DSRN: A Deep Scale Relationship Network for Scene Text Detection

Yuxin Wang, **Hongtao Xie***, **Zilong Fu** and **Yongdong Zhang**
School of Information Science and Technology, University of Science and Technology of China
{wangyx58, JeromeF}@mail.ustc.edu.cn, {htxie, zhyd73}@ustc.edu.cn

Abstract

Nowadays, scene text detection has become increasingly important and popular. However, the large variance of text scale remains the main challenge and limits the detection performance in most previous methods. To address this problem, we propose an end-to-end architecture called Deep Scale Relationship Network (DSRN) to map multi-scale convolution features onto a scale invariant space to obtain uniform activation of multi-size text instances. Firstly, we develop a Scale-transfer module to transfer the multi-scale feature maps to a unified dimension. Due to the heterogeneity of features, simply concatenating feature maps with multi-scale information would limit the detection performance. Thus we propose a Scale Relationship module to aggregate the multi-scale information through bi-directional convolution operations. Finally, to further reduce the miss-detected instances, a novel Recall Loss is proposed to force the network to concern more about miss-detected text instances by up-weighting poor-classified examples. Compared with previous approaches, DSRN efficiently handles the large-variance scale problem without complex hand-crafted hyperparameter settings (e.g. scale of default boxes) and complicated post processing. On standard datasets including ICDAR2015 and MSRA-TD500, the proposed algorithm achieves the state-of-art performance with impressive speed (8.8 FPS on ICDAR2015 and 13.3 FPS on MSRA-TD500).

1 Introduction

Recently, extracting and recognizing textual information in the wild have become increasingly important and popular because of its significant value in practical applications [Long *et al.*, 2018; Lyu *et al.*, 2018; Fang *et al.*, 2018; Xie *et al.*, 2019]. Scene text detection, playing a critical role in the whole process, is one of the main bottlenecks in recognition quality.

* Hongtao Xie (htxie@ustc.edu.cn)

(a) Cropped input (b) EAST (c) DSRN

Figure 1: We visualize the feature maps of different channels in training stage. The proposed method can obtain more uniform activation of multi-size text instances. Both models use ResNet50 as their basic network.

Despite the fact that recent multi-scale object detection algorithms [Liu *et al.*, 2016; Zhou *et al.*, 2018] gain great improvements, directly implementing these approaches for scene text detection, which are only for horizontal targets, may not be a good choice for multi-oriented text detection.

In the past few years, detecting objects with large-variance scale has been the main challenge in both general object detection and scene text detection. To address this problem, SSD [Liu *et al.*, 2016] combines predictions from multi-scale features to handle the objects with various scales. SAN [Kim *et al.*, 2018] constructs scale normalized patches to reduce the scale variation of objects in feature maps. Besides the scale variance, multi-oriented detection is another challenge in scene text detection. TextSnake [Long *et al.*, 2018] uses the sequence of overlapping disks to represent text instances of arbitrary orientation. Lyu *et al.* [2018] propose a novel approach for multi-oriented scene text detection by predicting the corners of a text instance. However, these previous methods usually simply fuse multi-scale features and require fairly complicated post processing, which would limit the performance and result in large amount of time consumption.

In this paper, we propose an end-to-end architecture called Deep Scale Relationship Network (DSRN) to handle the large-variance scale problem in scene text detection by mapping multi-scale convolution features onto a scale invariant space, which obtains uniform activation of multi-size text instances. Although different-scale features facilitate the detection of multi-size objects, simply concatenating features with multi-scale information would limit the performance of network due to the heterogeneity of features. To handle this issue, we firstly develop a Scale-transfer module to

transfer multi-scale features to a unified dimension. Then a Scale Relationship module is constructed on different layers to pass contextual scale information through bidirectional convolution operations. In the end, the proposed method essentially improves the quality of convolutional features by obtaining uniform activation of multi-size text instances (Figure 1 (c)).

The contributions of this paper are following:

- We develop a Scale Relationship module for multi-scale feature aggregation, which maps multi-scale convolution features onto a scale invariant space and obtains uniform activation of multi-size text instances.
- We propose a new loss function called Recall Loss, which up-weights the loss assigned to poor-classified examples. It efficiently reduces the miss-detected instances and boosts performance in recall.
- Our DSRN essentially improves the quality of convolutional features, and the proposed Scale Relation module can be easily embedded into other existing detection networks, boosting the performance without obvious speed sacrifice.

