TOWARDS ACCURATE INSTANCE-LEVEL TEXT SPOTTING WITH GUIDED ATTENTION

Haiyan Wang, Xuejian Rong, Yingli Tian

Department of Electrical Engineering,
The City College of New York, New York, NY 10031
hwang005@citymail.cuny.edu, {xrong, ytian}@ccny.cuny.edu

ABSTRACT

We tackle the text detection problem from the instance-aware segmentation perspective, in which text bounding boxes are directly extracted from segmentation results without location regression. Specifically, a text-specific attention model and a global enhancement block are introduced to enrich the semantics of text detection features. The attention model is trained with a weakly segmentation supervision signal and enforces the detector to focus on the text regions, while also suppressing the influence of neighboring background clutters. In conjunction with the attention model, a global enhancement block (GEB) is adapted to reason the relationship among different channels with channel-wise weights calibration. Our method achieves comparable performance with the recent state-of-the-arts on *ICDAR2013*, *ICDAR2015*, and *ICDAR2017-MLT* benchmark datasets.

Index Terms— Text detection, instance segmentation, attention, richer feature representation

1. INTRODUCTION

Scene text broadly exists in our daily life. It appears almost everywhere such as supermarket labels and traffic signs. In the recent decade, scene text detection becomes increasingly crucial in computer vision tasks such as image retrieval [1] and autonomous driving [2]. However, due to the large variance of aspect ratio, scale, and illumination, especially for the multi-oriented text regions, scene text detection is actually one of the most challenging tasks in computer vision fields. Because of these inevitable challenges and complexities, traditional methods [3, 4, 5] usually tend to first detect individual characters or parts of the text such as extracting extreme regions and then group them by exhaustive search methods.

In recent years, deep learning-based methods have been popular to solve the object detection problem. The most popular methods are based on proposal and multi-stages [6, 7, 8]. Through first extracting region-of-interest (ROI), and then performing the bounding box regression, the network could perform reasonably well on detecting text. However, certain problems still arise from these approaches. Early methods usually stem from the Faster-RCNN [7] model, such as [9, 10, 11, 12]. Ma at al. [10] proposed a method to detect texts in arbitrary orientations by injecting the angle information of the anchor bounding boxes. After the ROI pooling, regressions for both bounding box and the angle were conducted for prediction. Jiang et al. proposed to predict the axis-aligned bounding boxes and the inclined minimum area boxes together for multi-oriented text detection [11]. Instead of using the inclined anchor boxes to predict the angle information, they added one more output branch to predict the oriented bounding boxes and also introduce the inclined non-maximum suppression to filter the highest confidence score of the oriented results. However, due to the two-stage design with each ROI being independent, the high computational overhead is inevitable. The network has to do the classification and bounding boxes regression for each ROI, which introduces too many parameters to compute. Then, with the R-FCN [13] proposed, the position sensitive score map is designed to solve this problem. Different ROIs could share weights through position sensitive score map and significantly accelerate the computation speed. However, due to the fully-convolution design, the network loses the global information because there are no fully-connected layers within the network. In this way, the network could not effectively capture the global and context information which is crucial to the detection performance.

In parallel, following the design of the one-shot object detectors such as SSD [14] and YOLO [15], there are some other methods like textboxes [16], textboxes++[17] and seglink [18] that perform text detection in the one-stage manner. The text classification and detection are performed densely without objectness-based pruning. Because of sharing weights through all of the bounding boxes, the network could highly improve the computation efficiency. And meanwhile, they also apply the multiple convolution layers to detect text with different size and aspect ratios. Therefore, the semantic information is learned in a hierarchical manner. Smaller and larger texts are detected by the lower layers and higher layers respectively. However the lower level features are not semantically rich enough, and these sparse visual features are easily leaning to miss small text in detection. In addition, the high-level features are also possibly damaged by the poor low-level features. Deng et al. proposed Pixel-Link method based on the

