IncepText: A New Inception-Text Module with Deformable PSROI Pooling for Multi-Oriented Scene Text Detection

Qiangpeng Yang, Mengli Cheng, Wenmeng Zhou, Yan Chen, Minghui Qiu, Wei Lin
Alibaba Group
{qiangpeng.yqp, mengli.cml, wenmeng.zwm, chenyan.cy, minghui.qmh, weilin.lw}@alibaba-inc.com

Abstract

Incidental scene text detection, especially for multi-oriented text regions, is one of the most challenging tasks in many computer vision applications. Different from the common object detection task, scene text often suffers from a large variance of aspect ratio, scale, and orientation. To solve this problem, we propose a novel end-to-end scene text detector IncepText from an instance-aware segmentation perspective. We design a novel Inception-Text module and introduce deformable PSROI pooling to deal with multi-oriented text detection. Extensive experiments on ICDAR2015, RCTW-17, and MSRA-TD500 datasets demonstrate our method's superiority in terms of both effectiveness and efficiency. Our proposed method achieves 1st place result on ICDAR2015 challenge and the state-of-the-art performance on other datasets. Moreover, we have released our implementation as an OCR product which is available for public access. [1]

1 Introduction

Scene text detection is one of the most challenging tasks in many computer vision applications such as multilingual translation, image retrieval, and automatic driving. The first challenge is scene text contains various kinds of images, such as street views, posters, menus, indoor scenes, *etc*. Furthermore, the scene text has large variations in both foreground texts and background objects, and also with various lighting, burring, and orientation.

In the past years, there have been many outstanding approaches focus on scene text detection. The key point of text detection is to design features to distinguish text and non-text regions. Most of the traditional methods such as MSER [Neumann and Matas, 2010] and FASText [Busta *et al.*, 2015] use manually designed text features. These methods are not robust enough to handle complex scene text. Recently, Convolutional Neural Network (CNN) based methods achieve the state-of-the-art results in text detection and recognition [He *et al.*, 2016b; Tian *et al.*, 2016; Zhou *et al.*, 2017; He *et al.*, 2017]. CNN based models have a powerful capability of feature representation, and deeper CNN models are able to extract higher level or abstract features.

In the literature, there are mainly two types of approaches for scene text detection, namely indirect and direct regressions. Indirect regression methods predict the offsets from some box proposals, such as CTPN [Tian *et al.*, 2016] and RRPN [Ma *et al.*, 2017]. These methods are based on Faster-RCNN [Ren *et al.*, 2015] framework. Recently, direct regression methods have achieved high performance for scene text detection, e.g. East [Zhou *et al.*, 2017] and DDR [He *et al.*, 2017]. Direct regression usually performs boundary regression by predicting the offsets from a given point.

In this paper, we solve this problem from an instance-aware segmentation perspective that mainly draws on the experience of FCIS [Li *et al.*, 2016]. Different from common object detection, scene text often suffers from a large variance of scale, aspect ratio, and orientation. Therefore, we design a novel Inception-Text module to deal with these challenges. This module is inspired by Inception module [Szegedy *et al.*, 2015] in GoogLeNet, we choose multi branches of different convolution kernels to deal with the text of different aspect ratios and scales. At the end of each branch, we add a deformable convolution layer to adapt multi orientations. Another improvement is that we replace the PSROI pooling in FCIS with deformable PSROI pooling [Dai *et al.*, 2017a]. According to our experiments, deformable PSROI pooling has better performance in the classification task.

Our main contributions can be summarized as follows:

- We propose a new Inception-Text module for multi-oriented scene text detection. According to our experiments, this module shows a significant increase in accuracy with little computation cost.
- We propose to use deformable PSROI pooling module to deal with multi-oriented text. The qualitative study of learned offset parts in deformable PSROI pooling and quantitive evaluations show its efficiency to handle arbitrary oriented scene text.
- We evaluate our proposed method on three public datasets ICDAR2015, RCTW-17 and MSRA-TD500, and show that our proposed method achieves the state-of-the-art performance on several benchmarks without using any extra data.

[1] https://market.aliyun.com/products/57124001/cmapi020020.html

- Our proposed method has been implemented as an API service in our OCR product, which is available in public.

The rest of this paper is organized as follows: we first give a brief overview of scene text detection and mainly focus on multi-oriented scene text detection. Then we describe our proposed method in detail and present experimental results on three public benchmarks. We conclude this paper and discuss future work at the end.