2 Related Works

2.1 Indirect Regression based methods.

Indirect Regression based methods regress the offsets from a default box to the corresponding ground truth. As inspired by the object detection method [Liu *et al.*, 2016], Liao *et.al.*[2018] construct a fully convolutional architecture and adopt ARF [Zhou *et al.*, 2017b] to generate rotation-sensitive features, which obtains better representation for multi-oriented texts. He et.al.[2017a] develop a hierarchical inception module to aggregate multi-scale features, then predictions from multi-scale features are combined for better multi-size detection.

2.2 Direct Regression based methods.

In order to handle the large variance of text scale and multi-orientation issues, many previous works often contain multiple sequential steps or complicated post processing, which result in large time consumption and are difficult for practical application. Different from previous methods, direct regression based methods directly predict offsets from bounding box boundaries or vertexes to points without setting complex hand-crafted hyperparameters (e.g. scale of default boxes), providing a new approach for accurate and efficient text detection [Zhou *et al.*, 2017a; He *et al.*, 2017b].

Although direct regression based methods provide an efficient approach for scene text detection, simply concatenating features from different levels would limit the network performance due to the heterogeneity of features. Thus we propose an effective approach to aggregate multi-scale information through bidirectional convolution operations, which maps multi-scale convolution features onto a scale invariant space. Contributing to the aggregation of multi-scale information, our DSRN obtains more uniform activation of multi-size text instances (Figure 1).

2.3 Segmentation based methods.

Inspired by the segmentation methods [Long *et al.*, 2015; Min and Chen, 2018], some algorithms are proposed for scene text detection by using segmentation maps. TextSnake [Long *et al.*, 2018] uses the sequence of overlapping disks to represent text instances of arbitrary orientation. Pixel link [Deng *et al.*, 2018] predicts the links among every pixel and their neighbors which are valid when both of the linked pixels belong to text instances. By doing this, pixel link successfully separates text instances that are very close to each other. Although the segmentation based methods are able to handle the text with various scales, they usually require fairly complicated post processing which will slow down the speed.

3 The Proposed Method

3.1 Overview

In this section, we will introduce the details of DSRN. Our method is based on [Zeng *et al.*, 2016], which is originally proposed to use the contextual visual cues of ROIs in general object detection. We design a novel Scale Relationship module and implement a Scale-transfer module to extend this framework for scene text detection. Different from [Zeng *et al.*, 2016], our method is proposed to aggregate multi-scale feature maps and to map the multi-scale features onto a scale invariant space to obtain uniform activation of multi-size text instances. The total architecture of DSRN is illustrated in Figure 2.

We use ResNet50 [He *et al.*, 2016] as our basic network and reduce the channels of feature maps during up-sample to obtain a light framework. In order to aggregate multi-scale information from feature maps of different scales, Scale-transfer module is proposed to transfer the multi-scale convolution features to a unified dimension. Then Scale Relationship module passes contextual multi-scale information between unified features during both feature learning and extraction. Such message passing is conducted in different layers and carried out through bidirectional convolution operations. In the end, the Non-maximum suppression (NMS) is adopted to reduce redundant results and the whole network is optimized in an end-to-end way. The effective aggregation of multi-scale information and lightness of this architecture make our approach accurate and efficient.

3.2 Feature Extraction

The backbone of DSRN is adapted from a pre-trained ResNet50 [He *et al.*, 2016] network and designed with following considerations: 1) the backbone must have enough capacity to handle the large variance of text scale. 2) Features should contain more contextual information, due to the complexity of background within natural scene images. 3) The corresponding feature maps should include strong semantic information to represent multi-size texts. Inspired by FPN [Lin *et al.*, 2017a] which achieves the good performance on those problems, we adopt the backbone with a similar architecture to extract features.

Particularly, we convert the fc layers in ResNet50 to an attention module [Wang *et al.*, 2017] (F1 in Figure 2). Then