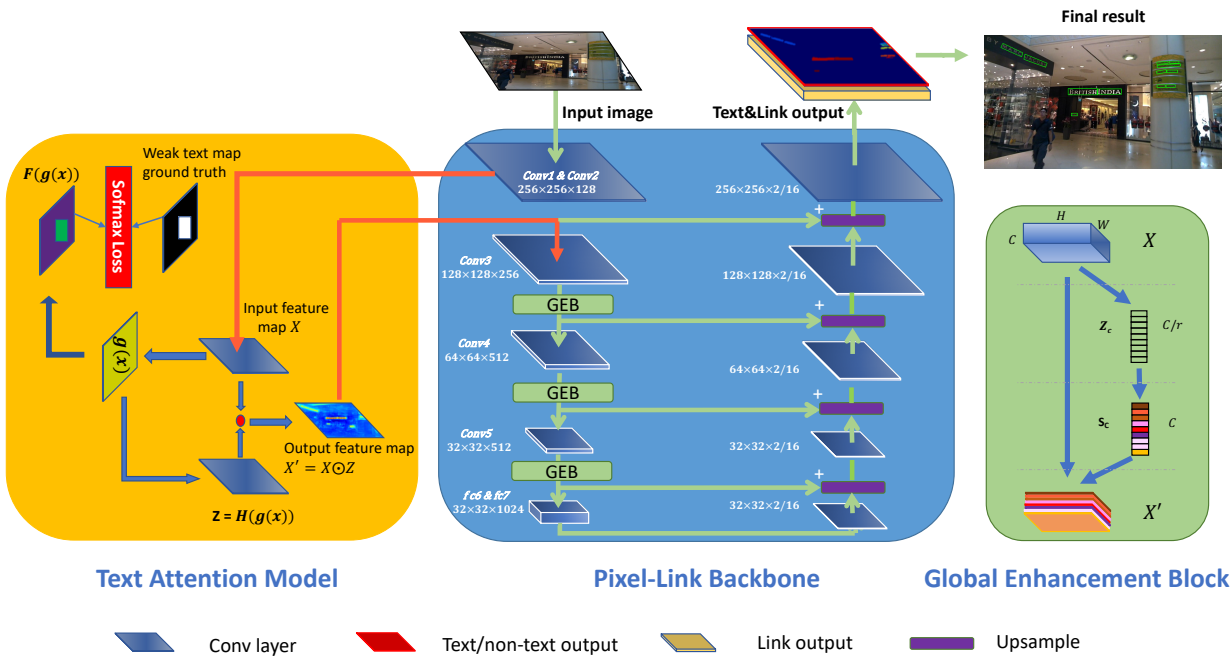

Fig. 1. Our proposed framework consists of three main components: Text attention model, Pixel-link Backbone and Global Enhancement Blocks (GEBs). Given an input image, the proposed framework can generate pixels and links outputs. Text regions can be detected by connecting the positive links.

instance segmentation for text detection [19] which is simple and effective. Their network structure follows the Deep Direct Regression design and changes the output of each stage directly to the text/non-text and link prediction. Inspired by Pixel-Link [19], our proposed approach frames the text detection task from the instance-aware segmentation perspective. First a text attention model is introduced to provide the network an attention and guide the network to focus on the scene text region at the lower level feature. Also, it brings in strong semantic information to lower layers. The lower layer is supervised by a weak text-map ground truth. Then a global enhancement block is proposed to solve the problem that the network is too local to be discriminate enough because there is no fully connected layer or global average pooling layer to provide the global context. The global enhancement block can explore the relationship of different channels and provide more semantic information to higher layers.

