# Strong-Background Restrained Cross Entropy Loss for Scene Text Detection

Randong Huang[1,2], Bo Xu[1*]
[1]*Institute of Automation, Chinese Academy of Sciences*, Beijing, China
[2]*University of Chinese Academy of Sciences*, Beijing, China
{huangrandong2015, xubo}@ia.ac.cn

***Abstract*—In this paper, we investigate the issue of class imbalance in scene text detection. Class Balanced Cross Entropy (CBCE) loss is often adopted for addressing this imbalance problem. We find that CBCE excessively restrains the backward gradients of background. Negative samples own extremely small weights which are offered by CBCE during training of text detectors. These tiny weight values lead to insufficient learning of background. As a result, the CBCE-based text detection methods only can achieve sub-optimal performance.**

**We propose a novel loss function, Strong-Background Restrained Cross Entropy (SBRCE), to deal with the disadvantage in CBCE. Specifically, SBRCE effectively down-weights the loss assigned to the strong background which means well-classified negative samples. Our SBRCE can make training focused on all positive samples and weak background(i.e., hard-classified negative samples). Moreover, it can prevent the enormous amount of strong background from overwhelming text detectors during training. Experimental results show that the proposed SBRCE can improve the performance of the efficient and accurate scene text detector (EAST) by F-score of 3.3% on ICDAR2015 dataset and 1.12% on MSRA-TD500 dataset, without sacrificing the training and testing speed of EAST.**

## I. Introduction

Scene text detection extracts text information from natural scene images and has obtained an increasing amount of attention in computer vision community. It plays a significant role in many computer applications such as automatic driving, image retrieval, scene understanding and product search. With the tremendous advance of object detection like Fast R-CNN [1], Faster R-CNN [2], YOLO [3] and SSD [4] and instance segmentation such as FCIS [5], MNC [6] and Mask R-CNN [7], many outstanding approaches concentrated on scene text detection are successively put forward by regarding text words or lines as objects. These novel text detection algorithms can be categorized into three groups: (1) Region proposal based scene text detection methods [8] [9] [10] [11] [12] [13], which apply state-of-the-art object detection approaches [4] [2] to regressing offset values from preestablished default proposals to the ground truth boxes. (2) Regression based scene text detection methods [14] [15] which output a score map and the corresponding offsets. The confidence of a pixel on the score map indicates its probability of being text. The offsets are the regression distances from one pixel location to its ground truth. (3) Instance segmentation based scene text detection [16] [17] [18] which directly extracts individual text instances from an input image. And then a minimal area rectangle algorithm is used to obtain oriented boxes as the final detection results.

Like object detection, class imbalance also emerges in scene text detection. As we all know, the root cause of this imbalance is that the area of background on the image is much larger than that of foreground. To address this problem, researchers have come up with many elegant methods such as OHEM [19], CBCE [20] [15] and Focal Loss [21]. These balancing sample techniques have already acquired impressive performance on various benchmarks in object detection and text detection fields.

However, the above-mentioned techniques also have their drawbacks. OHEM and Focal Loss contain several hyperparameters. These hyperparameters do not have specific ranges. So it is tiresome that the researchers have to spend an immense amount of time and effort to adjust hyperparameters for finding their optimal values. CBCE calculates one balancing factor between positive and negative samples to downscale standard cross entropy loss. However, CBCE does not differentiate between easy/hard negative examples. As a result, scene text detection adopting CBCE merely obtains sub-optimal performance.

In this paper, we propose a new loss function named SBRCE that serves as a more effective substitute to previous techniques for handling class imbalance. Our loss function also applies the balancing factor between positive and negative samples to down-weight cross entropy loss of negative samples, but only for strong background. The balancing factor is the same as that of CBCE. To seek out the strong background, we define a manually adjustable parameter named as $c\_strongbg$ whose value is in the range of [0, 1]. Since the $c\_strongbg$ parameter has a certain range, it is easier to find its optimal value. If the confidence of one negative sample is less than or equal to $c\_strongbg$, this negative sample is classified as strong background. The proposed SBRCE loss can restrain the loss of strong negative samples and rapidly focuses training of text detectors on the weak background and all positive samples. To validate the effectiveness of our SBRCE loss, we replace the CBCE in EAST [15] by the proposed loss, which will not sacrifice the training and testing speed of EAST. Our proposed loss increases the performance of EAST by the F-score of 3.3% on ICDAR2015 [22] dataset and 1.12% on MSRA-TD500 [23] dataset.

