Scene Text Detection with Supervised Pyramid Context Network

Enze Xie,[1,3*] **Yuhang Zang,**[2,3*] **Shuai Shao,**[3] **Gang Yu,**[3] **Cong Yao,**[3] **Guangyao Li**[1†]
[1]Department of Comuter Science and Technology, Tongji University
[2]School of Information and Software Engineering, University of Electronic Science and Technology of China
[3]Megvii (Face++) Technology Inc.
{xieenze, lgy}@tongji.edu.cn, yuhangzang@foxmail.com, {shaoshuai, yugang, yaocong}@megvii.com

Abstract

Scene text detection methods based on deep learning have achieved remarkable results over the past years. However, due to the high diversity and complexity of natural scenes, previous state-of-the-art text detection methods may still produce a considerable amount of false positives, when applied to images captured in real-world environments. To tackle this issue, mainly inspired by Mask R-CNN, we propose in this paper an effective model for scene text detection, which is based on Feature Pyramid Network (FPN) and instance segmentation. We propose a supervised pyramid context network (SPCNET) to precisely locate text regions while suppressing false positives.

Benefited from the guidance of semantic information and sharing FPN, SPCNET obtains significantly enhanced performance while introducing marginal extra computation. Experiments on standard datasets demonstrate that our SPCNET clearly outperforms start-of-the-art methods. Specifically, it achieves an F-measure of 92.1% on ICDAR2013, 87.2% on ICDAR2015, 74.1% on ICDAR2017 MLT and 82.9% on Total-Text.

Figure 1: Visualization of detection results and semantic segmentation feature maps. Left:The detection result with classification score and fusion score. The fused score is calculated by Re-Score mechanism. Right: The feature map for text segmentation.

Introduction

Reading text in the wild, as a fundamental task in the field of computer vision, has been widely studied. Many applications in the real world rely on accurate text localization, such as license plate recognition, autonomous driving, and document analysis. Recently, most previous works mainly focus on several challenging issues in natural scene text detection, such as multi-oriented text (Lyu et al. 2018b), large aspect ratios (Liao et al. 2018), and difficulty in separating adjacent text instances (Deng et al. 2018). However, due to the large differences in foreground text and background objects, as well as the variety of text changes in shape, color, font, orientation and scale, together with extreme illumination and occlusion, there are still many challenges to be addressed for text detection in natural scenes.

The first challenge is false positives (FP). Some specific scenarios such as autonomous driving require high precision in text detection. To the best of our knowledge, little research pays attention to false positive problem in scene text detection. Second, flexible locating text in arbitrary shape still remains challenges. Text in natural scenes can be in multi-oriented, multi-lingual or curved forms, making network difficult to distinguish FPs. Most of the existing methods are specifically designed to detect multi-oriented text and may fall short when handling with curved text. TextSnake (Long et al. 2018) uses ordered disks to represent curved text, but it still needs time-consuming and complicated post-processing.

To detect text with various forms, instance segmentation based method is adopted. Modern instance segmentation methods, such as Mask R-CNN (He et al. 2017a), are usually developed as a multi-task learning problem: (1) differentiate foreground object proposals from background and assign them with proper class labels. (2) perform regression and segmentation on each foreground proposal.

Nevertheless, simply transfer Mask-RCNN to the text detection scenario is prone to cause some problems, for the

*indicates equal contribution. This work was done when Enze Xie and Yuhang Zang were interns in Detection Group, Face++, Beijing, China.
†Corresponding author.

following two reasons: (1) **Lack of context information clues**. False positives in natural scene tend to be closely related to the surrounding scene. For instance, dishes often appear on the table, and fences usually appear in batches. However, Mask R-CNN distinguishes object in a single region of interest, which lacks global semantic information guide. Thence, it tends to cause classification errors on some objects who have similar texture information to text without the helping of context information clues. (2) **Inaccurate classification score**. The classification scores of Mask R-CNN are easily to be inaccurate when dealing with tilted text. Because for tilted text, Mask R-CNN gives classification score rudely based on horizontal proposal, while the background occupies a large proportion. Therefore, when facing tilted text, the classification score of Mask R-CNN tends to be low.

In this paper, we propose a shape robust text detector guided by semantic information. Inspired by Mask R-CNN, which can generate shape masks of objects, we use the output of the mask branch to locate the text area. Thus our method is flexible to detect text of arbitrary shapes.

In order to solve the FP problems of lacking context information clues and inaccurate classification score, we design the Text Context module and Re-Score mechanism. For Text Context module, we use the semantic segmentation branch to auxiliary guide the detection branch capturing the context information. Through compensating global semantic feature, the network discriminates FPs better. For Re-Score mechanism, we compensate activation values on segmentation map to classification score to get a fused score. When tackling with tilted text, although the classification score is relatively low, the response on the segmentation map remains strong, leading to an accurate high fused score. The Re-Score mechanism can further help to reduce FP numbers. This is because the response of FP on segmentation map is intensely weak, causing low fused score. Therefore, FPs with low scores will be more easily filtered out during inference. The visualization result of the Re-Score mechanism is shown in Fig. 1.