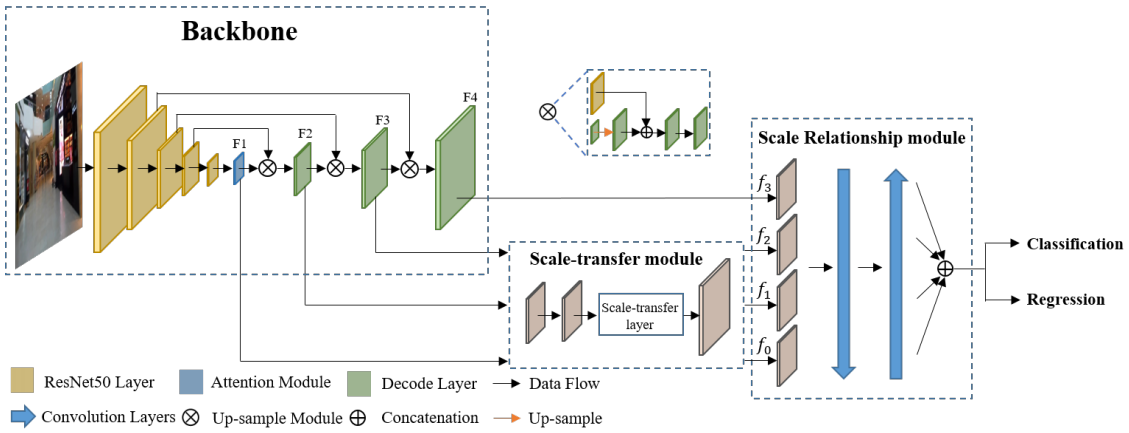


Figure 2: The architecture of proposed network. The network contains three parts: backbone, Scale-transfer module and Scale Relationship module. The backbone is adapted from ResNet50. Scale-transfer module and Scale Relationship module are built on multiple feature layers. Output features from Scale relationship module are used for classification and regression.

several extra convolutional layers (F2, F3 and F4) are stacked above the attention module in a top-down pathway.

3.3 Scale-transfer Module

Features from different stages have various scales, which makes it difficult for aggregation. Inspired by the object detection method [Zhou *et al.*, 2018], we develop a Scale-transfer module to generate feature maps with a unified dimension (1/4 of input image in our experiment).

As illustrated in Figure 3, channel matching layer firstly produces features with corresponding channels, which controls the output channels of scale-transfer layer. Then scale-transfer layer expands width and hight of feature maps simultaneously by compressing the number of channels. The dimensions of input tensor are $C_i \times H_i \times W_i$ $(i = 1, 2, 3)$ and those of outputs are $\frac{C_i}{m^2} \times m \times H_i \times m \times W_i$. We choose m to 8, 4, 2 for F1, F2, F3 respectively. Finally, the multi-scale features are mapped to a unified dimension through Scale-transfer module.

As channel matching layer should also be implemented following other expanding approaches for channel normalization (channels of features are usually related to stages), Scale-transfer module expands the scale of feature maps with less additional parameters, which can make our network more efficient.

3.4 Scale Relationship Module

The detection of large text requires features from late-stage with small scales, while accurate geometry prediction of small texts needs low-level information from early-stage [Zhou *et al.*, 2017a]. However, simply concatenating features with multi-scale information would limit the performance of network due to the heterogeneity of features. Different from previous methods [He *et al.*, 2017a; Lyu *et al.*, 2018; Liao *et al.*, 2018] which combine predictions from multi-scale features to handle the large-variance scale problem, we propose a Scale Relationship module to aggregate multi-scale features. It passes contextual scale information among multi-scale feature maps and maps the multi-scale features onto a

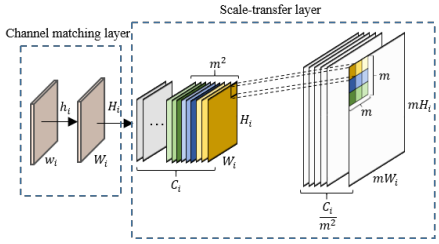


Figure 3: The architecture of Scale-transfer module. Channel matching layer produces features with corresponding channels to control output dimensions of scale-transfer layer.

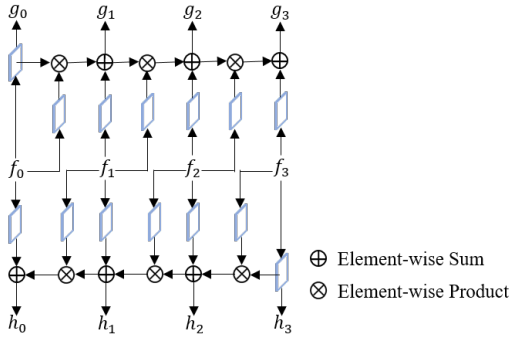


Figure 4: The architecture of bidirectional convolution. Inputs are feature maps from decoding layers in Figure 2. Contextual scale information passes through bidirectional convolution layers. Element-wise product is learned to control the multi-scale information passing.

scale invariant space, which results in uniform activation of multi-size text instances and essentially improves the quality of features (Figure 1(c)).