In summary, this study makes two main contributions:

*Corresponding author: Bo Xu (xubo@ia.ac.cn)

978-1-7281-2009-6/$31.00 ©2019 IEEE paper N-19894.pdf

(1) In order to better handle the class imbalance in scene text detection, we design a new loss function, namely SBRCE. This novel loss function can be seen as a variant of CBCE. SBRCE aims to balance the positive and negative samples in a straightforward way. Moreover, it only has one hyperparameter $c\_strongbg$ needed to be adjusted. The $c\_strongbg$ can be easily set because its value is between 0 and 1.

(2) We experimentally demonstrate that such a SBRCE loss can improve the performance of scene text detectors. Meanwhile, it can be easily implemented with only a few lines of code. This novel loss does not sacrifice the training and testing speed of EAST. To validate the superiority of SBRCE, we also adopt OHEM and Focal Loss for score map [15] in EAST, respectively. OHEM is only applied to negative samples. From the experimental results, SBRCE can obtain the best performance.

The rest of this paper is organized as follows: In Section II a brief review of scene text detection and class imbalance methods is given. In Section III we introduce the details of SBRCE loss. In Section IV we present the experimental results on two benchmarks, which show the effectiveness of proposed SBRCE loss. And in Section V we conclude this paper.

## II. RELATED WORK

As a significant task in the computer vision, scene text detection has been extensively studied for a long time. Numerous outstanding and effective methods [24] [25] [26] [27] [28] [20] have been investigated. Some methods pay attention to detecting horizontal or approximately horizontal text, while some recent approaches focus on multi-oriented scene text detection. Below we briefly introduce the related studies.

**CTPN.** CTPN [29] presents a novel text detection method. It firstly decomposes the text into many fine-scale text proposals, which can be detected directly on feature maps of Convolutional Neural Network (CNN). Then these fine-scale text proposals are combined into text lines according to special rules. Finally, a side-refinement approach to refine the text line bounding boxes is also needed in CTPN.

**SSTD.** SSTD [30] is modified from SSD to capture arbitrary oriented text in natural images. SSTD proposes a fancy attention mechanism which employs an automatically learned attention map to suppress background inference and highlight text region. To work reliably on multi-scale text, the authors design a hierarchical inception module to aggregate multi-scale inception features effectively.

**SegLink.** SegLink [10] is an also multi-oriented text detection approach whose main innovation is to disassemble text into two locally detectable elements, namely segments and links. A segment represents an oriented quadrangle enclosing a part of one word or text line, while a link indicates whether or not two neighboring segments belong to the same word or text line. This method employs a fully convolutional network to detect segments and links at multiple scales. The final detection outputs are generated by combining segments connected by links.

**DDR.** Deep Direct Regression (DDR) [14] detects the scene text by predicting the relative offsets from a given reference point. It does not need to design default boxes to match ground truth boxes. DDR takes advantage of a fully convolutional network to regress the oriented quadrangles of text. DDR consists of a segmentation branch predicting the text presence and a regression branch predicting final box for each word or text line.

**PiexlLink.** PixelLink [17] is based on instance segmentation for obtaining oriented text quadrilateral. This method trains a CNN to get two kinds of pixel-wise predictions, which are text/non-text prediction and link prediction. The concept of this link is similar to the link in SegLink. This link is to indicate whether or not two adjacent pixels lie within the same text instance (i.e., word or text line). Finally, it is essential to use the links to assemble the pixels into a Connected Components (CC), with each CC representing a detected word or text line. The final detection bounding boxes of text can be acquired by applying $minAreaRect$ in OpenCV to CCs directly.