2 Related Work

Scene text detection has been extensively studied in the last decades. Most of the previous work focused on horizontal text detection, while more recent research studies have concentrated on multi-oriented scene text detection. Below we briefly introduce the related studies.

HMP. HMP [Yao *et al.*, 2016] is inspired by Holistically-Nested Edge Detection (HED) [Xie and Tu, 2015]. It simultaneously predicts the probability of text regions, characters and the relationship among adjacent characters with a unified framework. Therefore, two kinds of label maps are needed: the label map of text line and the label map of characters. Graph partition algorithm is used to determine the retained and eliminated linkings, which is not robust enough for scene text detection.

SegLink. SegLink [Shi *et al.*, 2017] introduced a novel text detection framework which decomposes the text into two locally detectable elements, segments and links. A segment is an oriented box of a text line, while a link indicates the two adjacent segments belong to the same text line or not. The segments and links are detected at multiple scales by a fully convolutional network. However, a post-process step of combining segments is also needed in SegLink.

RRPN. RRPN [Ma *et al.*, 2017] is modified from Faster-RCNN [Ren *et al.*, 2015] for multi-oriented scene text detection. The main difference between RRPN and Faster-RCNN is that anchors with six different orientations are generated at each position of the feature map. The angle information is a regression target in regression task to get more accurate boxes.

EAST. EAST [Zhou *et al.*, 2017] proposed an efficient scene text detector which uses a single fully convolutional network. The network has two branches: a segmentation task predicts the text score map and a regression task which directly predicts the final box for each point in the text region. According to our experiments, this framework is not suitable for long text line, maybe a line grouping method is needed in post-processing.

DDR. Deep Direct Regression (DDR) [He *et al.*, 2017] is very similar to EAST. They use a fully convolutional network to directly predict the final quadrilateral from a given point. In testing, a multi-scale sliding window strategy is used, which is very time-consuming. The main limitation is the same as EAST.

In a nutshell, different from previous models, our method is an end-to-end trainable neural network from an instance-aware segmentation perspective. We design an new Inception-Text module for multi-oriented text detection. To handle arbitrary oriented text, we replace standard PSROI pooling with deformable PSROI pooling and demonstrate its efficiency. Below we present our method in detail.

3 The Proposed Method

3.1 Overview

Our proposed method is based on FCIS [Li *et al.*, 2016] framework, which is originally proposed for instance-aware segmentation. We design a novel Inception-Text module and use deformable PSROI pooling to extend this framework for scene text detection. Figure 1 shows an overview of our model architecture.

In Figure 1, the basic feature extraction module is ResNet-50 [He *et al.*, 2016a]. For scene text detection, finer feature information is very important especially for segmentation task, the final downsampling in res stage 5 may lose some useful information. Therefore we exploit hole algorithm [Long *et al.*, 2015] in res stage 5 to maintain to field of view. The stride=2 operations in the stage are changed to stride=1, and all convolutional filters on the stage use hole algorithm instead to compensate the reduced stride.

To predict accurate location of small text regions, low-level features also need to be taken into consideration. As illustrated in Figure 1, layer *res4f* and layer *res5c* are upsampled by a factor 2 and added with layer *res3d*. Then these two fused feature maps are followed by Inception-Text module which is designed for scene text detection. We replaced PSROI pooling layer in FCIS with deformable PSROI pooling, because standard PSROI pooling can only handle horizontal text while scene text always has arbitrary orientations. Similar to FCIS, we obtain text boxes with masks and classification scores as in Figure 1, and then we apply NMS on the boxes based on their scores. For each unsuppressed boxes, we find its similar boxes which are defined as the suppressed boxes that overlap with the unsuppressed box by $IoU >= 0.5$. The prediction masks of the unsuppressed boxes and its similar boxes are merged by weighted averaging, pixel-by-pixel, using the classification scores as their averaging weights. And then a simple minimal quadrilateral algorithm is used to generate the oriented boxes.