As illustrated in Figure 4, Scale Relationship module passes contextual scale information by sequentially convoluting feature maps through bidirectional convolution operations. f_0, f_1, f_2 and f_3 are corresponding features from decoding layers F_1, F_2, F_3 and F_4 in Figure 2 respectively. Convolution in the first direction starts from the last decoding layer (f_3) and ends at the first decoding layer (f_0), thus the

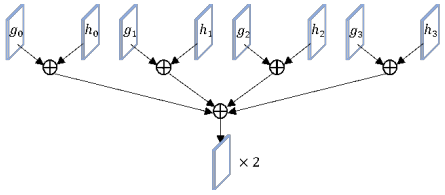


Figure 5: Feature aggregation. Two following convolution layers are implemented for better representation

sequential feature maps can receive scale information from larger-scale features, which is better for small text detection. In contrast, sequential features can also receive large-size textual information in the second direction. Element-wise product operation is learned to control how much contextual scale information can be received from previous feature maps. The outputs of bidirectional convolution (g_i and h_i, $i = 0, 1, 2, 3$ in Figure 4) have richer multi-scale and stronger semantic information than previous feature maps (f_i, $i = 0, 1, 2, 3$).

In the end, we concatenate the output features from bidirectional convolution operations and implement two sequential convolution layers to obtain better representation. In order to balance the accuracy and efficiency in the detection task, only four decoding layers are used in the proposed method.

4 Loss Functions

The multi-task loss is formulated as:

$$L = L_{cls} + \lambda_{reg} L_{reg} \tag{1}$$

Where L_{cls} and L_{reg} are classification loss and regression loss respectively. λ_{reg} is a hyperparameter to balance L_{cls} and L_{reg}. In our experiment, we set λ_{reg} to 1.

4.1 Loss for Clssification

In most previous detection approaches, processes such as hard negative mining and balanced sampling are implemented to handle the imbalanced distribution between background and target objects. Lin *et al.* [2017b] propose a Focal Loss (FL), which achieves better training by down-weighting the well-classified examples, as seen in (2).

$$FL(p_t) = -a_t(1 - p_t)^{\lambda} \log(p_t) \tag{2}$$

As direct regression based methods directly predict a score map of text regions, Focal Loss will distribute almost the same weight to miss-detected text instances (red region in Figure 6) and the border of text instances (green region in Figure 6), which makes the model pay identical attention to both regions during training and is harmful to the performance. To handle this issue, we propose a new loss function to force the network to pay more attention to the miss-detected text instances without concerning about the border of texts, which is called Recall Loss (RL).

$$RL = \begin{cases} -\alpha\eta_1 \log(p) & if\ y = 1 \\ -\eta_2 \log(1 - p) & otherwise \end{cases} \tag{3}$$

$$\alpha = \begin{cases} (e - \theta)^{\beta - IoU} & if\ \ IoU < \beta \\ 1 & otherwise \end{cases} \tag{4}$$

$$IoU = S \cap G / S \cup G \tag{5}$$

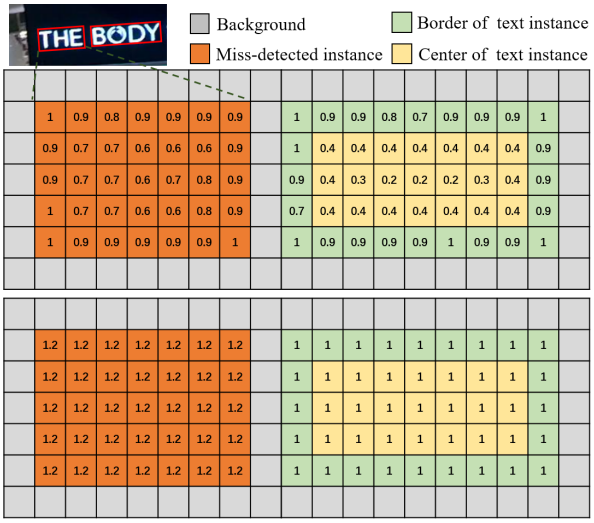


Figure 6: Pixels with confidence below 0.5 in score map are regarded as negative samples. Left regions are miss-detected text instances ($IoU = 0$ in (5)), and right regions are detected text instances. Both miss-detected text instances and border of detected text instances obtain the similar confidence the in score map. Top: weight in Focal Loss. We set λ and a_t to 1 for visualization. Down: weight in Recall Loss. We set β, θ in Recall Loss to 0.4 and 1 respectively for visualization.