In summary, our method is efficient and end-to-end trainable from an instance-aware segmentation perspective. In addition, since there is no anchor box limitation, the network is more robust to predict arbitrary oriented bounding boxes compared to the existing methods. The main contributions are: 1) the text attention model is designed to provide the guided attention to the low-level text features, therefore, the performance of text detection can be improved by introducing strong semantic information to lower layers; 2) the global enhancement block is able to re-calibrate the weights among channels and improve high-level features accordingly; and 3) the experiments show that our method achieves top-performing results on three widely-used public benchmarks: *ICDAR2013*, *ICDAR2015* and *ICDAR2017-MLT*, without adopting extra training data.

2. METHODOLOGY

In the Pixel-Link [19] paper, the network extracts features of input images and obtains the prediction of text and link map output, including positive, negative pixels and links between pixels and their neighbors through bottom-up and element-wise sum operations. Positive pixels are grouped together by positives links. Then the network clusters the text regions by connective components and cv2-minarea-recantangle methods to obtain the bounding boxes of the detected text regions. However, the pixel-link network is still a fully convolutional design and there is no fully connected or global average pooling structure to enforce the global context reasoning. In addition, the features extracted by lower layers are still relatively plain without providing sufficient semantic information. Therefore, we introduce the attention model and global enhancement blocks (GEB) to improve the Pixel-Link framework.

2.1. Overview

Following the design in Pixel-Link [19], VGG16 is employed as our backbone to extract image features. As shown in Figure 1, we integrate a text attention model at the end of Conv2 to provide richer semantic information to the low-level layer and then combine the global enhancement blocks [20] after the Conv3, Conv4, Conv5 to capture more global context information respectively.

The attention model is trained with weak text maps to guide the network to focus on text regions, while blocking the influence of the surrounding noisy background. As shown in Figure 1, the text attention model takes the feature map of the Conv2 layer as the input, and generates a newly activated Conv2′ feature map as the output and continuously performs the following detection steps.

As for the higher layers, GEBs are applied. Since the text features are mainly learned from the lower layers, the GEBs contribute on introducing more global context information to enrich the feature representation in a self-supervised manner. Take the feature map from Conv3 to Conv5 as input, the GEBs output the re-calibrated feature maps as the enriched semantic features.

2.2. Text Attention Model

In order to train the text attention model, we first generate weak text maps from the word level ground truth bounding boxes. Then the attention model can be trained by taking the lower layer feature map X and the weak text map as the input, and output the feature map which is enhanced by the semantic information. The enhanced feature map guides the network to pay more attention to the scene text regions and carries more semantic information.

As shown in Figure 1, for a input lower layer feature map with size $X^{C\times H\times W}$, the network is designed to generate the intermediate layer feature maps $g(x)$, mainly through 4 dilated convolutional layers with a dilated rate = 2 (kernel size = 3×3). After the intermediate feature maps $g(x)$ have been generated, the attention model is divided into two streams. One stream passes the sigmoid layer and predicts the text map. The prediction $P=\mathcal{F}(g(x))$ has the size of $X^{2\times H\times W}$, which is same as the $G\in\mathbb{R}^{2\times H\times W}$(text and background). As for the other stream, $g(x)$ is used to generate the local activation map, $Z=\mathcal{H}(g(x))\in\mathbb{R}^{C\times H\times W}$. The generated activation map guides the network's attention to the scene text regions via an element-wise multiplication: $X'=X\odot Z$. X' is the enhanced feature map which combines both low-level general text features and the high-level semantic information. Therefore, the original feature maps X will be replaced by the enhanced feature maps X' before performing the remaining convolutions.

Examples of the proposed attention model and the improved text detection results compared to Pixel-Link are shown in the Fig 2. For better visualization of the results comparison, several text detection regions are enlarged. Compared to the Pixel-Link network [19], more missing text regions are detected by our method with the attention model.