Compared with baseline, the proposed algorithm enhances performance significantly, while adding little computation. Furthermore, the proposed algorithm achieves an F-measure of 92.1% on ICDAR2013, 87.2% on ICDAR2015, 74.1% on ICDAR2017MLT and 82.9% on Total-Text, outperforming previous state-of-the-art algorithms in various kinds of scene text benchmarks (e.g., horizontal, oriented, multi-lingual and curved).

The contributions of this work are three-fold: (1) We propose Text Context module and Re-Score mechanism, which can effectively suppress false positives. (2) The proposed method can flexibly detect text in various shapes, including horizontal, oriented and curved text. (3) The proposed algorithm significantly outperforms state-of-the-art methods on several benchmarks containing text instances of different forms.

Related Work

Scene text detection, as one of the most important problems in computer vision, has been extensively studied. Most of the previous deep learning methods can be roughly divided into two branches: segmentation-based text detection and regression-based text detection.

Mainstream segmentation-based approaches are inspired by fully convolutional networks (FCN) (Long, Shelhamer, and Darrell 2015). (Zhang et al. 2016) first uses FCN to extract text blocks and detect character candidates from those text blocks with MSER. (Yao et al. 2016) treats one text region as consisting of three parts:text/non-text, character classes, and character linking orientations, then use them as labels for FCN. PixelLink (Deng et al. 2018) performs text/non-text and link prediction on an input image, then adds some post-processing to get text box and filter noise. PSENET (Li et al. 2018) finds text kernels and uses progressive scale expansion to position text boundary. (Peng et al. 2017b) argues that using large kernel can help boosting semantic segmentation performance. The main difference between these methods is the generation of different labels for the text. Segmentation-based approaches often need time-consuming post-processing steps while obtained performance is still unsatisfying.

General object detection and instance segmentation methods, e.g., Faster R-CNN (Ren et al. 2015), SSD (Liu et al. 2016) and FCIS (Li et al. 2016), are widely applied to text detection. TextBoxes (Liao et al. 2017) modifies anchors and kernels of SSD to detect large-aspect-radio scene text. EAST (Zhou et al. 2017) adopts FCN to predict a text score map and a final box for each point in the text region. RRD (Liao et al. 2018) extracts two types of feature for classification and regression respectively for long text line detection. Based on Faster R-CNN, (Ma et al. 2018) adds rotation to both anchors and RoIPooling to detect multi-oriented text region. IncepText (Yang et al. 2018) uses FCIS to detect multi-oriented text boxes from the perspective of instance segmentation.

However, most of the above methods lack attention to false positives problem in scene text detection, and these methods are often not flexible enough to adapt to arbitrary shapes of text detection. In this paper, we devise a pipeline that uses deep supervised semantic information to guide Mask R-CNN finding text area accurately and suppress false positives efficiently. The model combines instance segmentation with semantic segmentation and allows training in an end-to-end manner. Moreover, the proposed method can flexibly detect text of arbitrary shape. Results on several benchmarks show that our method significantly surpasses all previous methods by an obvious gap in performance.

Proposed method

Our pipeline is composed of two key parts: a Text Context module and a post Re-Score mechanism. The basis of this pipeline is based on Mask R-CNN. The text-context module contains two modules: a text attention module and a deep feature fusion module. This section is organized as follows: In Section 3.1, we examine the method for text detection based on Mask R-CNN. In Section 3.2, we illustrate the effectiveness of the Text Context module in suppressing false positives. In Section 3.3, we show the irrationality of the original scoring method and propose a method of Re-Score

to further suppress FPs. In Section 3.4, we explain the loss function design.

Mask R-CNN

Why Mask R-CNN? Mask R-CNN is the state of the art in instance segmentation. Most of the winners in MS COCO instance segmentation challenge are based on Mask R-CNN. A recent work (Lyu et al. 2018a) also uses Mask R-CNN for end-to-end text detection and recognition. Hence Mask R-CNN makes a strong baseline to compare against.

Label Generation The ground truth of text instance is exemplified in Fig. 3. Different from common instance segmentation datasets, pixel-level text/non-text annotations are not provided. We treat the pixels in the polygon as text, and the pixels outside the polygon as non-text, then we get an instance of the text area. The minimum bounding horizontal rectangle of the polygon will be treated as a bounding box. We generate the global binary map in the same way as the instance generation.

Mask R-CNN architecture The overall architecture of our proposed method is presented in Fig. 2. Our network is composed of five parts: feature pyramid network (FPN), region proposal network(RPN), R-CNN branch, mask prediction branch and global text segmentation prediction branch. Feature Pyramid Network (FPN) is a feature fusion structure widely used in current mainstream detection models. FPN uses a top-down architecture with lateral connections to build an in-network feature pyramid from a single-scale input. Region Proposal Network (RPN) generates bounding boxes likely to contain an object as proposals. Through RoiAlign, all proposals are resized to 7×7 for R-CNN branch and 14×14 for mask prediction branch. The global text segmentation branch acts on each stage of the FPN to generate a semantic segmentation map of the text.