3.2 Inception-Text

Our proposed Inception-Text module mainly has two parts, which is shown in Figure 2. The first part has three branches which is very similar to Inception module in GoogLeNet. Firstly, we add a 1×1 conv-layer to decrease the number of channels in the original feature map. To deal with the different scales of text, three different convolution kernels: 1×1, 3×3 and 5×5 are selected. Different scales of text are activated in different branches. We further factorize the $n \times n$ convolution into two convolutions, which is a $1 \times n$ convolution followed by a $n \times 1$ convolution. According to [Szegedy *et al.*, 2016], these two structures have same receptive fields, while the factorization has a lower computational cost.

Comparing to standard Inception module, another important difference is that we add a deformable convolution layer at the end of each branch of the first part. Deformable convolution layer is firstly introduced in [Dai *et al.*, 2017a], where the spatial sampling locations are augmented with additional

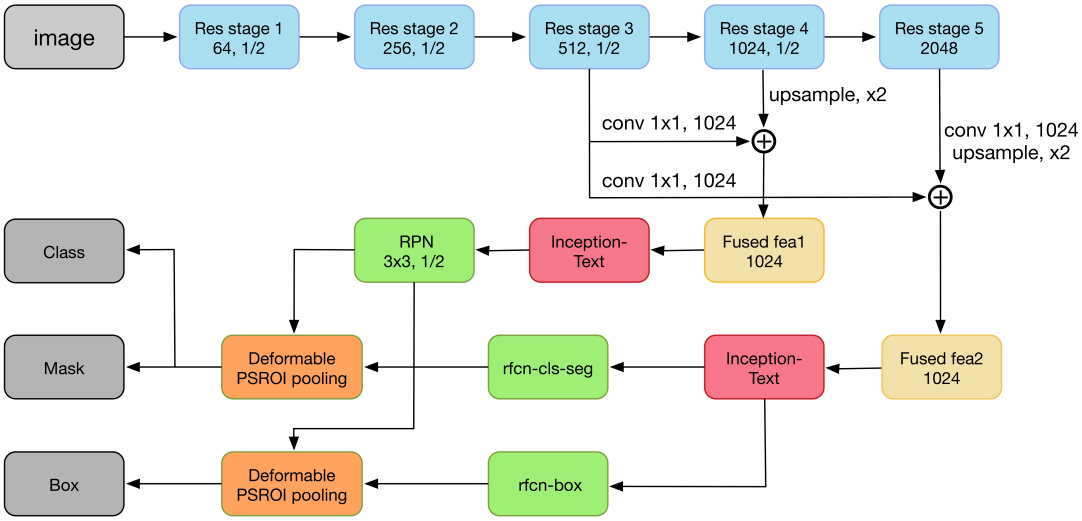


Figure 1: An overview of IncepText architecture. The basic feature extraction module in this figure is ResNet-50. Inception-Text module is appended after feature fusion, and the original PSROI pooling is replaced by deformable PSROI pooling.

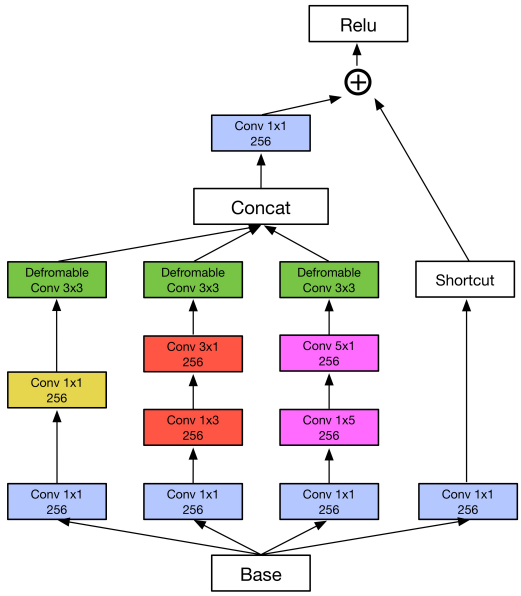
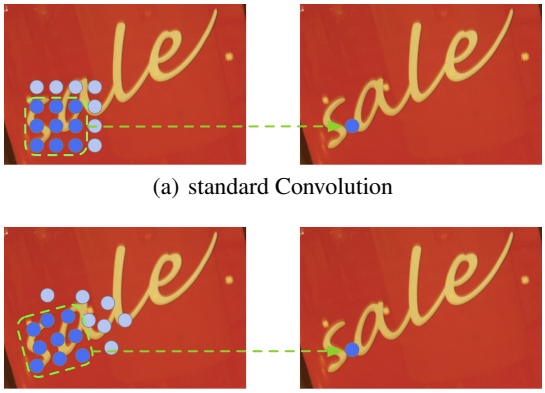