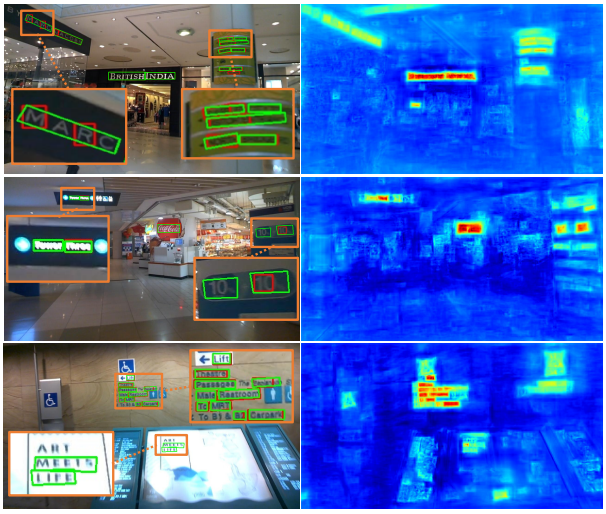

Fig. 2. Comparison between our proposed method and Pixel-Link [19] baseline method. In the first column, our detection results are represented in green boxes and the Pixel-Link results are shown in red boxes. The orange boxes show the enlarged text regions for a better visualization. Second column shows the corresponding attention maps generated by our attention model with high attention on text regions.

2.3. Global Enhancement Blocks

To capture more global context information, in conjunction with the attention model, a global activation model is introduced at the end of the higher convolutional stage. It contains several GEBs to re-calibrate weights (i.e., measure the importance) among different channels. Each GEB consists of three parts: squeeze, excitation, and fusion.

For the input feature map of GEBs with size $X=H\times W\times C$, the embedded vector Z_c of the global average pooling can be generated by Eq. (1).

$$Z_c=F_{sq}(X)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}X(i,j). \qquad (1)$$

Therefore, the local descriptors could be collected and thus global information is able to be squeezed into the vector. Here $X(i,j)$ represents the value of each pixel in the feature map. After squeezing the information into the vector, an excitation network is employed to extract the global information from $Z_c: S_c=F_{ex}(z,W)=\sigma(g(z,W))=\sigma(W_2\delta(W_1z))$, where σ is the sigmoid function and δ is the ReLU function. W_1z is the first fully connected layer with length C/r (r is the reduction radio) and $W_2\delta(W_1z)$ is the second fully connected layer with length C. Finally, the outputs of the exci-

tation network need to be fused to the heading feature map X: $X' = F_{scale}(X, S_c) = S_c \odot X$, where F_{scale} means the channel-wise multiplication between the original feature map X and the scalar S_c.

2.4. Loss function

In addition to the pixel loss and link loss in Pixel-Link, attention loss is added in our framework to supervise the text attention model with the weak text map ground truth. The loss function is as follows:

$$L = \lambda_1 L_{\text{pixel}} + L_{\text{link}} + \lambda_2 L_{\text{attention}} \tag{2}$$

Due to the importance of the pixel and attention loss, we set the $\lambda_1 = \lambda_2 = 2$, and the $L_{\text{pixel}}, L_{\text{link}}$ are the *Instance-Balanced Cross-Entropy Loss* proposed in the Pixel-Link and *Class-Balanced Cross-Entropy Loss*, respectively. The $L_{\text{attention}}$ is used to optimize the global text attention:

$$L_{\text{attention}} = \frac{1}{N} \sum_i -\log\left(\frac{e^{p_{\text{i}}}}{\sum_j e^{p_{\text{j}}}}\right) \tag{3}$$

where $L_{\text{attention}}$ is the cross entropy loss and p is the predicted output of the attention model.

3. EXPERIMENTS

3.1. Evaluation Datasets

The effectiveness of our proposed framework is evaluated on three public benchmark datasets from International Conference on Document Analysis and Recognition (ICDAR) competitions in years of 2013, 2015, and 2017.

ICDAR 2013 [21] contains 229 images for training, and 233 for testing. Text instances in this dataset are mostly in horizontal orientation and annotated as rectangles at character and word-level.