Text Context Module

Suppressing false positives is a challenging issue for general object detection and text detection. In natural scenes, some regular objects, such as discs, fences, etc., are easily detected as text by the detection network. Mask R-CNN uses region of interests (ROIs) to classify whether the proposal is text or background. However, the text region classification is performed with features extracted from only one region of interest. Since false positives in natural scenes often do not appear unexpectedly, such as plates are more likely appear on the table, introducing contextual information helps the network extract more discriminative features and accurately classify proposals. Our Text Context Module (TCM) is composed of two sub-modules: Pyramid Attention Module (PAM) and Pyramid Fusion Module (PFM). The feature maps are feed to TCM, which produces text segmentation as output.

Pyramid Attention Module Our pyramid attention module is inspired by SSTD (He et al. 2017b). We additionally add a global text segmentation branch after FPN from stage2 to stage5. It generates a saliency map of pixel-level text/non-text regions for each FPN layer. The attention module and the fusion module share a branch, named text context module, including two 3×3 convolutional layers and one 1×1 convolutional layer. The output saliency map includes two channels, which means text/non-text map. We enhance the saliency map and use it to activate the text area on the feature map. Specifically, take stage2 as an example, giving an input sample of 512×512, the feature map $S_2 \in R^{128\times128\times256}$. The generation of saliency map is as follows:

$$map = Text\_Context\_Module(S_2) \tag{1}$$

$$saliency\_map = e^{Softmax(map)} \tag{2}$$

where Text Context module generates the saliency map with 2 channels. Then after the channel-wise softmax, we obtain the text saliency map. Through the Exponential activation, the saliency map is enhanced, that is, the response gap in text/non-text areas becomes larger. The saliency map will act on the feature map as follows:

$$saliency\_map^* = Broadcast(saliency\_map) \tag{3}$$

$$S_2^* = saliency\_map^* \odot S_2 \tag{4}$$

where saliency_map is broadcast to the same 256 channel as S_2, and "$\odot$" represents the pixel-by-pixel multiplication of the two maps S_2 and $saliency\_map^*$.

Pyramid Fusion Module Next we introduce the pyramid fusion module. The PFM combines detection feature with the deep supervised semantic feature, makes the network more discriminative to distinguish text from non-text. Specifically, semantic segmentation examines text from the perspective of a single pixel and determines the text region by combining the information of surrounding pixels, and the detection classifies the text region by ROIs. There is a natural complementary relationship between the two branches.

After first 3×3 convolutional layers of Text Context module, we get the feature map(GTF) of global text segmentation. These features capture complementary information like context, semantic segmentation of background and of text. Both computer vision (Divvala et al. 2009) and cognitive psychology (Oliva and Torralba 2007) research show that identifying the local surrounding of an object helps to better identify itself. This is because the category of object are often correlated with surrounding stuff, e.g. discs often appear on the table. Although there is only textual annotation information, this encoding method allows the network to implicitly learn more discriminative semantic information. Introducing it into the original feature map makes Mask R-CNN performing stronger on the classification task. The specific details are as follows:

$$GTF = Conv_{3\times3}(S_2) \tag{5}$$

$$\hat{S_2} = S_2^* + GTF \tag{6}$$

where the $Conv_{3\times3}$ is the first Conv layer in Text Context module and GTF represent global text feature. Then "+" represents element-wise addition operation.

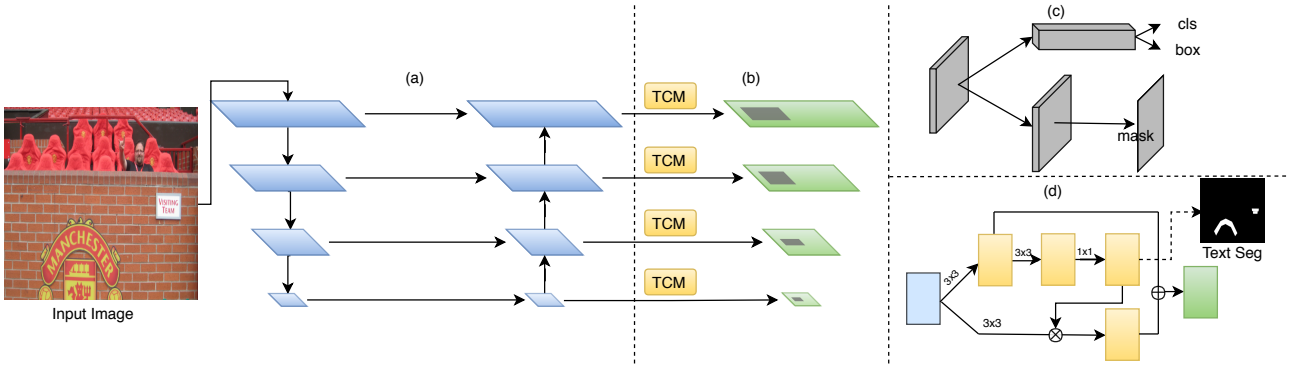


Figure 2: The architecture of our method. (a) The Feature Pyramid Network (FPN) architecture. (b) Pyramid Feature fusion via TCM. (c) Mask R-CNN branch for text classification, bounding box regression and instance segmentation. (d) The proposed Text-Context Module(TCM). Dotted line indicates the text semantic segmentation branch. The text segmentation map is upsampled to the input image size and calculates the loss with Ground Truth.