**Class Imbalance.** Both outstanding object detection methods, like YOLO and SSD, and some novel scene text detection approaches, like EAST and IncepText [18], have to accept the truth of class imbalance during training. The imbalance between foreground and background class causes two problems: (1) training is insufficient as there are too many strong negative samples, which have few useful learning signals; (2) the strong negative samples can overwhelm training and lead to sub-optimal models. This class imbalance is addressed by some novel methods such as OHEM, CBCE and Focal Loss. OHEM automatically selects a certain number of hard samples to train the text detection network more effectively and efficiently. CBCE directly calculates a balancing factor between positive and negative samples to downscale standard cross entropy loss straightforwardly. CBCE applys this balancing factor to decreasing the weights of backward gradients of all negative samples. Focal Loss is proposed to reshape the standard cross entropy loss to alleviate this class imbalance situation by down-weighting the loss of easy examples. In contrast, we show that our proposed SBRCE naturally deals with this imbalance problem and makes the training concentrated on weak negative samples which contribute more useful sample information, without strong negative samples overwhelming the loss and backward gradients.

## III. OUR WORK

SBRCE is designed to dispose the class imbalance problem of scene text detection. We introduce SBRCE starting from the standard Cross Entropy (CE) loss for binary classification corresponding to foreground/background classification.

$$CE(\hat{Y}, Y^*) = -Y^* \log \hat{Y} - (1 - Y^*) \log(1 - \hat{Y}) \quad (1)$$

where $Y^* \in \{1, 0\}$ indicates the ground-truth class, and $\hat{Y} \in [0, 1]$ is the prediction probability of text detector for the class with label $Y^* = 1$.

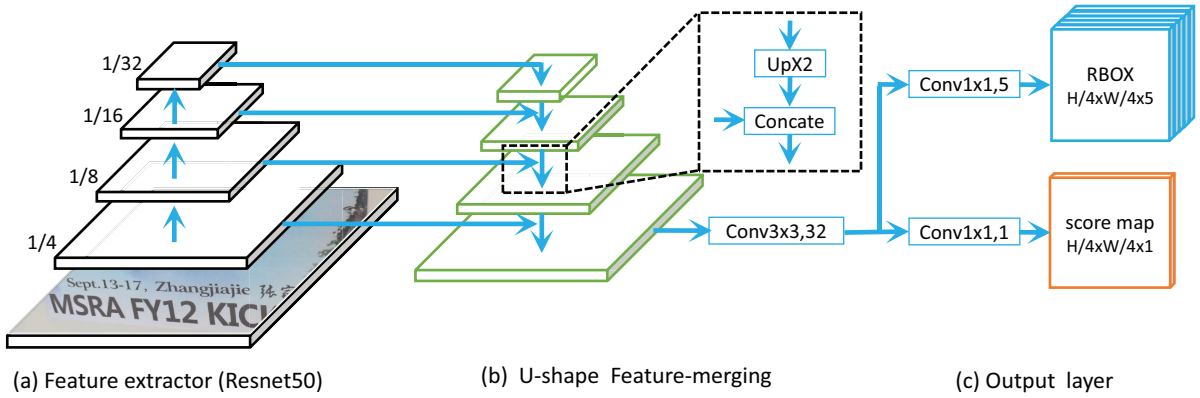


Fig. 1. Reimplemented version of EAST Framework. EAST uses Resnet50 [31] as backbone to extract feature maps. Then the idea from U-shape [32] is adopted to merge the feature maps gradually. The merged feature maps are followed by two convolution layers to output score map and RBOX.

CE loss measures the classification performance of text detectors. If one text detector only adopts CE and does not contain other sample balancing procedure such as hard negative mining, the learning is inefficient and the detection performance is not optimal.