ICDAR 2015 [22] consists of 1000 scene text images collected from internet or captured by volunteers for training. The texts are annotated as text line polygons, 500 images are selected for testing. These text are mostly in arbitrary orientations and more complicated than the ICDAR 2013 dataset.

ICDAR2017-MLT [23] provides more challenging scene text images including multi-oriented, multi-scripting, and multi-lingual texts. There are 9 different kinds of lingual texts and $7,200$ training images, $1,800$ validation images, and $9,000$ testing images.

3.2. Implementation Details

All experiments are trained only on ICDAR 2013, 2015, and 2017 datasets without employing pretrained weights on SynthText 80k dataset or other external datasets as many recent methods and tested on the test subsets. `VGG16` is used as the backbone network to extract features and Stochastic Gradient Descent (SGD) is adopted as the optimization method with momentum $= 0.9$, batch size $= 8$. The `VGG` model and the new layers we added are all initialized by Xavier initialization. The learning rate is set to 0.008 for the first $1,000$ iterations and then 0.01 for the following steps. First, we train the model on ICDAR 2015 and obtained the testing results for ICDAR 2015. Then the ICDAR 2015 pre-trained model is employed and finetuned on the ICDAR 2013 for another $80K$ iterations. For ICDAR 2017-MLT, both the training and validation splits are used for training and we train the network with $300K$ iterations. All input images are resized to 512×512 after rotation augmentation.

In our experiments, the reduction radio r in the GEB equals to 16. Considering the limited ability of single fully connected layer and the computation overhead of stacking `FC` layers, we choose to adopt 2 `FC` layers.

Table 1. Text detection results on ICDAR 2013 dataset.

Method	Recall (%)	Precision (%)	F-measure (%)
Seglink[18]	83.00	88.00	85.00
TextBoxes MS [16]	83.00	89.00	86.00
TextBoxes++_MS [17]	86.00	92.00	89.00
CTPN [9]	83.00	83.00	88.00
WordSup [26]	88.00	93.00	90.00
Lyu et al. [27]	79.40	93.30	85.80
Pixel-Link [19]	87.50	88.60	88.10
Mask Textspotter [24]	94.10	88.10	91.00
Ours	89.20	91.50	90.30

3.3. Experiment Results and Analysis

Here, both qualitative and quantitative results are provided to demonstrate the effectiveness of our proposed method to detect horizontal, multi-oriented, and multi-lingual scene texts. Figure 3 shows some text detection examples by our proposed network, and Tables 1, 2, 3 show the detailed performance compared with the state-of-the-arts.

Horizontal Text, ICDAR 2013 Results: As shown in Table 1, our proposed method achieves 90.3% f-score which outperforms the baseline Pixel-Link more than 2% and is comparable to the state-of-the-art results. The performance of Mask Textspotter [24] method is 0.7% higher than our method, however, it used SynthText 80k dataset for pre-training.

Oriented Text, ICDAR 2015 Results: As shown in Table 2, our proposed method obtains 85.94% on the ICDAR 2015 dataset, which is about 2.3% higher than the Pixel-Link baseline. It is still around 1% lower than the PSENET [25]. Empirically, the gain of PSENET most likely results from the advantages of ResNet and FPN backbone networks adopted in PSENET.

Multi-Lingual Text, ICDAR2017-MLT Results: As shown in Table 3, our method outperforms all the state-of-the-arts

Fig. 3. Examples of scene text detection results by our proposed method.

Table 2. Text detection results on ICDAR 2015 dataset.

Methods	Recall (%)	Precision (%)	F-measure (%)
CTPN [9]	51.60	74.20	60.90
Seglink [18]	70.80	73.10	75.00
East [28]	72.80	80.50	76.40
SSTD [29]	73.00	80.00	77.00
WordSup [26]	77.00	79.30	78.20
RRPN [10]	77.00	84.00	80.00
R2CNN [11]	79.68	85.62	82.54
Textboxes++ [17]	78.50	87.80	82.90
East [28]	78.30	83.30	80.70
Text-snake [30]	80.40	84.90	82.60
Pixel-Link [19]	82.00	85.50	83.70
FTSN [31]	80.00	88.60	84.10
IncepText [32]	80.60	90.50	85.30
PSENET [25]	89.30	85.22	87.21
Ours	84.91	87.00	85.94

methods and achieves 67.48% f-score which is 7% higher than the Pixel-Link baseline by using the same training settings. Compared to FOTS [33] which used the recognition results to supervise the training process, our model achieves a better performance.