Figure 3: Ground truth. Left: Image sample with green bounding box and yellow polygon. Right: Corresponding binary text segmentation map.

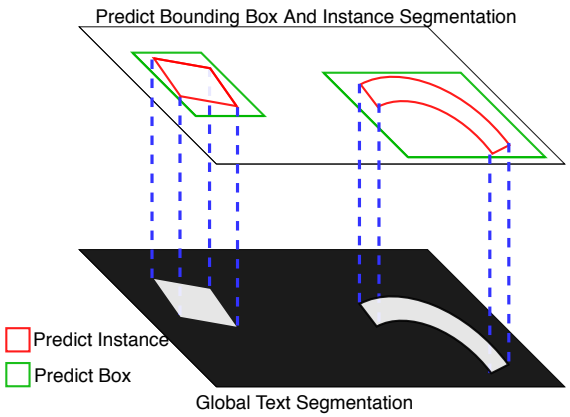


Figure 4: Overview of the Re-Score Mechanism. Upon:The predicted text boxes and instances of the input images; Bottom:The global text segmentation map output from TCM. For each text instance, we project them onto the segmentation map and calculate the activation value of the projected area.

Re-Score Mechanism

For standard Mask R-CNN inference processing, the predicted top-K(e.g., 1000) bounding boxes are sorted by the classification confidence, then after standard NMS processing, Up to top-M(e.g., 300) bounding boxes with highest classification confidence are retained. These bounding boxes are feed to Mask R-CNN as proposals to generate predicted text instance maps. This method treats one horizontal bounding box's classification confidence as the score, then artificially sets a threshold to filter out background boxes. However, this method will filter out some true positives with low scores, because if a horizontal bounding box encloses a titled text instance, it also accompanies a lot of background information. At the same time, some FPs with relatively high confidence will be retained.

We re-assign scores for each text instance. The visualization diagram is shown in Fig. 4. The fused score of text instance is composed of two parts: classification score(CS) and instance score(IS). Formally, the fused score for the ith proposal, given the predicted 2-class scores CS = $\{s_{i0}^{cs}, s_{i1}^{cs}\}$ and IS = $\{s_{i0}^{is}, s_{i1}^{is}\}$ is computed via the following softmax function:

$$s_i = \frac{e^{(s_{i1}^{cs}+s_{i1}^{is})}}{e^{(s_{i1}^{cs}+s_{i1}^{is})} + e^{(s_{i0}^{cs}+s_{i0}^{is})}} \tag{7}$$

where CS is directly obtained by Mask R-CNN classification branch, and IS is the activation value of the text instance on the global text segmentation map. In details, for each text instance, it is projected onto text segmentation map, containing $P_i = \{p_i^1, p_i^2 ... p_i^n\}$, and the mean of p_i in the text instance area is calculated:

$$s_{i1}^{cs} = \frac{\sum_j p_i^j}{N} \tag{8}$$

where P_i is the set of the pixels' value of ith text instance on text segmentation map. The fused score combines the classification score with the instance score, which can reduce the FP confidence effectively, because FP instances tend to have weaker response than text on the segmentation map.

This mechanism is also more friendly for titled text, because the titled text instance also has a strong response on the segmentation map, high instance score will compensate for low classification score.

Loss Function Design

Similar to Mask R-CNN, our network includes multi-task. Following the loss function design of Mask R-CNN, we additionally add a global text segmentation loss based on it. The loss expression is as follows:

$$L = L_{rpn} + \lambda_1 \cdot L_{cls} + \lambda_2 \cdot L_{box} + \lambda_3 \cdot L_{mask} + \lambda_4 \cdot L_{gts} \quad (9)$$

where L_{rpn}, L_{cls}, L_{box} and L_{mask} are the standard loss in Mask R-CNN. The L_{gts} is used to optimize global text segmentation, defined as :

$$L_{gts} = \frac{1}{N} \sum_i -\log(\frac{e^{p_i}}{\sum_j e^{p_j}}) \quad (10)$$

The L_{gts} is Softmax loss, where p is the output prediction of the network.

Multitask learning is the process of learning useful representations of multiple complementary tasks from the same input, and has been found to improve the performance of both tasks. This method enables the network to learn text detection and global text segmentation by end-to-end joint training, allowing gradients from two tasks to influence shared feature maps.

Experiments

We evaluate our approach on four standard benchmarks: ICDAR2013, ICDAR2015, ICDAR2017 MLT, Total-Text, and compare with other state-of-the-art methods.