### A. Class Balanced Cross Entropy

A common approach to deal with class imbalance is to design a weight factor $\beta \in [0, 1]$ for text with class label 1 and $(1-\beta)$ for background with class label 0. $\beta$ may be computed by inverse class frequency or regarded as a hyperparameter to set manually. This CBCE loss function can facilitate a pretty simple training procedure:

$$CBCE(\hat{Y}, Y^*) = -\beta Y^* \log \hat{Y} - (1-\beta)(1-Y^*) \log(1-\hat{Y}) \tag{2}$$

In the above, parameter $\beta$ is the balanced factor between positive and negative samples, given by

$$\beta = 1 - \frac{\sum_{y^* \in Y^*} y^*}{|Y^*|} \tag{3}$$

CBCE is a straightforward method to balance the samples by taking advantage of the number of positive and negative samples to produce a balanced factor $\beta$. When adopting CBCE as the objective function for text/non-text classification prediction, scene text detector can work better than detector using CE in practice. But CBCE does not consider the influence of strong background that prevents the text detectors from learning the weak background and all positive samples better, and reduces the detector's ability to distinguish between foreground and background.

### B. Strong-Background Restrained Cross Entropy

Like object detection, the extreme class imbalance also exists during training of scene text detectors. This imbalance restrains the learning of weak negative samples. Easily classified strong negatives comprise the majority of the loss and dominate the gradient. Notwithstanding CBCE balances the importance of positive/negative examples, it does not differentiate between strong background and weak background. Therefore, we propose a novel SBRCE loss by modifying CBCE such that it down-weights the loss assigned to the strong background and thus the training can be focused on all positive samples and weak background.

More formally, we discard the parameter $\beta$ for positive samples. Because $\beta$ is pretty close to 1 and it almost has no influence on the training of scene text detector. And then we make a new balancing factor $G$ to replace original $(1-\beta)$ for negative samples to differentially treat strong background and weak background. We define the SBRCE loss as:

$$SBRCE(\hat{Y}, Y^*) = -Y^* \log \hat{Y} - G(1-Y^*) \log(1-\hat{Y}) \tag{4}$$

$$G = \begin{cases} (1-\beta), & pred\_neg <= c\_strongbg \\ 1, & otherwise \end{cases} \tag{5}$$

where $pred\_neg = \hat{Y}(1-Y^*)$ is the prediction for negative samples. $\beta$ is the same as that of CBCE.

If the prediction of negative samples $pred\_neg$ is less than or equal to $c\_strongbg$, these negative samples are strong background and other negatives are weak background. $c\_strongbg$ is a hyperparameter to set manually, whose value is in the range of [0, 1]. Since the $c\_strongbg$ parameter has a certain range, it is easier to find its optimal value(we found $c\_strongbg = 0.5$ to work best for ICDAR2015 and $c\_strongbg = 0.6$ for MSRA-TD500). The $(1-\beta)$ is approximately equal to 0 and it can down-weight the loss and gradient of strong background effectively. As a result, SBRCE can make training focused on weak negative samples and all positive samples. It lets the text detector capture text region more easily and promotes the detector's discriminant ability

between text and background. Finally, We directly use SBRCE to replace CBCE in EAST without sacrificing the training and testing speed of EAST.

### C. Anchor-Free Scene Text Detector

In this paper, EAST is adopted as our base scene text detector. It is a light-weight and anchor-free text detector. And it is a very simple yet powerful pipeline. EAST can achieve superior performance and maintain an approximately real-time speed. The algorithm of EAST follows the general design of DenseBox [33], which is a simple and efficient object detection method. Here we briefly introduce EAST algorithm.

An overview of the reimplemented version of EAST framework is illustrated in Fig. 1. The backbone of the reimplemented version is Resnet50 [31]. Given an image $I$, a fully convolutional network (FCN) is used to extract multi-layers feature maps with different heights and widths. Since the scales of text regions vary enormously, capturing small words would require feature maps from the low-level layers of the FCN, while obtaining an accurate quadrangle enclosing a large word would need the high-level feature maps of the FCN. Hence, the FCN has to employ feature maps from multi-layers to fulfill these requirements. The idea of U-shape [32] is adopted by EAST to fuse different levels feature maps stage by stage. EAST merges low-level feature maps and high-level semantic features by making use of some concatenation and bilinear upsampling operations, as shown in Fig. 1. The fused feature maps are fed into several $conv_{1X1}$ operations to be transformed into a score map [15] and a multi-channel geometry map [15]. The geometry output can be either one of RBOX [15] or QUAD [15] and here RBOX is chosen.