3.4. Effects of Attention Model and GEBs

To verify the effects of the proposed attention model and GEBs, as shown in Table 4, a series of experiments are conducted on the ICDAR 2015 dataset with different settings. Without pretrained weights and external dataset, the best performance of Pixel-Link-2S baseline method (The size of final output feature map is half of the original image) achieves 83.7% f-score. By only adding the **attention model**, the performance (f-score) increases to 85.16%. The high improvement of recall shows that attention model could significantly assist in finding the hard text examples in natural scene images. By only adding the **GEBs**, the F-score of the detection is boosted to 84.53%. By adding both the **attention model** and the **GEBs**, the performance is boosted up to 85.94%. This proves that with the attention model and GEBs, the network is able to provide more semantic information in lower layers and global information in higher layers and achieve better performance of text detection.

Table 3. Detection results on ICDAR 2017-MLT dataset with multi-lingual texts.

Method	Recall	Precision	F-measure
TH-DL [23]	34.80	67.80	46.00
Pixel-Link [19]	55.37	67.07	60.66
SARI FDU RRPN V1 [10]	55.50	71.20	62.40
Sensetime OCR [23]	69.40	56.90	62.60
SCUT DLVClab1 [23]	54.50	80.30	65.00
Lyu et al. [27]	55.60	83.80	66.80
FOTS [33]	57.51	80.95	67.25
Ours	63.50	72.00	67.48

4. CONCLUSION

In this paper, we have presented an effective end-to-end framework for detecting multi-lingual scene texts in arbitrary orientations by integrating text attention model and global enhancement block with the pixel-link method without adopting pretrained weights or extra synthetic datasets. Under the guidance of the text region attention and the global context

Table 4. Effects of the Proposed Attention Model and Global Enhancement Blocks (GEBs) on ICDAR 2015 Dataset.

Attention	GEB	Recall(%)	Precision(%)	F-measure(%)
×	×	82.00	85.50	83.70
✓	×	84.39	85.94	85.16
×	✓	83.50	85.60	84.53
✓	✓	84.91	87.00	85.94

of the global enhancement block, our method achieves comparable performance with the recent state-of-the-arts on three benchmark datasets. In the future, we aim to improve the proposed pipeline to better handle irregular curved text instances.

5. ACKNOWLEDGEMENT

This work was supported in part by NSF grant IIS-1400802.

6. REFERENCES

[1] Xuejian Rong, Chucai Yi, and Yingli Tian, "Unambiguous text localization and retrieval for cluttered scenes," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3279–3287, 2017. 1

[2] Xuejian Rong, Chucai Yi, and Yingli Tian, "Recognizing text-based traffic guide panels with cascaded localization network," *ECCV Workshop, 2016*, pp. 1484–1493, 2016. 1

[3] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, pp. 1–20, 2015. 1

[4] Boris Epshtein, Eyal Ofek, and Yonatan Wexler, "Detecting text in natural scenes with stroke width transform," *CVPR*, 2010. 1

[5] Huizhong Chen et al., "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," *2011 18th IEEE International Conference on Image Processing*, pp. 2609–2612, 2011. 1

[6] Ross Girshick, "Fast r-cnn," in *International Conference on Computer Vision (ICCV)*, 2015. 1

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015. 1

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition*, 2014. 1

[9] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao, "Detecting text in natural image with connectionist text proposal network," in *ECCV*, 2016. 1, 4, 5