Datasets

The datasets used for the experiments in this paper are briefly introduced below:

SynthText (Gupta, Vedaldi, and Zisserman 2016) is a synthetically generated dataset composed of 800000 synthetic images. We use the dataset with word-level labels to pre-train our model.

ICDAR2017 MLT (Nayef et al. 2017) is a dataset focuses on multi-oriented, multi-scripting, and multi-lingual aspects of scene text. It consists of 7200 training images, 1800 validation images, and 9000 test images. Image annotations are labeled as word-level quadrangles. We use both training set and validation set to train our model.

ICDAR2015 (Karatzas et al. 2015) is a dataset proposed for incidental scene text detection. There are 1000 training images and 500 tests images with annotations labeled as word-level quadrangles.

ICDAR2013 (Karatzas et al. 2013) is a dataset points at horizontal text in the scene. It contains 229 training images and 233 testing images with only horizontal texts.

Total-Text (Ch'ng and Chan 2017) is a newly-released benchmark for curved text detection. The dataset is split into training and testing sets with 1255 and 300 images, respectively.

Implementation Details

Training We set hyper-parameters mainly following Mask R-CNN. Our base-model is ResNet50 and pre-trained on ImageNet. All new layers are initialized with a zero-mean Gaussian distribution with standard deviation 0.001. We use Adam as optimizer with batch size 16, momentum 0.9 and weight decay 1e-4 in training. Similar to (Yu et al. 2018), we apply the "poly" learning rate strategy in which the initial rate is multiplied by $(1 - \frac{iter}{max\_iter})^{power}$ each iteration with power 0.9. The initial learning rate is 2×10^{-3} for all experiments. We first adopted the warmup strategy in (Peng et al. 2017a), then we found without warmup the net can still convergence fast. The network only takes 6h and 1h to complete training when use 8 GPUs. The aspect ratios of anchors are set to $1/5, 1/2, 1, 2, 5$ for all experiments.

Data Augmentation We follow the data augmentation strategy of Mask R-CNN. Short edges of the images are randomly resized to three scales (640, 720, 800). Then each image is randomly flipped with a probability of 0.5.

Post Processing Our post processing is simple. We re-score all text instances, then find the minimum bounding rectangle for each text instance. Finally a polygon NMS is utilized to suppress redundant boxes. Methods like $minAreaRect$ in OpenCV can be applied to obtain the bounding boxes of text instances as the final detection result.

Ablation Study

To verify the effectiveness of our approach, we do a series of comparative experiments on the ICDAR2017 MLT validation set. These experiments mainly focus on evaluating two essential methods in our model: Text Context module(TCM) and Re-Score mechanism(RS). Table 1 summarizes the results of our models with different settings on ICDAR2017 MLT.

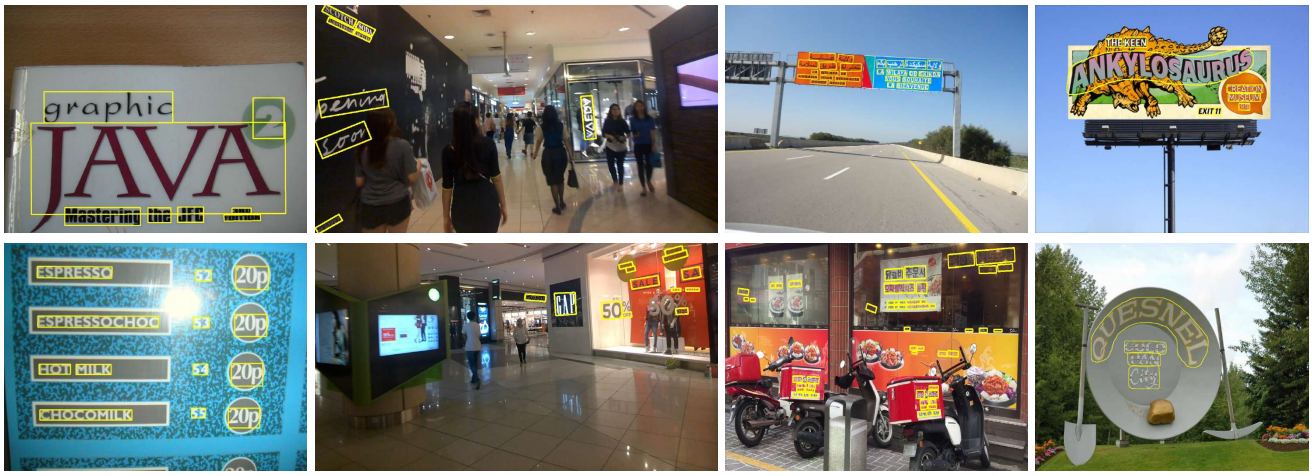
Method	Recall	Precision	F-measure
Baseline	73.4	76.2	74.7
Ours+TCM	73.4	80.3	76.8
Ours+TCM+RS	73.4	84.2	**78.5**

Table 1: Effectiveness of several modules on ICDAR2017 MLT incidental scene text location task.