For RBOX, the network produces output maps of six channels representing the score map and geometry map. The first one of the generated channels is the score map with each pixel valued from [0,1]. It indicates the probability that each location is text. Note that, before making the score map label, the ground truth quadrangles need to be shrunk to ignore the gray zone which is defined on the margin of text and background region. This label generation of the score map is like that of semantic segmentation. CBCE loss is adopted for score map. We define this loss as $Ls$:

$$Ls = CBCE(\hat{Y}, Y^*) \tag{6}$$

The rest five output channels of the generated channels describe the geometric information of the detected bouding boxes. Four channels indicate the distances from each pixel having positive score to the 4 boundaries of the rotated rectangle that encapsulates a word or text line with minimal area. These distances are used as ground truth. And IOU loss [34] is adopted for calculating loss, since it possesses invariance against multi-scale scene text. The loss is denoted as $L_{IOU}$. Next, the inclination angle of text $\hat{\theta}$ is also utilized to compute another loss:

$$L_\theta(\hat{\theta}, \theta^*) = 1 - \cos(\hat{\theta}, \theta^*) \tag{7}$$

where $\theta^*$ stands for the ground truth.

TABLE I
RESULTS ON ICDAR 2015 CHALLENGE 4 INCIDENTAL SCENE TEXT LOCALIZATION TASK

| Algorithm | Recall | Precision | F-score |
|---|---|---|---|
| StradVision1 [22] | 0.4627 | 0.5339 | 0.4957 |
| StradVision2 [22] | 0.3674 | 0.7746 | 0.4984 |
| Zhang et al. [35] | 0.4309 | 0.7081 | 0.5358 |
| Tian et al. [29] | 0.5156 | 0.7422 | 0.6085 |
| Yao et al. [20] | 0.5869 | 0.7226 | 0.6477 |
| SegLink [10] | 0.768 | 0.731 | 0.75 |
| RRPN [36] | 0.732 | 0.822 | 0.774 |
| EAST [15] | 0.7347 | 0.8357 | 0.7820 |
| DDR [14] | 0.800 | 0.820 | 0.810 |
| **EAST+CBCE** | 0.7665 | 0.8032 | 0.7844 |
| **EAST+OHEM** | 0.7256 | **0.8834** | 0.7967 |
| **EAST+Focal Loss** | 0.7675 | 0.8425 | 0.8032 |
| **EAST+SBRCE** | **0.7756** | 0.8638 | **0.8174** |

Finally, the total geometry loss $Lg$ is the weighted sum of IOU loss $L_{IOU}$ and angle loss $L_\theta$ , given by

$$Lg = L_{IOU} + \lambda_\theta L_\theta \tag{8}$$

where $\lambda_\theta$ is set to 20 in our experiments.

The total training loss $L_{det}$ is the weighted sum of $Ls$ and $Lg$, given by

$$L_{det} = Ls + \lambda_g Lg \tag{9}$$

where $\lambda_g$ is set to 1 in our experiments. The original paper of EAST shows more details.

### D. Implementation Details

All experiments are implemented in Tensorflow [37]. For the sake of fairness, we use original CBCE to reimplement the origin pipeline of EAST except for using the Resnet50 [31] to replace the original PVANET2x [15] as the CNN backbone to extract the feature of input images. Our reimplemented version of EAST can be regarded as our baseline. The proposed improved version of EAST is to replace CBCE in the baseline with SBRCE. Data augmentation of our all experiments is consistent with EAST with randomly sampling 512x512 crops from images. These experiments were conducted on a server (CPU: Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz; GPU: Titan Xp; RAM: 32G). We train our model with the batch size of 14 on 1 GPU and evaluate our model on 1 GPU with the batch size set as 1.

## IV. EXPERIMENTS

To compare the proposed SBRCE loss with previous CBCE loss, we conducted qualitative and quantitative experiments on two public benchmarks: ICDAR2015 and MSRA-TD500.