[10] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue, "Arbitrary-oriented scene text detection via rotation proposals," *CoRR*, vol. abs/1703.01086, 2017. 1, 5

[11] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo, "R2cnn: Rotational region cnn for orientation robust scene text detection," *CoRR*, vol. abs/1706.09579, 2017. 1, 5

[12] Linjie Deng, Yanxiang Gong, Yi Lin, Jingwen Shuai, Xiaoguang Tu, Yufei Zhang, Zheng Ma, and Mei Xie, "Detecting multi-oriented text with corner-based region proposals," *CoRR*, vol. abs/1804.02690, 2018. 1

[13] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *NIPS*, 2016. 1

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg, "SSD: Single Shot MultiBox Detector," in *ECCV*, 2016. 1

[15] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," *CVPR*, 2016. 1

[16] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu, "Textboxes: A fast text detector with a single deep neural network," in *AAAI*, 2017. 1, 4

[17] Baoguang Shi Minghui Liao and Xiang Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3676–3690, 2018. 1, 4, 5

[18] Baoguang Shi, Xiang Bai, and Serge J. Belongie, "Detecting oriented text in natural images by linking segments," *CVPR*, 2017. 1, 4, 5

[19] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai, "Pixellink: Detecting scene text via instance segmentation," *CoRR*, vol. abs/1801.01315, 2018. 2, 3, 4, 5

[20] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," *CoRR*, vol. abs/1709.01507, 2017. 3

[21] Dimosthenis Karatzas et al., "Icdar 2013 robust reading competition," *2013 12th International Conference on Document Analysis and Recognition*, pp. 1484–1493, 2013. 4

[22] Dimosthenis Karatzas et al., "Icdar 2015 competition on robust reading," *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1156–1160, 2015. 4

[23] Nibal Nayef et al., "Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification - rrc-mlt," *ICDAR*, vol. 01, pp. 1454–1459, 2017. 4, 5

[24] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," *CoRR*, vol. abs/1807.02242, 2018. 4

[25] Xiang Li, Wenhai Wang, Wenbo Hou, Ruo-Ze Liu, Tong Lu, and Jian Yang, "Shape robust text detection with progressive scale expansion network," *CoRR*, vol. abs/1806.02559, 2018. 4, 5

[26] Han Hu, Chengquan Zhang, Yuxuan Luo, Yuzhuo Wang, Junyu Han, and Errui Ding, "Wordsup: Exploiting word annotations for character based text detection," *ICCV*, 2017. 4, 5

[27] Pengyuan Lyu, Cong Yao, Wenhao Wu, Shuicheng Yan, and Xiang Bai, "Multi-oriented scene text detection via corner localization and region segmentation," *CoRR*, vol. abs/1802.08948, 2018. 4, 5

[28] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang, "East: An efficient and accurate scene text detector," *CVPR*, 2017. 5

[29] Pan He, Weilin Huang, Tong He, Qile Zhu, Yu Qiao, and Xiaolin Li, "Single shot text detector with regional attention," 2017, Proceedings of Internatioanl Conference on Computer Vision (ICCV). 5

[30] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao, "Textsnake: A flexible representation for detecting text of arbitrary shapes," *CoRR*, vol. abs/1807.01544, 2018. 5

[31] Yuchen Dai, Zheng Huang, Yuting Gao, and Kai Chen, "Fused text segmentation networks for multi-oriented scene text detection," *CoRR*, vol. abs/1709.03272, 2017. 5

[32] Qiangpeng Yang, Mengli Cheng, Wenmeng Zhou, Yan Chen, Minghui Qiu, Wei Lin, and Wei Chu, "IncepText: A New Inception-Text Module with Deformable PSROI Pooling for Multi-Oriented Scene Text Detection," in *IJCAI*, 2018. 5

[33] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan, "Fots: Fast oriented text spotting with a unified network," *CoRR*, vol. abs/1801.01671, 2018. 5