False positive problems often appear in complex natural scenes. The MLT dataset is composed of complete scene images which come from 9 languages. According to our statistics, the smallest text box size on the MLT is less than 20 pixels, and the largest is more than 3000 pixels. So size range of the text box is very different. To the best of our knowledge, it is the most challenging public scene text benchmark, hence the experiment results in MLT are convincing. The detailed comparison is given in the following.

Baseline Mask R-CNN architecture without Text Context Module and Re-Score Mechanism.

Text Context Module Compared with baseline, the Text Context module achieves an improvement of 4.1 percents on precision while keeping the recall identical. This implies that



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

Figure 5: Qualitative results of the proposed algorithm. (a) ICDAR2013. (b) ICDAR2015. (c)ICDAR2017. (d) Total-Text.

the TCM helps network extract more discriminative features of text/non-text and reduced the number of FPs.

Re-Score Mechanism In the post-processing stage, we use our proposed re-score mechanism to re-rank the scores of all text instances during inference. Table 1 shows our re-score mechanism can further improve precision of 3.9% based on TCM. This brings in total 3.8% F-measure of revenue compared with baseline. The experimental result proves that the Re-Score mechanism can further suppress FPs with weakly response on the global text segmentation in post-processing stage.

Results on Scene Text Benchmarks

Detecting MultiLingual Text We first pretrain the proposed network on SynthText for one epoch then fine-tuned on MLT 9000 train and val images for 40 epochs. With single scale of 848(short edge), our proposed method achieves an F-measure of 70.0%, outperforming state of the art methods over 3%. Since there are many small words on the MLT, we apply a simple multi-scale test method with scale $\in [720, 1920]$. By merging the results of two scales, the F-measure is 74.1%, which outperforms all competing methods by at least 1.7%. To our best knowledge, this is the best reported result in literature. The result is shown in Table 2.

Detecting Oriented Text On ICDAR2015, we use pretrained model from MLT and fine-tune another 40 epochs. The comparison with the state of the art results on ICDAR2015 dataset is given in Table 3. All setting are same as MLT except we only use single scale test. Experimental results show the results of our method surpasses the state of the art results by more than 1.5% percents with single scale setting. Moreover, Fig. 6 shows our methods can suppress false positives effectively compared with prior arts.

Detecting Horizontal Text On ICDAR2013, the proposed model is pre-trained from MLT and fine-tune on 299 training images for another 40 epochs. All test settings are the same as ICDAR2015. Although our method is specifically designed for text detection of arbitrary shapes, our method also shows superiority in horizontal text detection compared to prior arts. Table 4 shows the experiment results of different methods. Similarly, in ICDAR2013 dataset, our approach achieves the state of the art result at 92.1%, experiments prove the effectiveness of our method.

Method	Recall	Precision	F-measure
TH-DL(Nayef et al. 2017)	34.8	67.8	46.0
SARI FDU RRPN V1(Nayef et al. 2017)	55.5	71.2	62.4
Sensetime OCR(Nayef et al. 2017)	69.4	56.9	62.6
SCUT DLVClab1(Nayef et al. 2017)	54.5	80.3	65.0
Lyu *et al.(Lyu et al. 2018b)*	55.6	83.8	66.8
Lyu *et al.*(Lyu et al. 2018b)*	70.6	74.3	72.4
Baseline	62.2	69.2	65.5
Ours	66.9	73.4	70.0
Ours*	68.6	80.6	**74.1**

Table 2: Effectiveness of several methods on ICDAR2017 MLT incidental scene text location task. * means multi scale test.

Detecting Curved Text We evaluate the ability of our model to detect curved text on Total-Text dataset. Similar to the above training methods, we use the MLT pretrained weights to initialization model and fine-tune on Total-Text for 40 epochs. All test settings are the same as ICDAR2015 and ICDAR2013. Our method is shape robust for text detection. The proposed method can be flexibly applied to different types of scene text detection datasets without special modifications. Experimental results show that our method surpasses prior art methods. The detail results are shown in Table 5. Note that the results of SegLink and EAST are referenced from TextSnake.

Method	Recall	Precision	F-measure
CTPN(Tian et al. 2016)	51.6	74.2	60.9
SegLink (Shi, Bai, and Belongie 2017)	76.8	73.1	75.0
MCN(Liu et al. 2018)	72.0	80.0	76.0
SSTD(He et al. 2017b)	73.0	80.0	77.0
WordSup*(Hu et al. 2017)	77.0	79.3	78.2
EAST*(Zhou et al. 2017)	78.3	83.3	80.7
Lyu *et al.(Lyu et al. 2018b)*	70.7	94.1	80.7
DeepReg(He et al. 2017c)	80.0	82.0	81.0
RRD*(Liao et al. 2018)	80.0	88.0	83.8
TextSnake(Long et al. 2018)	80.4	84.9	82.6
PixelLink(Deng et al. 2018)	82.0	85.5	83.7
FTSN(Dai et al. 2017)	80.0	88.6	84.1
IncepText(Yang et al. 2018)	80.6	90.5	85.3
Baseline	83.8	87.4	85.5
Ours	85.8	88.7	**87.2**

Table 3: Effectiveness of several methods on ICDAR2015. * means multi scale test.

