Fused Text Segmentation Networks for Multi-oriented Scene Text Detection

Yuchen Dai*, Zheng Huang†, Yuting Gao, Kai Chen
Shanghai Jiao Tong University
Email: xinghedyc@sjtu.edu.cn

Abstract—**In this paper, we introduce a novel end-end framework for multi-oriented scene text detection from an instance-aware segmentation perspective. We present Fused Text Segmentation Networks, which combine multi-level features during the feature extracting as text instance may rely on finer feature expression compared to general objects. It detects and segments the text instance jointly and simultaneously, leveraging merits from both semantic segmentation task and region proposal based object detection task. Not involving any extra pipelines, our approach surpasses the current state of the art on multi-oriented scene text detection benchmarks: ICDAR2015 Incidental Scene Text and MSRA-TD500 reaching Hmean 84.1% and Hmean 82.0% respectively which suggests effectiveness of the proposed approach.**

I. INTRODUCTION

Recently, scene text detection has drawn great attention from computer vision and machine learning community. Driven by many content-based image applications such as photo translation and receipt content recognition, it has become a promising and challenging research area both in academia and industry. Detecting text in natural images is difficult, because both text and background may be complex in the wild and it often suffers from disturbance such as occlusion and uncontrollable lighting conditions[47].

Previous text detection methods[4], [6], [1], [39], [15] have achieved promising results on several benchmarks. The essential problem in text detection is to represent text region using discriminative features. Conventionally, hand-crafted features are designed[6], [30], [32], [42] to capture the properties of text region such as texture and shape, while in the past few years, deep learning based approaches[14], [16], [15], [7], [44], [22] directly learn hierarchical features from training data, demonstrating more accurate and efficient performance in various benchmarks such as ICDAR series contests[35], [18], [17].

Existing methods[16], [14], [15], [22] have obtained decent performance for detecting horizontal or near-horizontal text. While horizontal text detection has constraints of axis-aligned bounding-box ground truth, the multi-oriented text is not restrictive to a particular orientation and usually uses quadrilaterals for annotations. Due to this unconstrained condition, it reports relatively lower accuracies in ICDAR 2015 Competition Challenge 4 Incidental scene text localization[17] compared to horizontal scene text detection benchmarks[35], [18].

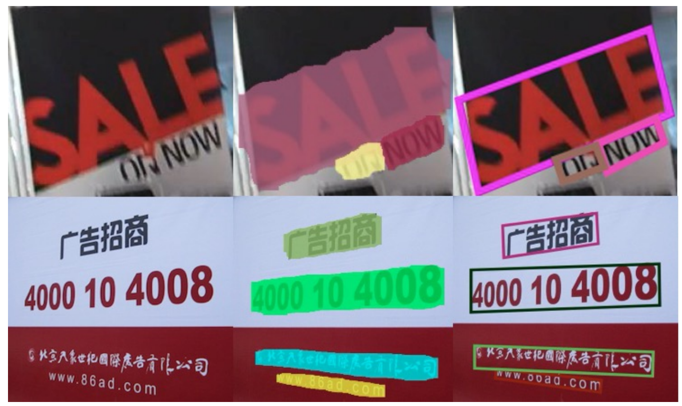

Fig. 1. **FTSN Work Flows**. From left to right, input images, text instance segmentation results and final processed quadrilateral results are shown in figure.

Recently, a few approaches[33], [12], [36], [46], [25], [27] have been proposed to address the multi-oriented text detection. In general, there are currently four different types of methods. Region based methods[36], [27], [25] leverage advanced object detection techniques such as Faster RCNN[34], R-FCN[20] and SSD[24], and segmentation-based methods[45], [11] mainly utilize fully convolutional neural networks (FCN) for generating text score maps, which often need several stages and components to achieve final detections. Direct regression based method[12] regresses the position and size of an object from a given point. Finally, hybrid method[46] combines text scores map and rotated/quadrangle bounding boxes generation to collaboratively obtain the efficient and accurate performance of multi-oriented text detection.

