance with other methods and higher stability.

ACKNOWLEDGMENT

This work is supported by the Key Programs of the Chinese Academy of Sciences under Grant No.ZDBS-SSW-JSC003, No.ZDBS-SSW-JSC004 and No.ZDBS-SSW-JSC005, and the National Natural Science Foundation of China (NSFC) under Grant No.61601462, No.61531019 and No.71621002.

REFERENCES

- [1] L. Gómez and D. Karatzas, “Object proposals for text extraction in the wild,” in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, Aug 2015, pp. 206–210.
- [2] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, “Object detectors emerge in deep scene cnns,” *CoRR*, vol. abs/1412.6856, 2015.
- [3] B. Zhou, A. Khosla, g. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” 12 2015.
- [4] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, “Object region mining with adversarial erasing: A simple classification to semantic segmentation approach,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6488–6496, 2017.
- [5] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, “Adversarial complementary learning for weakly supervised object localization,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1325–1334, 2018.

- [6] L. Rong, E. MengYi, L. JianQiang, and Z. HaiBin, "Weakly supervised text attention network for generating text proposals in scene images," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, Nov 2017, pp. 324–330.
- [7] L. Sun, Q. Huo, W. Jia, and K. Chen, "Robust text detection in natural scene images by generalized color-enhanced contrasting extremal region and neural networks," in *2014 22nd International Conference on Pattern Recognition*, Aug 2014, pp. 2715–2720.
- [8] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 3538–3545.
- [9] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Computer Vision – ACCV 2010*, R. Kimmel, R. Klette, and A. Sugimoto, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 770–783.
- [10] A. González, L. M. Bergasa, J. J. Yebes, and S. Bronte, "Text location in complex images," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Nov 2012, pp. 617–620.
- [11] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 2963–2970.
- [12] G. Zhou, Y. Liu, Z. Tian, and Y. Su, "A new hybrid method to detect text in natural scene," in *2011 18th IEEE International Conference on Image Processing*, Sep. 2011, pp. 2605–2608.
- [13] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [14] K. Wang and S. J. Belongie, "Word spotting in the wild," in *ECCV*, 2010.
- [15] S. M. Hanif, L. Prevost, and P. A. Negri, "A cascade detector for text detection in natural scene images," in *2008 19th International Conference on Pattern Recognition*, Dec 2008, pp. 1–4.
- [16] A. Mishra, K. Alahari, and C. V. Jawahar, "Top-down and bottom-up cues for scene text recognition," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 2687–2694.
- [17] K. Wang, B. Babenko, and S. J. Belongie, "End-to-end scene text recognition," *2011 International Conference on Computer Vision*, pp. 1457–1464, 2011.
- [18] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 391–405.
- [19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, 06 2015.
- [20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *ECCV*, 2016.
- [21] Z. Zhong, L. Jin, S. Zhang, and Z. Feng, "Deeptext: A unified framework for text proposal generation and text detection in natural images," *CoRR*, vol. abs/1605.07314, 2016.
- [22] Z. Tian, W. Huang, H. Tong, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," vol. 9912, 10 2016, pp. 56–72.
- [23] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," in *AAAI*, 2016.
- [24] M. Liao, B. Shi, and X. Bai, "Textboxes++: A single-shot oriented scene text detector," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3676–3690, Aug 2018.
- [25] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: An efficient and accurate scene text detector," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2642–2651, 2017.
- [26] D. Deng, H. Liu, X. Li, and D. Cai, "Pixellink: Detecting scene text via instance segmentation," 01 2018.
- [27] Y. Li, Y. Yu, Z. Li, Y. Lin, M. Xu, J. Li, and X. Zhou, "Pixel-anchor: A fast oriented scene text detector with combined networks," *CoRR*, vol. abs/1811.07432, 2018.
- [28] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," 06 2014.
- [29] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [30] 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, 12 2014.
- [31] K. Kumar Singh and Y. Jae Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," 04 2017.
- [32] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. M. Romeu, D. F. Mota, J. Almazán, and L.-P. de las Heras, "Icdar 2013 robust reading competition," *2013 12th International Conference on Document Analysis and Recognition*, pp. 1484–1493, 2013.
- [33] R. Nagy, A. Dicker, and K. Meyer-Wegener, "Neocr: A configurable dataset for natural image text recognition," in *Camera-Based Document Analysis and Recognition*, M. Iwamura and F. Shafait, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 150–163.
- [34] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 1083–1090.
- [35] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>.
- [36] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 3485–3492.
- [37] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 725–739.
- [38] T. He, W. Huang, Y. Qiao, and J. Yao, "Text-attentional convolutional neural network for scene text detection," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2529–2541, June 2016.