Figure 2: Inception-Text module.

offsets learned from data. In scene text detection, arbitrary text orientation is one of the most challenging problems, deformable convolution allows free form deformation of sampling grid instead of regular sampling grid in standard convolution. This deformation is conditioned over the input features, thus the receptive field is adjusted when the input text is rotated. To illustrate this, we compare standard convolution and deformable convolution in Figure 3. Clearly, the standard convolution layer can only handle horizontal text regions, while the deformable convolution layer is able to use an adaptive receptive field to capture regions with different orientations. More quantitive results are illustrated in Table 1.

Furthermore, similar to Inception-ResNet V2 [Szegedy *et al.*, 2017], we also apply the shortcut design followed by a 1×1 conv-layer.



(a) standard Convolution

(b) Deformable Convolution

Figure 3: Comparison between standard convolution and deformable convolution. The receptive field in standard convolution (a) is fixed while deformable convolution (b) has adaptive receptive field.

3.3 Deformable PSROI Pooling

PSROI pooling [Dai *et al.*, 2016] is a variant of regular ROI pooling, which operates on position-sensitive score maps with no weighted layers following. The position-sensitive property encodes useful spatial information for classification and object location.

However, for multi-oriented text detection task, PSROI pooling can only deal with axis-aligned proposals. Hence we use deformable PSROI pooling [Dai *et al.*, 2017a] to add offsets to the spatial binning positions in PSROI pooling. These

offsets are learned purely from data. The deformable PSROI pooling is defined as:

$$r_c(i,j) = \sum_{(x,y)\in bin(i,j)} \frac{\hat{z}_{i,j}(x + x_0 + \Delta x, y + y_0 + \Delta y)}{n}, \tag{1}$$

where $r_c(i,j)$ is the pooled response in the (i,j)-th bin, $\hat{z}_{(i,j)}$ is the transformed feature map, (x_0, y_0) is the top-left corner of an ROI, n is the number of pixels in the bin. Δx and Δy are learned from a fc layer.

Figure 4: Visualization of learned offsetted parts in deformable PSROI pooling. We have 21×21 bins (red) for the each input ROI (yellow). Deformable PSROI pooling tend to learn the context surrounding the text.

Deformable PSROI Pooling is proposed for non-rigid object detection, and we apply it in multi-oriented scene text detection. In Figure 4, we take a brief visualization of how the parts are offset to cover the text with arbitrary orientation. More quantitative analyses are shown in Table 1.

3.4 Ground Truth and Loss Function

The ground truth of text instance is exemplified in Figure 5. Different from general instance-aware segmentation task, we do not have the pixel-wise label of text and non-text. Instead, the pixels in the quadrilateral are all positive, while the left pixels are negative.

Figure 5: Ground Truth. The target of regression task is colored in yellow dashed lines, and the mask target is filled with gray quadrilateral.

The loss function is similar to FCIS [Li *et al.*, 2016], which can be formulated as:

$$L = L_{rcls} + L_{rbox} + L_{cls} + L_{box} + \lambda_m L_{mask} \tag{2}$$

where L_{rcls} and L_{rbox} are classification and regression loss in RPN stage, while L_{cls} and L_{box} are in RCNN stage. L_{mask} is cross-entropy loss for segmentation task, where λ_m in our experiments is set to 2.

4 Experiments

4.1 Benchmark Datasets

We evaluated our method on three public benchmark datasets. These datasets all have scene text with arbitrary orientations.

ICDAR2015. This dataset was used in challenge 4 of ICDAR2015 Robust Reading competition [Karatzas *et al.*, 2015]. It contains 1000 images for training while 500 images for testing. These images were collected by Google Glass, which suffers from motion blur and low resolution. The bounding boxes of text have multi-orientations and they are specified by the coordinates of their four corners in a clock-wise manner.

RCTW-17. This is a competition on reading Chinese Text in images, which contains a various kinds of images, including street views, posters, menus, indoor scenes and screenshots. Most of the images are taken by phone cameras. This dataset contains about 8000 training images and 4000 test images. Annotations of RCTW-17 are similar to ICDAR2015.