### A. Benchmark Datasets

ICDAR 2015 is Challenge 4 of ICDAR 2015 Robust Reading Competition. This dataset comprises 1000 training images and 500 test images, which are obtained by using Google Glass in a casual way. Hence, text in these images can be multi-oriented, and undergo motion blur and low resolution in some degree. Each word of every image is annotated by 4 vertices


(a) (b) (c)

(d) (e) (f)

Fig. 2. The detection results of **EAST+SBRCE**. Top row: ICDAR2015 dataset. Bottom row: MSRA-TD500 dataset.

of the quadrilateral. Then RBOX output can be produced by using the methods like $minAreaRect$ in OpenCV to fit one oriented rectangle which has the minimum area. The 229 training images from ICDAR2013 [38] are also used as the training data.

MSRA-TD500 is a dataset including a total of 500 images, 300 of which are the training data and the remaining are for evaluating. Text in this dataset is in arbitrary orientations and annotated at text line level. Unlike ICDAR2015 dataset, it consists of scene text in both Chinese and English. The text regions is annotated in RBOX format. Following the original EAST, 400 images from HUST-TR400 dataset [39] are also included as the training data, since the training set is pretty small.

### B. Quantitative Results

As shown in Table I and Table II, **EAST+CBCE** represents our reimplemented version of EAST, which uses original CBCE as loss function for score map [15]. **EAST+OHEM** takes advantage of standard CE to replace CBCE of **EAST+CBCE** and only applies OHEM to the negative samples of score map. For each image, $N$ hard negative samples and all positive samples are selected for classfication. **EAST+Focal Loss** adopts the Focal Loss [21] to replace CBCE of **EAST+CBCE**.

TABLE II
RESULTS ON MSRA-TD500

| Algorithm | Recall | Precision | F-score |
|---|---|---|---|
| TD-ICDAR [23] | 0.52 | 0.53 | 0.50 |
| TD-Mixture [23] | 0.63 | 0.63 | 0.60 |
| Yin et al. [40] | 0.63 | 0.81 | 0.71 |
| Zhang et al. [35] | 0.67 | 0.83 | 0.74 |
| DDR [14] | 0.700 | 0.770 | 0.74 |
| Yao et al. [20] | **0.7531** | 0.7651 | 0.7591 |
| EAST [15] | 0.6743 | 0.8728 | 0.7608 |
| SegLink [10] | 0.700 | 0.860 | 0.770 |
| **EAST+CBCE** | 0.7062 | 0.8527 | 0.7726 |
| **EAST+OHEM** | 0.6873 | **0.8734** | 0.7692 |
| **EAST+Focal Loss** | 0.6976 | 0.8602 | 0.7704 |
| **EAST+SBRCE** | 0.7165 | 0.8651 | **0.7838** |

**EAST+SBRCE** directly uses SBRCE as a substitute of CBCE of **EAST+CBCE**. **EAST+SBRCE** increases the performance of EAST by F-score of 3.3% on ICDAR2015 dataset (0.8174 vs. 0.7844) and F-score of 1.12% on MSRA-TD500 dataset (0.7838 vs. 0.7726).

We also compare proposed SBRCE with OHEM and Focal Loss [21]. **EAST+SBRCE** outperforms **EAST+Focal Loss** by F-score of 1.42% (0.8174 vs. 0.8032) and **EAST+OHEM** by F-score of 2.07% (0.8174 vs. 0.7967) on the ICDAR2015 dataset. **EAST+SBRCE** exceeds **EAST+Focal Loss** by F-score of 1.34% (0.7838 vs. 0.7704) and **EAST+OHEM** by F-score of 1.46% (0.7838 vs. 0.7692) on the MSRA-TD500 dataset. These experiments prove that the proposed SBRCE can significantly improve the performance of EAST. Note that Our all experimental results are based on single-scale testing.

### C. Qualitative Results

*1) Detection with SBRCE:* Fig. 2 gives some detection results of **EAST+SBRCE**. As shown in Fig. 2, our SBRCE can help EAST effectively capture text instance for these cases. Furthermore, many patterns similar to text strokes are hard to classify, such as fences, lattices, etc. SBRCE can distinguish these patterns well. The detailed analysis of SBRCE is presented in IV-C2.