Inspired by recent advance of instance-aware segmentation[21], [9], we present a novel perspective to handle the task of multi-oriented text detection. In this work, we leverage the merits from accurate region proposal based methods[34], [20], and flexible segmentation based methods which can easily generate arbitrary-shaped text mask[45], [11] . It is an end-to-end trainable framework excluding redundant and low-efficient pipelines such as the use of text/nontext salient map[45] and text-line generation[11]. Based on region proposal network (RPN), our approach detects and segments text instance simultaneously, followed by non-maximum suppression (NMS) to suppress overlapping instances. Finally,

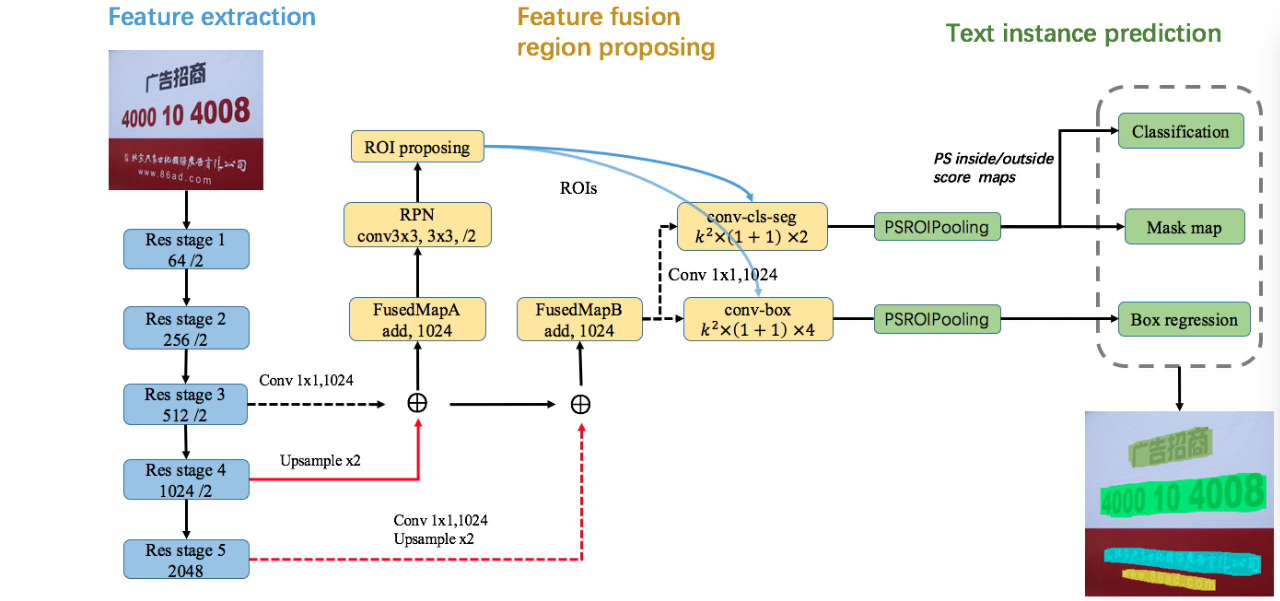

Fig. 2. **The proposed framework** consists of three parts: feature extraction, feature fusion along with region proposing and text instance prediction. The dashed line represents convolution with 1x1 kernel size and 1024 output channels. The line in red is for upsampling operation and blue lines indicate on which feature maps PSROIPooling are performed using ROIs.

a minimum quadrangle bounding box to fit each instance area is generated as the result of the whole detection process.

Our main contributions are summarized as follows:

- We present an end-end efficient and trainable solution for multi-oriented text detection from an instance aware segmentation perspective, excluding any redundant pipelines.
- During feature extraction, feature maps are composed in a fused fashion to adaptively satisfy the finer representation of text instance.
- Mask-NMS is introduced to improve the standard NMS when facing heavily inclined or line-level text instances.
- Without many bells and whistles, our approach outperforms state of the art on current multi-oriented text detection benchmarks.

