Supplementary Materials for "Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network"

Wenhai Wang*[1], Enze Xie*[2,4], Xiaoge Song[1], Yuhang Zang[3], Wenjia Wang[2], Tong Lu†[1], Gang Yu[4], and Chunhua Shen[5]

[1]National Key Lab for Novel Software Technology, Nanjing University
[2]Tongji University
[3]University of Electronic Science and Technology of China
[4]Megvii (Face++) Technology Inc.
[5]The University of Adelaide
{wangwenhai362, Johnny_ez, sxg514}@163.com, yuhangzang@foxmail.com, wwj940312@126.com, lutong@nju.edu.cn, yugang@megvii.com, chunhua.shen@adelaide.edu.au

A. Robustness Analysis

To further demonstrate the robustness of the proposed PAN, we evaluate the model by training on one dataset and testing on other datasets. Based on the annotation level, we divide the datasets into two groups which are word level and text line level datasets. SynthText, ICDAR 2015 and Total-Text are annotated at word level, while CTW1500 and MSRA-TD500 are annotated at text line level. For fair comparisons, we train all model without any external dataset, and the short side of test images in ICDAR 2015, MSRA-TD500, CTW1500 and Total-Text are set to 736, 736, 640, 640 respectively.

The cross-dataset results of PAN are shown in Table I. Notably, the proposed PAN trained on SynthText (a synthetic dataset) have reluctantly satisfied performance on ICDAR 2015 and Total-Text, which indicates that even without any manually annotated data, PAN can satisfy the scene with low precision requirements. The PAN trained on manually annotated dataset has over 64% F-measure in the cross-dataset evaluation, which is still competitive. Furthermore, in the cross-dataset evaluation at text line level, all models achieve the F-measure of nearly 75% even the training and the testing are performed on quadrangle and curved text datasets respectively. These cross-dataset experiments demonstrate that the proposed PAN is robust in generalizing to brand new datasets.

Annotation	Train Set→Test Set	P	R	F
Word	SynthText → ICDAR 2015	65.9	46.9	54.8
	SynthText → Total-Text	69.1	40.8	51.3
	ICDAR 2015 → Total-Text	72.0	57.8	64.1
	Total-Text → ICDAR 2015	77.6	65.5	71.1
Text Line	CTW1500 → MSRA-TD500	76.6	73.1	74.8
	MSRA-TD500 → CTW1500	82.4	69.1	75.2

Table I. Cross-dataset results of PAN on word-level and line-level datasets. "P", "R" and "F" represent the precision, recall and F-measure respectively.

B. Comparisons with Other Semantic Segmentation Methods

Unlike common semantic segmentation tasks, text detection needs to distinguish different text instances that lie closely. So feature map resolution matters and cannot be too small. However, most of high efficiency segmentation methods (i.e. BiSeNet [3]) make prediction on $1/8$ feature map, sacrificing accuracy for speed. Their speed will reduce sharply if using $1/4$ feature map directly. Thus, 'how to keep the high efficiency and the high resolution feature map simultaneously?" is a challenging problem, and our answer is "ResNet18 + 2FPEM + FFM". We compare our method with two methods BiSeNet [3] and CU-Net [2] on CTW1500. For fair comparisons, we set the backbone of BiSeNet to ResNet18 and use one of default settings of CU-Net, which has the similar speed with our method. As shown in Table II, our method enjoys obviously better accuracy (+2.2% and +4.6%) at the similar speed.

*Authors contributed equally.
†Corresponding author.

Methods	Ext.	F (%)	FPS
BiSeNet (ResNet18) [3]	-	78.8	25.9
CU-Net-2 (m=128, n=32) [2]	-	76.4	39.3
Ours (ResNet18 + 2FPEM + FFM)	-	**81.0**	**39.8**

Table II. The results on CTW1500 of different segmentation methods. "F" means F-measure. "Ext." indicates external data.

(a) (b) (c) (d)

Figure I. Failure Samples.

C. Failure Samples

As demonstrated in previous experiments, the proposed PAN works well in most cases of arbitrary-shaped text detection. It still fails for some difficult cases, such as large character spacing (see Fig. I (a)), symbols (see Fig. I (b)) and false positives (see Fig. I (c)(d)). Large character spacing is an unresolved problem which also exists in other state-of-the-art methods such as RRD [1]. For symbol detection and false positives, PAN is trained on small datasets (about 1000 images) and we believe this problem will be alleviated when increasing training data.