Method	Recall	Precision	F-measure
CTPN(Tian et al. 2016)	83.0	83.0	88.0
TextBoxes(Liao et al. 2017)	74.0	88.0	81.0
SegLink (Shi, Bai, and Belongie 2017)	83.0	87.7	85.3
MCN(Liu et al. 2018)	87.0	88.0	88.0
SSTD(He et al. 2017b)	86.0	89.0	88.0
WordSup*(Hu et al. 2017)	88.0	93.0	90.0
Lyu *et al.(Lyu et al. 2018b)*	79.4	93.3	85.8
DeepReg(He et al. 2017c)	81.0	92.0	86.0
RRD(Liao et al. 2018)	75.0	88.0	81.0
PixelLink*(Deng et al. 2018)	87.5	88.6	88.1
Baseline	88.1	91.0	89.6
Ours	90.5	93.8	**92.1**

Table 4: Effectiveness of several methods on ICDAR2013. * means multi scale test.

Method	Recall	Precision	F-measure
SegLink (Shi, Bai, and Belongie 2017)	23.8	30.3	26.7
EAST(Zhou et al. 2017)	36.2	50.0	42.0
DeconvNet (Ch'ng and Chan 2017)	40.0	33.0	36.0
TextSnake(Long et al. 2018)	74.5	82.7	78.4
FTSN(Dai et al. 2017)	78.0	84.7	81.3
Baseline	80.5	81.5	81.0
Ours	82.8	83.0	**82.9**

Table 5: Effectiveness of several methods on Total-Text dataset. Note that EAST and SegLink were not fine-tuned on Total-Text. Therefore their results are included only for reference.

Conclusion

In this work, we have presented a shape robust text detector that can detect text with arbitrary shapes. It is an end-to-end trainable framework with semantic segmentation guidance. We effectively alleviate the false positive problem via introducing context semantic information and re-score mechanism for all predicted text instances. By sharing convolutional features, the text segmentation branch is nearly cost-free. The results on different scene text benchmarks demonstrate the effectiveness and generalization of our approach.

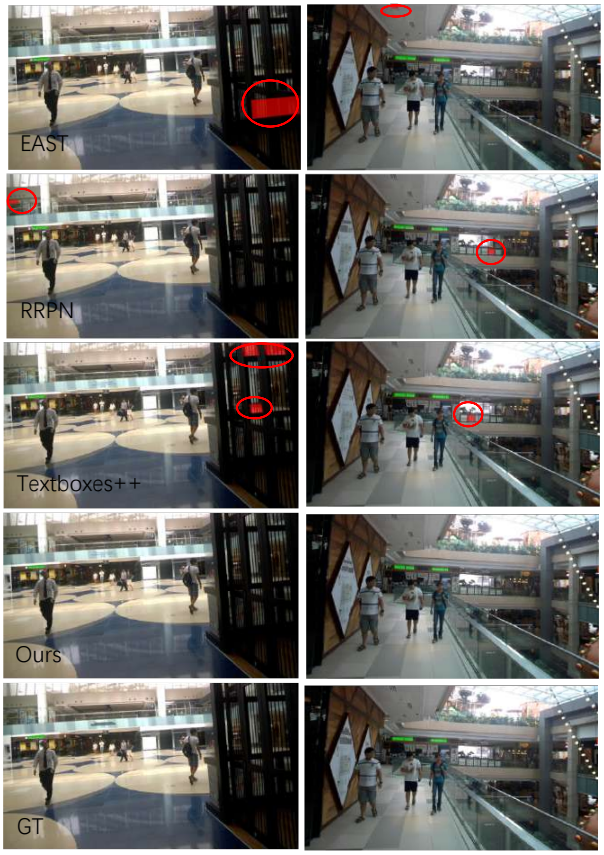


Figure 6: Qualitative detection results of EAST, RRPN(Ma et al. 2018), TextBoxes++(Liao, Shi, and Bai 2018) and our method. The green and red regions represent true positive and false positive results respectively. Visualizations are captured from the ICDAR official online evaluation system (http://rrc.cvc.uab.es/?ch=4&com=evaluation&task=1).

In the future, we are interested in multiple directions as below: (1) We will attempt to integrate the Re-Score mechanism into the network in an end-to-end manner. (2) We are interested in exploring our method on other multi-oriented or curved object detection task, such as an aerial scene. (3) We will investigate more efficient fast text detection networks that running on mobile phones.

Acknowlegement

This work was supported by the National Natural Science Foundation of China under Grant number 61771346. Special thanks Mr. Mengxiao Lin in Megvii base-model group for all his kindness and great help to us.

References

Ch'ng, C. K., and Chan, C. S. 2017. Total-text: A comprehensive dataset for scene text detection and recognition. In *ICDAR*. IEEE.