II. Related work

A. Scene Text Detection

Detecting text in natural images has been widely studied in past few years, motivated by many text-related real-world applications such as photo OCR and blind navigation. One of the mainstream traditional methods for scene text detection are Connected Components (CCs) based methods[31], [40], [16], [13], [37], [29] , which consider text as a group of individual components such as characters. Within these methods, stroke width transform (SWT)[6], [13] and maximally stable extremal region (MSER)[28], [29], [32] are usually used to seek character candidates, followed by several steps to filter no-text areas. Finally, these candidates are combined to obtain text objects. Although these bottom up approaches may be accurate on some benchmarks[35], [18], they often suffer from too many pipelines, which may cause inefficiency. Another mainstream traditional methods are sliding window based[19], [4], [8], [16]. These methods often use a fixed-size or multi-scale window to slide through the image searching the region which most likely contains text. However, the process of sliding window may involve large computational cost which results in inefficiency. Generally, traditional methods often require several steps to obtain final detections, and hand-designed features are usually used to represent properties of text. Therefore, they may suffer from inefficiency and low generalization ability against complex situations such as non-uniform illumination[43].

Recent progress on deep learning based approaches for object detection and semantic segmentation has provided new techniques for reading text in the wild. Scene text detection can be also seen as an instance of general object detection. Driven by the advance of object detection frameworks such as Faster RCNN[34] and SSD[24], these methods achieved state of the art by either using a region proposal network to first classify some text region proposals[27], [33], or directly regress text bounding boxes coordinates from a set of default boxes[22], [36]. These methods are able to achieve leading performance on horizontal or multi-oriented scene text detection benchmarks. However, they may also be restricted to rectangular bounding box constraints even with appropriate rotation[25]. Different from these methods, FCN based approaches generate text/non-text map which classifies text at the pixel level[45]. Though it may be suited well for arbitrary shape of text in natural images, it often involves several pipelines which leads to inefficiency[45], [33].

Inspired by recent advance on instance-aware segmentation[21], [9], we present an end-end trainable framework called Fused Text Segmentation Networks (FTSN) to handle arbitrary-shape text detection with no extra pipelines involved. It inherits merits from both object detection and semantic segmentation architecture to be efficient which detects and segments an text instance simultaneously and accurate which gives predictions in the pixel level. As text may rely on finer feature representation, a fused structure formed by multi-level feature maps is set to fit this property.

III. Methods

The proposed framework for multi-oriented scene text detection is diagrammed in Fig.2. It is a deep CNN model which mainly consists of three parts. Feature representations of each image are extracted through resnet-101 backbone[10], then multi-level feature maps are fused as FusedMapA which is fed to the region proposed network (RPN) for text region of interest (ROI) generation and FusedMapB for later rois' PSROIPooling. Finally the rois also known as text instances are sent to the detection, segmentation and box regression branches to output text instances in pixel level along with their corresponding bounding boxes. The post-processing part includes NMS and minimal quadrilateral generation.

A. Network Architecture

The convolutional feature representation is designed in a fusion fashion. The text instance is not like the general object such as people and cars which have relatively strong semantics. On the contrary, texts often vary tremendously in intra-class geometries. Consequently, low-level features should be taken into consideration. Basically, resnet-101 consists of five stages. Before region proposing, stage3 and upsampled stage4 feature maps are combined to form FusedMapA through element-wise adding, then upsampled feature maps from stage5 are fused with FusedMapA to form FusedMapB. It is noted that downsampling is not involved during stage5. Instead, we use the hole algorithm[2], [26] to keep the feature stride and maintain the receptive field. The reason for this is that both text properties and the segmentation task may require finer features and involving final downsampling may lose some useful information.

Because using feature stride of stage3 may cause millions of anchors in original RPN[34] which makes model training hard, so we add a 3×3 with stride 2 convolution to reduce such huge number of anchors.