MSRA-TD500. This was collected from indoor and outdoor scenes using a pocket camera [Yao *et al.*, 2012]. This dataset contains 300 training images and 200 testing images. Different from ICDAR2015, the basic unit in this dataset is text line rather than word and the text line may be in different languages, Chinese, English, or a mixture of both.

4.2 Experimental Setup

Our proposed network was trained end-to-end using ADAM optimizer [Kingma and Ba, 2014]. We adopted the multi-step strategy to update learning rate, and the initial learning rate is 10^{-3}. Each image is randomly cropped and scaled to have short edge of $\{640, 800, 960, 1120\}$. The anchor scales are $\{2, 4, 8, 16\}$, and ratios are $\{0.2, 0.5, 2, 5\}$. And we also applied online hard example mining (OHEM) for balancing the positive and negative samples.

4.3 Experimental Results

Impact of Inception-Text and Deformable PSROI pooling.

We conducted several experiments to evaluate the effectiveness of our design. These experiments mainly focus on evaluating two important modules in our model: Inception-Text and deformable PSROI pooling. Table 1 summarizes the results of our models with different settings on ICDAR 2015.

Inception-Text. We designed this module to handle text with multiple scales, aspect ratios and orientations. To evaluate this module, we set the input image with text of three different scales and visualize the feature maps at the end of each branch. An interesting phenomenon is exemplified in Figure 6. The left branch (kernel size = 1) in Figure 2 is activated with three scales, some channels of the middle branch (kernel size = 3) are not activated with the smallest text, and some channels of the right branch (kernel size = 5) are only activated with the largest text.

We also conducted another experiment in Figure 7. We found that, if we used all three branches in testing, all words will be detected with high confidence. When we remove the

left branch, the scores of three words are decreased simultaneously and the smallest word decreased farthest. If we remove the middle branch, the influence of the smallest word is reduced. When we remove the right branch, the biggest word is missed, while the other two words have little influence.

These two experiments demonstrate that different scales of text are activated in different branches. The branch with large kernel size has more influence on large text, while the small text is mainly influenced by branch of small kernel size.

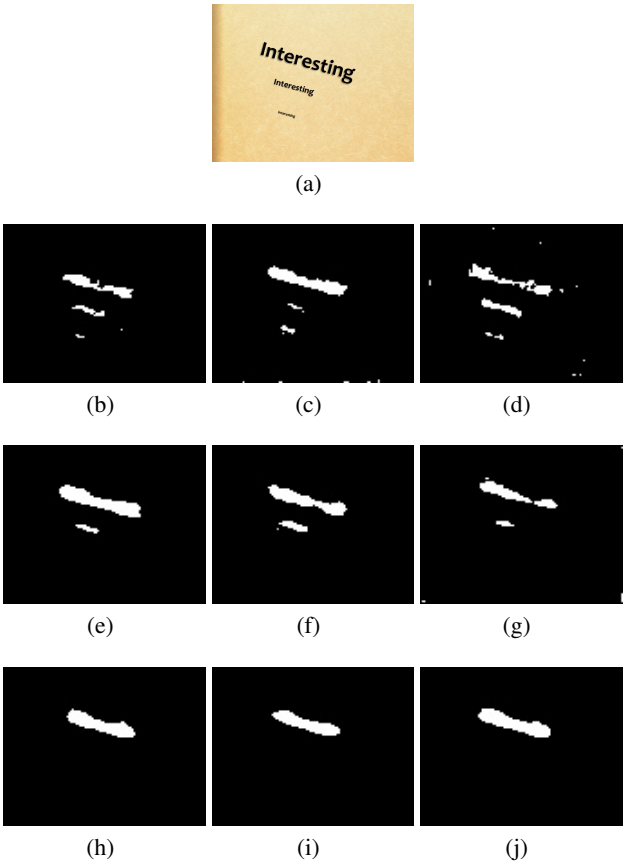


Figure 6: Feature maps of each branch in Inception-Text module. (a) Input image. (b)(c)(d) Feature maps of the left branch. (e)(f)(g) Feature maps of the middle branch. (h)(i)(j) Feature maps of the right branch.

In addition, we compared the origin Inception module and our Inception-Text module, our design has about 0.017 improvement in recall (0.803 vs. 0.786) while the precision is almost the same. Comparing to the origin design (without Inception or Inception-Text), both recall (0.803 vs. 0.775) and precision (0.891 vs. 0.873) have large improvements. The final F-measure has over 0.02 improvement.