*2) Comparison with Baseline:* Fig. 3 shows some detection results from **EAST+SBRCE** and **EAST+CBCE**. We can see that the proposed SBRCE can predict text regions and restrain redundant background better, because proposed SBRCE makes the network training focused on the weak negative samples and all positive samples. To analyze in detail, we summarize four

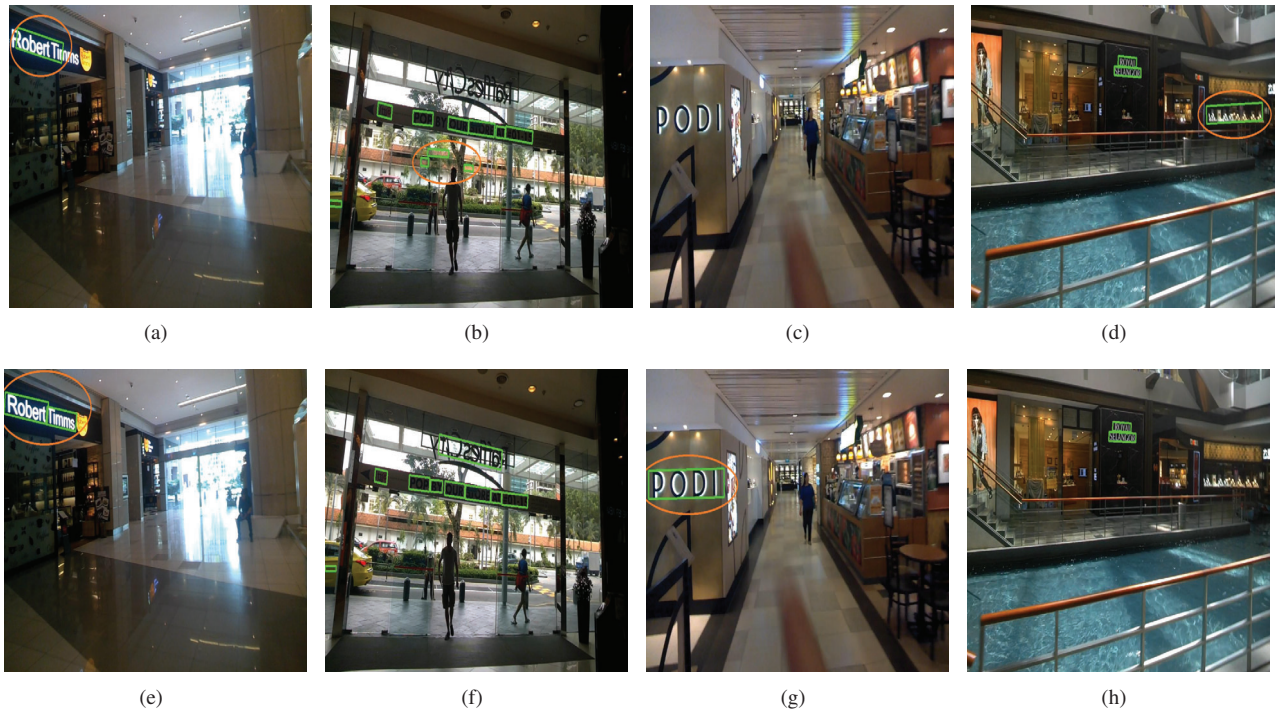


Fig. 3. Comparison of the detection results from **EAST+SBRCE** and **EAST+CBCE**: (a)-(d) from **EAST+CBCE** and (e)-(h) from **EAST+SBRCE**.

advantages of SBRCE loss compared with CBCE loss: (1) **Better text localization**: Fig. 3(e) can localize text intances better than Fig. 3(a). So it can be infered that better training of classification branch helps to train better localization branch, because they share merged features in EAST framework. (2) **Remove text-like patterns**: Fig. 3(b) contains some text-like patterns which are actually background regions, while Fig. 3(f) can avoid the mistake after training with proposed SBRCE which makes training focused on weak background. (3) **Don't miss text**: Fig. 3(g) would detect more text instances than Fig. 3(c), due to SBRCE making the training concentrated on all positive samples as well. (4) **Restrain background**: as Fig. 3(h) and Fig. 3(d) show, SBRCE can clear up more background regions than CBCE.