Followed FCIS[21], we use Joint Mask Prediction and Classification to simultaneously classify and mask the text instance on $2 \times (1 + 1)$ inside/outside score maps generated through PSROIPooling on conv-cls-seg feature maps, and box regression branch utilizes $4 \times (1 + 1)$ feature maps from conv-box after PSROIPooling ("$1 + 1$" means one class is for text and the other for background). we use $k = 7$ shown in Fig.2 in our experiments by default. It is noted that after PSROIPooling, the resolution of feature maps becomes 21×21. Therefore, we use global average pooling[23] for classification (after pixel-wise max) and box regression branches, and pixel-wise softmax on mask branch.

Fig. 3. **Ground truth.** Left: original image. Right: corresponding ground truth in which dashed lines are for bounding boxes and quadrilaterals filled with different colors are for masks

B. Ground Truth and Loss Function

The whole multi-task loss $\mathcal{L}$ can be interpreted as

$$\mathcal{L} = \mathcal{L}_{rpn} + \mathcal{L}_{ins} \tag{1}$$

$$\mathcal{L}_{rpn} = \mathcal{L}_{rcls} + \lambda_r \mathcal{L}_{rbox} \tag{2}$$

$$\mathcal{L}_{ins} = \mathcal{L}_{cls} + \lambda_m \mathcal{L}_{mask} + \lambda_b \mathcal{L}_{box} \tag{3}$$

The full loss $\mathcal{L}$ consists of two sub stage losses: RPN loss $\mathcal{L}_{rpn}$ where $\mathcal{L}_{rcls}$ is for region proposal classification and $\mathcal{L}_{rbox}$ is for box regression, and text instance loss $\mathcal{L}_{ins}$ based on each ROI, where $\mathcal{L}_{cls}, \mathcal{L}_{mask} and \mathcal{L}_{box}$ represent losses for instance classification, mask and box regression task respectively. λ is the hyper-parameter to control the balance among each loss term. They are set as $\lambda_r = 0.2, \lambda_m = 2, \lambda_b = 0.2$ in our experiments.

Classification and mask task both use cross-entropy as loss function, whereas we use smooth-L1 for box regression task formulated as

$$\text{smooth}_{L_1}(x) = \begin{cases} (\sigma x)^2 & \text{if } |x| < 1/\sigma^2, \\ |x| - 0.5/\sigma^2 & \text{otherwise.} \end{cases} \tag{4}$$

σ is set to 3 in our experiments which makes the box regression loss less sensitive to outliers.

Ground truth of each text instance is presented by bounding boxes and masks shown in Fig.3. In most multi-oriented text detection dataset, annotations are given in quadrilaterals such as IC15 or can be converted to quadrilaterals such as TD500. For each instance, we directly generate mask from quadrilateral coordinates and use the minimal rectangle containing the mask as the bounding box.

C. Post Processing

Mask-NMS To obtain final detection results, we use Non-Maximum Suppression mechanism (NMS) to filter overlapped text instances and preserve those with highest scores. After NMS, we generate a minimum quadrilateral for each text instance covering the mask as shown in Fig.1.

Standard NMS computes IOU among bounding boxes, which may be fine for word-level and near-horizontal results' filtering. However, it may filter some correct line-level detections when they are close and heavily inclined as shown in

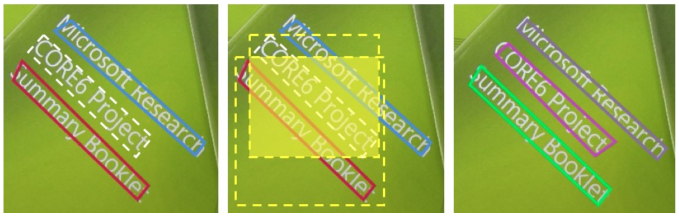

Fig. 4. **Heavily inclined text filtering results.** Left: standard NMS may filter the correct detection as shown in white dashed line. Middle: dashed lines in yellow show the bounding boxes used in IOU computation of standard NMS, and the yellow transparent area illustrates that the IOU between the two closed inclined instance is big enough to make the correct detection filtered. Right: Mask-NMS takes the mask area of each instance to compute MMI replacing IOU. Therefore, there is no intersection between each instance. No correct detections are fitered.