Deformable PSROI pooling. When we replace the standard PSROI pooling with deformable PSROI pooling, the precision has a large improvement (0.905 vs. 0.891). This indicates the power of our model for distinguishing text and non-text has been enhanced, more difficult regions have been correctly classified. After adding this module, the final F-measure increases from 0.845 to 0.853.

Inception	Inception-Text	Deformable PSROI pooling	Recall	Precision	F-measure
✗	✗	✗	0.775	0.873	0.821
✓	✗	✗	0.786	0.886	0.833
✗	✓	✗	0.803	0.891	0.845
✗	✓	✓	**0.806**	**0.905**	**0.853**

Table 1: Effectiveness of Inception-Text module and deformable PSROI pooling on ICDAR2015 incidental scene text location task.

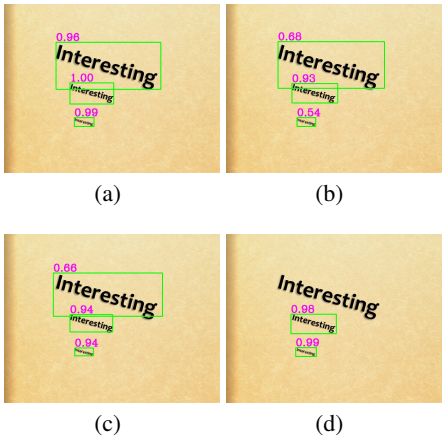


Figure 7: Influence of different branch. (a) Result with three branches. (b) Result without the left branch. (b) Result without the middle branch. (c) Result without the right branch.

Experiments on Scene Text Benchmarks.
We proceed to compare our method with the state-of-the-art methods on the public benchmark datasets, in Table 2 (ICDAR2015), Table 3 (RCTW17) and Table 4 (MSRA-TD500).

On ICDAR2015, we only used 1000 original images without extra data to train our network. With single scale of 960, our proposed method achieves an F-measure of 0.853. When testing with two scales $[960, 1120]$, the F-measure is 0.868 which is over 0.02 higher than the second best method in terms of absolute value. To the best of our knowledge, this is the best reported result in literature. Similar to [He *et al.*, 2016a], we utilized an ensemble of 5 networks, while the backbones are ResNet101 (2 networks), ResNet50 (2 networks) and VGG (1 network). We used an ensemble of these 5 networks for proposing regions. And the union set of the proposals is processed by an ensemble for mask prediction and classification. The final F-measure is 0.905, which is the 1st place result in ICDAR2015 leaderboard. [2]. The inference time of the ensemble approach is about 1.3s for the input image with resolution (1280 x 720) in ICDAR2015.

[2] http://rrc.cvc.uab.es/?ch=4&com=evaluation&task=1

(a) ICDAR 2015 (b) RCTW-17 (c) MSRA-TD500 (d) Failure cases

Figure 8: Detection results of our proposed method on ICDAR2015(a), RCTW-17(b), MSRA-TD500(c). Some failure cases are presented in (d). Red boxes are ground-truth boxes, while green boxes ard the predict results. Bounding boxes in yellow ellipses represent the failures.

Method	ExtraData	Recall	Precision	F-measure
IncepText ensemble	✗	**0.873**	**0.938**	**0.905**
IncepText MS [3]	✗	0.843	0.894	0.868
IncepText	✗	0.806	0.905	0.853
FTSN [Dai *et al.*, 2017b]	✓	0.800	0.886	0.841
R2CNN [Jiang *et al.*, 2017]	✓	0.797	0.856	0.825
DDR [He *et al.*, 2017]	✓	0.800	0.820	0.810
EAST [Zhou *et al.*, 2017]	-	0.783	0.832	0.807
RRPN [Ma *et al.*, 2017]	✓	0.732	0.822	0.774
SegLink [Shi *et al.*, 2017]	✓	0.731	0.768	0.749

Table 2: Results on ICDAR2015 incidental scene text location task.

Method	Recall	Precision	F-measure
IncepText	**0.569**	**0.785**	**0.660**
FTSN [Dai *et al.*, 2017b]	0.471	0.741	0.576
SegLink [Shi *et al.*, 2017]	0.404	0.760	0.527

Table 3: Results on RCTW-17 text location task.