D. More Detected Results on CTW1500, Total Text, ICDAR 2015 and MSRA-TD500

In this section, we show more test examples produced by PAN on different datasets in Fig. II (CTW1500) Fig. III (Total-Text), Fig. IV (ICDAR 2015) and Fig. V (MSRA-TD500). From these results, we can find that the proposed PAN have the following abilities: i) separating adjacent text instances with narrow distances; ii) locating the arbitrary-shaped text instances precisely; iii) detecting the text instances with various orientations; iv) detecting the long text instances; v) detecting the multiple Lingual text. Meanwhile, thanks to the strong feature representation, PAN can also locate the text instances with complex and unstable illumination, different colors and variable scales.

References

[1] Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.

[2] Zhiqiang Tang, Xi Peng, Shijie Geng, Yizhe Zhu, and Dimitris Metaxas. Cu-net: Coupled u-nets. In *BMVC*, 2018.

[3] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proc. Eur. Conf. Comp. Vis.*, 2018.

Figure II. Detection results on CTW1500.

Figure III. Detection results on Total-Text.

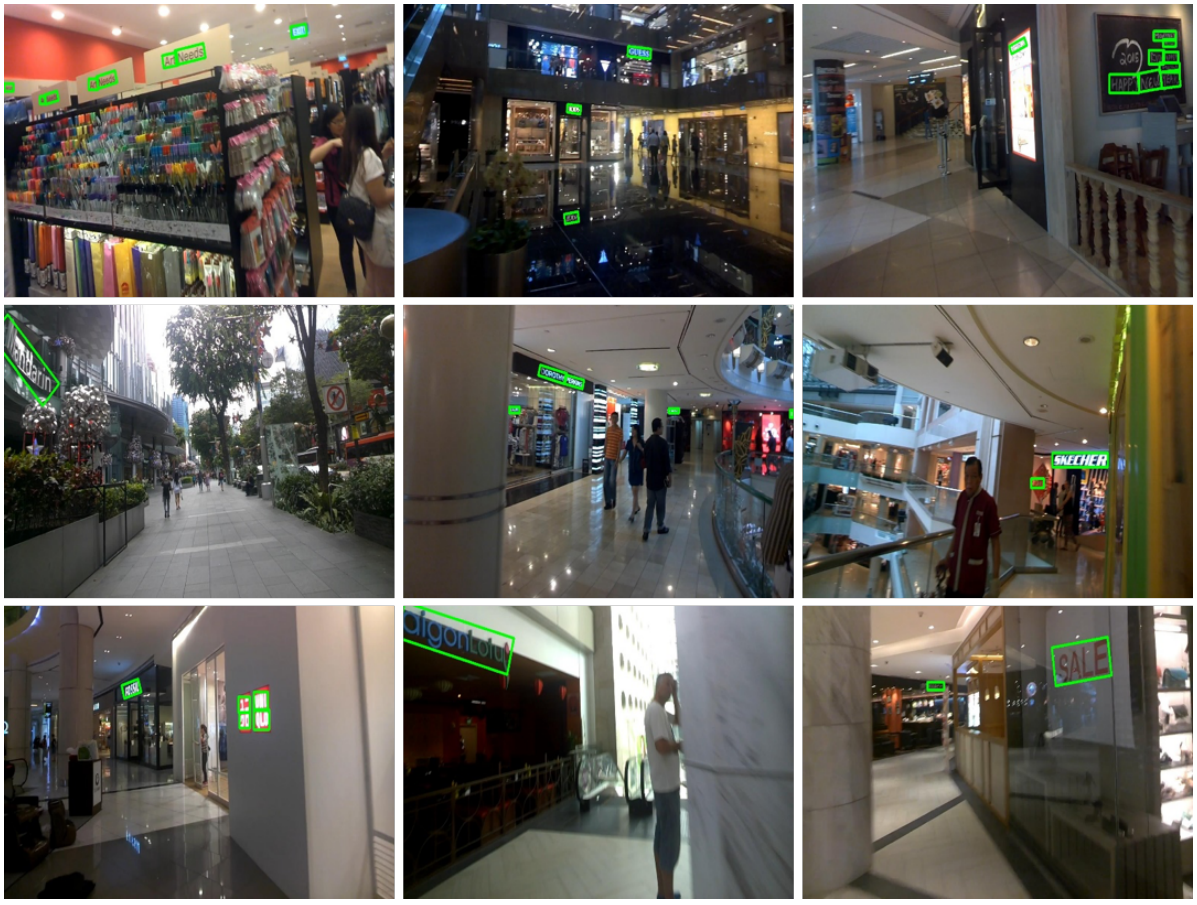

Figure IV. Detection results on ICDAR 2015.

Figure V. Detection results on MSRA-TD500.