Fig.4 or when words stay close in the same line as shown in Fig.5. Consequently, we propose a modified NMS called Mask-NMS to handle such situations. Mask-NMS mainly changes bounding box IOU computation to so-called mask-maximum-intersection (MMI) as formulated:

$$MMI = max(I/I_A, I/I_B) \tag{5}$$

I_A, I_B are mask areas of two text instances to be computed, I is the intersection area between the masks. Maximum intersection over the mask areas are used to replace original IOU for the reason that detections may easily involve line-level and word-level text instances simultaneously at the same line as shown in Fig.5. The proposed Mask-NMS has significantly improved performance for multi-oriented scene text detection as shown in section.5.

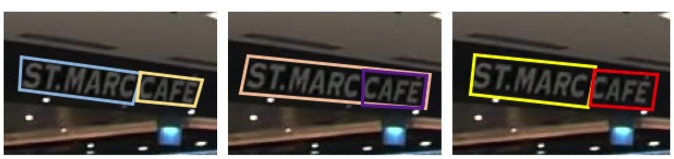

Fig. 5. **line-level text filtering results**. From left to right: ground truth, standard NMS and Mask-NMS results. Benchmarks may provide annotations at different levels such as word-level in IC15 and line-level in TD500. However, model may be confused about them and make predictions at different levels as shown in the middle. Using maximum intersection can greatly avoid this situation as the MMI over the two text instances in the middle image is 1 so one of them is certainly filtered.

IV. Experiments

To evaluate the proposed framework, we conduct quantitative experiments on two public benchmarks: ICDAR2015 and MSRA-TD500.

A. Datasets

ICDAR 2015 Incidental Text (IC15) the Challenge 4 of ICDAR 2015 Robust Reading Competition[17]. IC15 contains 1000 training and 500 testing incidental images taken by Google Glasses without paying attention to viewpoint and image quality. Therefore, large variations in text scale, orientation and resolution lead to difficulty for text detection. Annotations of the dataset are given in word-level quadrilaterals.

MSRA-TD500 (TD500) is early presented in [41]. The dataset is multi-oriented and multi-lingual including both Chinese and English text which consists of 300 training and 200 testing images. Different from IC15, annotations of TD500 are at line level which are rotated rectangles.

SynthText in the Wild (SynthText) The dataset contains 800,000 synthetic images[7], text with random color, fonts, scale and orientation are rendered on natural images carefully to have a realistic look. Annotations are given in character, word and line level. We only pick up 160,000 images for model's pretraining.

B. Implementation Details

Training We pretrain the proposed FSTN on a subset of SynthText containing 160,000 images, then finetune on IC15 and TD500. For optimization, standard SGD is used during training with learning rate 5×10^{-3} for first 5 epochs and 5×10^{-4} for the last epoch,and we also apply online hard example mining (OHEM)[38] for balancing the positive and negative samples. Different from original RPN anchor ratios and scales setting for object detection, anchor scales of $[32^2, 64^2, 128^2, 256^2]$ and ratios of [1/3,1/2,1,2,3,5,7] are set because text often has a large aspect ratio and a small scale[22].

Data augmentation Multi-scale training, rotation and color jittering are applied during training. Scales are randomly chosen from [600,720,960,1100] and each number represents the short edge of input images. Rotation with $15°$,$30°$ and $45°$ are applied with horizontal flip. Consequently, it enlarges 8x dataset size than the original one. Random brightness, contrast and saturation jittering are applied for input images.

Testing Input images are resized to 848×1500 when testing. After NMS, mask voting[5] is used to obtain an ensemble text instance mask by averaging all reasonable detections.

Experiments are conducted on MXNet[3] and run on a server with Intel i7 6700K CPU, 64GB RAM, GTX 1080 and Ubuntu 14.04 OS.