## V. CONCLUSION

In this work, we study the issue of class imbalance in scene text detection and analyze the drawback of CBCE loss. Due to too small weights assigned to hard negative samples, text detectors adopting CBCE for classification can not effectively learn background information. CBCE leads to sub-optimal performance of scene text detectors. To handle this, We have proposed a novel loss function, named SBRCE, which down-weights the loss of easy negative samples to focus training on hard negatives and all positive samples. Our method is simple yet highly effective. SBRCE only has a hyperparameter $c\_strongbg$ whose value is in the range of [0, 1]. Therefore, the adjustment of $c\_strongbg$ is very simple and the optimal value of $c\_strongbg$ can be found out in no time. We replace the CBCE of EAST with the SBRCE to demonstrate its efficacy. The experiments on the standard benchmarks sincerely validate the effectiveness of our SBRCE loss, which also outperforms the classical OHEM methods and influential Focal loss. Without loss of generality, SBRCE can be applied to other computer vision fields such as image classification, object detection and semantic segmentation.

## ACKNOWLEDGEMENTS

This work is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB32070000).

## REFERENCES

[1] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.
[5] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," *arXiv preprint arXiv:1611.07709*, 2016.
[6] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3150–3158.
[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.
[8] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2cnn: rotational region cnn for orientation robust scene text detection," *arXiv preprint arXiv:1706.09579*, 2017.

[9] M. Liao, B. Shi, and X. Bai, "Textboxes++: A single-shot oriented scene text detector," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3676–3690, 2018.

[10] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," *arXiv preprint arXiv:1703.06520*, 2017.

[11] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7553–7563.

[12] M. Liao, Z. Zhu, B. Shi, G.-s. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5909–5918.

[13] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in *Proc. CVPR*, 2017, pp. 3454–3461.

[14] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," *arXiv preprint arXiv:1703.08289*, 2017.

[15] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: an efficient and accurate scene text detector," in *Proc. CVPR*, 2017, pp. 2642–2651.

[16] Y. Dai, Z. Huang, Y. Gao, Y. Xu, K. Chen, J. Guo, and W. Qiu, "Fused text segmentation networks for multi-oriented scene text detection," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3604–3609.

[17] D. Deng, H. Liu, X. Li, and D. Cai, "Pixellink: Detecting scene text via instance segmentation," *arXiv preprint arXiv:1801.01315*, 2018.

[18] Q. Yang, M. Cheng, W. Zhou, Y. Chen, M. Qiu, and W. Lin, "Inceptext: A new inception-text module with deformable psroi pooling for multi-oriented scene text detection," *arXiv preprint arXiv:1805.01167*, 2018.

[19] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 761–769.

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

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *arXiv preprint arXiv:1708.02002*, 2017.

[22] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu *et al.*, "Icdar 2015 competition on robust reading," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015, pp. 1156–1160.

[23] Z. Tu, Y. Ma, W. Liu, X. Bai, and C. Yao, "Detecting texts of arbitrary orientations in natural images," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1083–1090.

[24] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2963–2970.

[25] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Asian Conference on Computer Vision*. Springer, 2010, pp. 770–783.

[26] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced mser trees," in *European Conference on Computer Vision*. Springer, 2014, pp. 497–511.

[27] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, 2016.

[28] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2315–2324.

[29] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *European conference on computer vision*. Springer, 2016, pp. 56–72.

[30] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *The IEEE International Conference on Computer Vision (ICCV)*, vol. 6, no. 7, 2017.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[33] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "Densebox: Unifying landmark localization with end to end object detection," *arXiv preprint arXiv:1509.04874*, 2015.

[34] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 516–520.

[35] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4159–4167.

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

[37] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[38] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. De Las Heras, "Icdar 2013 robust reading competition," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 1484–1493.

[39] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4737–4749, 2014.

[40] X.-C. Yin, W.-Y. Pei, J. Zhang, and H.-W. Hao, "Multi-orientation scene text detection with adaptive clustering," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 9, pp. 1930–1937, 2015.