Dai, Y.; Huang, Z.; Gao, Y.; Xu, Y.; Chen, K.; Guo, J.; and Qiu, W. 2017. Fused text segmentation networks for multi-oriented scene text detection. *ICPR*.

Deng, D.; Liu, H.; Li, X.; and Cai, D. 2018. Pixellink: Detecting scene text via instance segmentation. *AAAI*.

Divvala, S. K.; Hoiem, D.; Hays, J. H.; Efros, A. A.; and Hebert, M. 2009. An empirical study of context in object detection. In *CVPR*.

Gupta, A.; Vedaldi, A.; and Zisserman, A. 2016. Synthetic data for text localisation in natural images. In *CVPR*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017a. Mask r-cnn. In *ICCV*.

He, P.; Huang, W.; He, T.; Zhu, Q.; Qiao, Y.; and Li, X. 2017b. Single shot text detector with regional attention. In *The IEEE International Conference on Computer Vision (ICCV)*.

He, W.; Zhang, X.-Y.; Yin, F.; and Liu, C.-L. 2017c. Deep direct regression for multi-oriented scene text detection. *ICCV*.

Hu, H.; Zhang, C.; Luo, Y.; Wang, Y.; Han, J.; and Ding, E. 2017. Wordsup: Exploiting word annotations for character based text detection. In *ICCV*.

Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. Icdar 2013 robust reading competition. In *ICDAR*.

Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V. R.; Lu, S.; et al. 2015. Icdar 2015 competition on robust reading. In *ICDAR*.

Li, Y.; Qi, H.; Dai, J.; Ji, X.; and Wei, Y. 2016. Fully convolutional instance-aware semantic segmentation. *arXiv preprint arXiv:1611.07709*.

Li, X.; Wang, W.; Hou, W.; Liu, R.-Z.; Lu, T.; and Yang, J. 2018. Shape robust text detection with progressive scale expansion network. *arXiv preprint arXiv:1806.02559*.

Liao, M.; Shi, B.; Bai, X.; Wang, X.; and Liu, W. 2017. Textboxes: A fast text detector with a single deep neural network. In *AAAI*.

Liao, M.; Zhu, Z.; Shi, B.; Xia, G.-s.; and Bai, X. 2018. Rotation-sensitive regression for oriented scene text detection. In *CVPR*.

Liao, M.; Shi, B.; and Bai, X. 2018. Textboxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *ECCV*.

Liu, Z.; Lin, G.; Yang, S.; Feng, J.; Lin, W.; and Goh, W. L. 2018. Learning markov clustering networks for scene text detection. *CVPR*.

Long, S.; Ruan, J.; Zhang, W.; He, X.; Wu, W.; and Yao, C. 2018. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *ECCV*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*.

Lyu, P.; Liao, M.; Yao, C.; Wu, W.; and Bai, X. 2018a. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *ECCV*.

Lyu, P.; Yao, C.; Wu, W.; Yan, S.; and Bai, X. 2018b. Multi-oriented scene text detection via corner localization and region segmentation. In *CVPR*.

Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; and Xue, X. 2018. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*.

Nayef, N.; Yin, F.; Bizid, I.; Choi, H.; Feng, Y.; Karatzas, D.; Luo, Z.; Pal, U.; Rigaud, C.; Chazalon, J.; et al. 2017. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *ICDAR*. IEEE.

Oliva, A., and Torralba, A. 2007. The role of context in object recognition. *Trends in cognitive sciences*.

Peng, C.; Xiao, T.; Li, Z.; Jiang, Y.; Zhang, X.; Jia, K.; Yu, G.; and Sun, J. 2017a. Megdet: A large mini-batch object detector. *CVPR*.

Peng, C.; Zhang, X.; Yu, G.; Luo, G.; and Sun, J. 2017b. Large kernel matters–improve semantic segmentation by global convolutional network. In *CVPR*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

Shi, B.; Bai, X.; and Belongie, S. 2017. Detecting oriented text in natural images by linking segments. *CVPR*.

Tian, Z.; Huang, W.; He, T.; He, P.; and Qiao, Y. 2016. Detecting text in natural image with connectionist text proposal network. In *ECCV*.

Yang, Q.; Cheng, M.; Zhou, W.; Chen, Y.; Qiu, M.; and Lin, W. 2018. Inceptext: A new inception-text module with deformable psroi pooling for multi-oriented scene text detection. *IJCAI*.

Yao, C.; Bai, X.; Sang, N.; Zhou, X.; Zhou, S.; and Cao, Z. 2016. Scene text detection via holistic, multi-channel prediction. *arXiv preprint arXiv:1606.09002*.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. *ECCV*.

Zhang, Z.; Zhang, C.; Shen, W.; Yao, C.; Liu, W.; and Bai, X. 2016. Multi-oriented text detection with fully convolutional networks. In *CVPR*.

Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; and Liang, J. 2017. East: an efficient and accurate scene text detector. In *CVPR*.