TABLE I
ICDAR 2015 INCIDENTAL DATASET

Method	Precision (%)	Recall (%)	Hmean (%)
HUST[17]	44.0	37.8	40.7
Zhang *et al.*[45]	71.0	43.0	54.0
DMPNet[25]	73.2	68.2	70.6
Qin *et al.*[33]	79.0	65.0	71.0
SegLink[36]	73.1	76.8	75.0
RRPN[27]	73.2	82.2	77.4
EAST[46]	83.3	78.3	80.7
He *et al.*[12]	82.0	80.0	81.0
Proposed FSTN+SNMS	87.1	80.0	83.4
Proposed FSTN+MNMS	**88.6**	80.0	**84.1**

C. Results

Tabel.1 shows results of the proposed FSTN on IC15 compared with previous state of art published methods. SNMS and

Fig. 6. Results of ICDAR2015 and MSRA-TD500

TABLE II
MSRA-TD500 DATASET

Method	Precision (%)	Recall (%)	Hmean (%)
Yao *et al.*[41]	63.0	63.0	60.0
Zhang *et al.*[45]	83.0	67.0	74.0
RRPN[27]	82.0	68.0	74.0
He *et al.*[12]	77.0	70.0	74.0
EAST[46]	87.3	67.4	76.1
SegLink[36]	86.0	70.0	77.0
Proposed FSTN+SNMS	86.6	**77.3**	81.7
Proposed FSTN+MNMS	**87.6**	77.1	**82.0**

MNMS represent standard NMS and Mask-NMS respectively. Our FSTN with Mask-NMS outperforms former best result by 5.3% in Precision and 3.1% in Hmean. It is evaluated by the official submission server[1].

Results on TD500 are shown in Table.2 along with other state of art methods. It is shown that our methods outperform the current state of art approaches by a large margin in Hmean and Recall, without adding extra real-world training images.

Outperforming the current state of the art, our approach runs about 4 FPS on 848×1500 images, which presents efficiency and accuracy.

It is noted that the proposed Mask-NMS significantly improved Hmean by 0.7 and 0.3 percent on IC15 and TD500, which mainly target the situations in Fig.4 and Fig.5.

Fig.6 shows example results of FSTN. Upper row illustrates results on IC15 and lower row exhibits TD500's results. The decent performance for both word-level and line-level text detection with large variation in resolution, view point, scale and linguistics suggests excellent generalization ability.

[1] http://rrc.cvc.uab.es/?ch=4

D. Conclusion

we present FSTN, an end-end efficient and accurate multi-oriented scene text detection framework. It has outperformed previous state of the art approaches on word-level or line-level annotated benchmarks demonstrating decent generalization ability and flexibility. In the future, we will explore the potentials to extend FSTN to an end-end recognition architecture and make it a more efficient solution towards real-time applications.

REFERENCES

[1] Michal Buta, Luka Neumann, and Jiri Matas. Fastext: Efficient unconstrained scene text detector. In *IEEE International Conference on Computer Vision*, pages 1206–1214, 2015.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[3] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

[4] Xiangrong Chen and Alan L. Yuille. Detecting and reading text in natural scenes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 366–373, 2004.

[5] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016.

[6] Boris Epshtein, Eyal Ofek, and Yonatan Wexler. Detecting text in natural scenes with stroke width transform. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2963–2970. IEEE, 2010.

[7] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. pages 2315–2324, 2016.

[8] Shehzad Muhammad Hanif and Lionel Prevost. Text detection and localization in complex scene images using constrained adaboost algorithm. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 1–5. IEEE, 2009.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *arXiv preprint arXiv:1703.06870*, 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Tong He, Weilin Huang, Yu Qiao, and Jian Yao. Accurate text localization in natural image with cascaded convolutional text network. *arXiv preprint arXiv:1603.09423*, 2016.

[12] Wenhao He, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Deep direct regression for multi-oriented scene text detection. *arXiv preprint arXiv:1703.08289*, 2017.

[13] Weilin Huang, Zhe Lin, Jianchao Yang, and Jue Wang. Text localization in natural images using stroke feature transform and text covariance descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1241–1248, 2013.

[14] Weilin Huang, Yu Qiao, and Xiaoou Tang. *Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees*. Springer International Publishing, 2014.

[15] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

[16] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Deep features for text spotting. In *European Conference on Computer Vision*, pages 512–528, 2014.

[17] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, and Faisal Shafait. Icdar 2015 competition on robust reading. In *International Conference on Document Analysis and Recognition*, pages 1156–1160, 2015.