RCTW-17 is a new challenging benchmark on reading Chinese Text in images. Our proposed method achieves the F-measure of 0.66, which is a new state-of-the-art and significantly outperforms the previous methods.

On MSRA-TD500, our best performance achieves 0.790, 0.875 and 0.830 in recall, precision and F-measure, respectively. It exceeds the second best method by 0.01 in terms of F-measure.

For the original resolution (1280×720) image in ICDAR2015, our proposed method takes about $270ms$ on a Nvidia Tesla M40 GPU. The computation cost of the Inception-Text module is about $20ms$.

[3] We only use two scales [960, 1120] of the short side.

Method	Recall	Precision	F-measure
IncepText	**0.790**	0.875	**0.830**
FTSN [Dai *et al.*, 2017b]	0.771	**0.876**	0.820
SegLink [Shi *et al.*, 2017]	0.700	0.860	0.770
EAST [Zhou *et al.*, 2017]	0.674	0.873	0.761
DDR [He *et al.*, 2017]	0.700	0.770	0.740

Table 4: Results on MSRA-TD500 text location task.

Some detection samples of our proposed method are visualized in Figure 8. In ICDAR2015, the text is mainly in word level, while the text is both in word and line level in RCTW-17 and MSRA-TD500. IncepText performs well in most situations, however it still fails in some difficult cases. A main limitations is that it fails to split two words with small word spacing, which is shown at the top of Figure 8 (d). Another weakness of IncepText is that it may miss the words which are occluded as illustrated at the bottom of Figure 8 (d).

5 Conclusion

In this paper, we proposed a novel end-to-end approach for multi-oriented scene text detection based on instance-aware segmentation framework. The main idea is to design a new Inception-Text module to handle scene text which suffers from a large variance of scale, aspect ratio and orientation. Another improvement comes from using deformable PSROI pooling to handle scene text. We demonstrated its efficiency on three public scene text benchmarks. Our proposed method achieves the state-of-the-art performance in comparison with the competing methods. As for future work, we would like to combine our detection framework with recognition framework to further boost the efficiency of our model.

References

[Busta *et al.*, 2015] Michal Busta, Lukas Neumann, and Jiri Matas. Fastext: Efficient unconstrained scene text detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1206–1214, 2015.

[Dai *et al.*, 2016] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[Dai *et al.*, 2017a] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. *arXiv preprint arXiv:1703.06211*, 2017.

[Dai *et al.*, 2017b] Yuchen Dai, Zheng Huang, Yuting Gao, and Kai Chen. Fused text segmentation networks for multi-oriented scene text detection. *arXiv preprint arXiv:1709.03272*, 2017.

[He *et al.*, 2016a] 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.

[He *et al.*, 2016b] Tong He, Weilin Huang, Yu Qiao, and Jian Yao. Text-attentional convolutional neural network for scene text detection. *IEEE transactions on image processing*, 25(6):2529–2541, 2016.

[He *et al.*, 2017] 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.

[Jiang *et al.*, 2017] 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. *arXiv preprint arXiv:1706.09579*, 2017.

[Karatzas *et al.*, 2015] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 1156–1160. IEEE, 2015.

[Kingma and Ba, 2014] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Li *et al.*, 2016] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. *arXiv preprint arXiv:1611.07709*, 2016.

[Long *et al.*, 2015] 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.

[Ma *et al.*, 2017] 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.

[Neumann and Matas, 2010] 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.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[Shi *et al.*, 2017] Baoguang Shi, Xiang Bai, and Serge Belongie. Detecting oriented text in natural images by linking segments. *arXiv preprint arXiv:1703.06520*, 2017.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[Szegedy *et al.*, 2017] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284, 2017.

[Tian *et al.*, 2016] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. Detecting text in natural image with connectionist text proposal network. In *European Conference on Computer Vision*, pages 56–72. Springer, 2016.

[Xie and Tu, 2015] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

[Yao *et al.*, 2012] 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.

[Yao *et al.*, 2016] Cong Yao, Xiang Bai, Nong Sang, Xinyu Zhou, Shuchang Zhou, and Zhimin Cao. Scene text detection via holistic, multi-channel prediction. *arXiv preprint arXiv:1606.09002*, 2016.

[Zhou *et al.*, 2017] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: An efficient and accurate scene text detector. *arXiv preprint arXiv:1704.03155*, 2017.