[18] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez I Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazn Almazn, and Llus Pere De Las Heras. Icdar 2013 robust reading competition. In *International Conference on Document Analysis and Recognition*, pages 1484–1493, 2013.

[19] Kwang In Kim, Keechul Jung, and Jin Hyung Kim. Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1631–1639, 2003.

[20] Yi Li, Kaiming He, Jian Sun, et al. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, pages 379–387, 2016.

[21] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[22] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *Association for the Advancement of Artificial Intelligence*, 2017.

[23] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[25] Yuliang Liu and Lianwen Jin. Deep matching prior network: Toward tighter multi-oriented text detection. *arXiv preprint arXiv:1703.01425*, 2017.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[27] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *arXiv preprint arXiv:1703.01086*, 2017.

[28] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image & Vision Computing*, 22(10):761–767, 2004.

[29] Lukas Neumann and Jiri Matas. A method for text localization and recognition in real-world images. In *Asian Conference on Computer Vision*, pages 770–783. Springer, 2010.

[30] Lukas Neumann and Jiri Matas. A method for text localization and recognition in real-world images. In *Computer Vision - ACCV 2010 - Asian Conference on Computer Vision, Queenstown, New Zealand, November 8-12, 2010, Revised Selected Papers*, pages 770–783, 2011.

[31] Lukáš Neumann and Jiří Matas. Real-time lexicon-free scene text localization and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1872–1885, 2016.

[32] Luk Neumann and Ji Matas. Real-time scene text localization and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3538–3545, 2012.

[33] Siyang Qin and Roberto Manduchi. Cascaded segmentation-detection networks for word-level text spotting. 2017.

[34] 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 & Machine Intelligence*, 39(6):1137, 2017.

[35] Asif Shahab, Faisal Shafait, and Andreas Dengel. Icdar 2011 robust reading competition challenge 2: Reading text in scene images. In *International Conference on Document Analysis and Recognition*, pages 1491–1496, 2011.

[36] Baoguang Shi, Xiang Bai, and Serge Belongie. Detecting oriented text in natural images by linking segments. In *Computer Vision and Pattern Recognition*, 2017.

[37] Cunzhao Shi, Chunheng Wang, Baihua Xiao, Yang Zhang, Song Gao, and Zhong Zhang. Scene text recognition using part-based tree-structured character detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2968, 2013.

[38] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.

[39] Shangxuan Tian, Yifeng Pan, Chang Huang, Shijian Lu, Kai Yu, and Chew Lim Tan. Text flow: A unified text detection system in natural scene images. In *IEEE International Conference on Computer Vision*, pages 4651–4659, 2016.

[40] Kai Wang and Serge Belongie. Word spotting in the wild. In *European Conference on Computer Vision*, pages 591–604. Springer, 2010.

[41] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1083–1090. IEEE, 2012.

[42] Alessandro Zamberletti, Lucia Noce, and Ignazio Gallo. *Text Localization Based on Fast Feature Pyramids and Multi-Resolution Maximally Stable Extremal Regions*. Springer International Publishing, 2014.

[43] Shuye Zhang, Mude Lin, Tianshui Chen, Lianwen Jin, and Liang Lin. Character proposal network for robust text extraction. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 2633–2637. IEEE, 2016.

[44] Zheng Zhang, Chengquan Zhang, Wei Shen, Cong Yao, Wenyu Liu, and Xiang Bai. Multi-oriented text detection with fully convolutional networks. In *Computer Vision and Pattern Recognition*, 2016.

[45] Zheng Zhang, Chengquan Zhang, Wei Shen, Cong Yao, Wenyu Liu, and Xiang Bai. Multi-oriented text detection with fully convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4159–4167, 2016.

[46] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: An efficient and accurate scene text detector. In *Computer Vision and Pattern Recognition*, 2017.

[47] Yingying Zhu, Cong Yao, and Xiang Bai. Scene text detection and recognition: recent advances and future trends. *Frontiers of Computer Science*, 10(1):19–36, 2016.