As seen in (3)(4)(5). Where S is the predicted region with confidence of pixels above 0.5, G is the corresponding ground truth, e is a constant, and α and β are hyperparameters which will be illustrated in ablation studies. If $IoU < \beta$, then $\alpha > 1$ exists, which means all the pixels in this region will be assigned a larger weight, and the border and the center of text instances with $IoU > \beta$ will obtain the same weight that is lower than miss-detected text instances. This property forces the network to focus on regions with small IoU, which can effectively reduce the miss-detected instances (Figure 7 (a)). η_1 and η_2 are balancing factors between positive and negative samples. In our experiment, we combine Recall Loss with Dice Loss (6) and balance them by λ_R and λ_D respectively (o_i and y_i are pixels in score map and ground truth respectively). In the end, the classification loss is formulated as (7).

$$L_{Dice} = 1 - \frac{2\sum_i o_i y_i}{\sum_i o_i + \sum_i y_i} \tag{6}$$

$$L_{cls} = \lambda_R RL + \lambda_D L_{Dice} \tag{7}$$

4.2 Loss for Regression

We choose the same regression approach in [Zhou *et al.*, 2017a], which is invariant against scales of objects (8)(9).

$$L_{loc} = -\log IoU(P, G) = -\log \frac{|P \cap G|}{|P \cup G|} \tag{8}$$

$$L_{\theta} = 1 - \cos(\theta' - \theta^*) \tag{9}$$

We combine L_{loc} with L_{θ} for final regression loss. Distance from a pixel to 4 boundaries and value of angle are outputs for both training and inference.

5 Experiment

In this section, we evaluate the performance of Scale Relationship module and of Recall Loss respectively. We compare our DSRN with recent state-of-art methods on benchmark datasets to prove our accuracy and efficiency.

5.1 Benchmark Datasets

ICDAR2015. This is a dataset for incidental scene text detection proposed in the Challenge 4 of ICDAR 2015 Robust Reading Competition [Karatzas *et al.*, 2015]. It includes 1000 training images and 500 test images with annotations labeled as 4 vertices of a word level quadrangle. We fit a rotated rectangle with the minimum area for training. Different from previous dataset with only horizontal annotations, this benchmark is proposed for evaluating multi-orientation text detection and contains text of different scales, ambiguities, resolutions, perspectives, and directions.

HUST. [Yao *et al.*, 2014] is a dataset contains 400 images, which consists of Arabic numbers and English letters of different fonts with text line level labels.

MSRA-TD500. This is a dataset proposed in [Yao *et al.*, 2012] for detecting arbitrary-oriented and multi-lingual long text lines. It contains 300 images for training and 200 images for testing. Since the size of training data is too small to learn a deep newtwork, we also use 400 images from HUST [Yao *et al.*, 2014] in training stage.

5.2 Experimental Setup

Our proposed network is trained end-to-end on NVIDIA TITAN X GPU using ADAM[Kingma and Ba, 2014] optimizer. We perform data augmentation by randomly cropping each image and resize it to 512×512 for training. We update the learning rate by a multi-step strategy. The initial learning rate is 1e-3, and decays by 0.94 every 10k steps. We set batchsize to be 14 and training continues until convergence. In test stage, NMS is implemented to reduce the redundant results.

5.3 Ablation Studies

Recall Loss. θ and β are important hyperparameters which allow us to vary the capacity of Recall Loss in our model. To investigate this relationship, several experiments are conducted using different θ and β values on ICDAR2015 in Table 1. We also conduct an experiment without Recall Loss (α = 1 in (3)) for optimizing. The comparison reveals that Recall Loss can boost the performance in recall and obtain higher F-measure. Furthermore, we compare Recall Loss with Focal Loss [Lin *et al.*, 2017b] and Lovasz-Softmax Loss [Berman *et al.*, 2018] in Table 2. The comparison reveals that Recall Loss achieves better performance in both recall and F-measure.

θ	β	R	P	F
0	0	0.772	**0.846**	0.807
1	0.3	0.787	0.788	0.788
1.5	0.3	**0.796**	0.82	0.808
1.7	0.3	**0.796**	0.832	**0.814**
1.7	0.2	0.766	0.81	0.787
1.7	0.4	0.777	0.83	0.803

Table 1: Performance on ICDAR2015 with different settings of θ and β.

Loss	R	P	F
Focal Loss	0.592	**0.923**	0.721
Lovasz-Softmax	0.767	0.834	0.799
Recall Loss	**0.796**	0.832	**0.814**

Table 2: Compared with other loss functions on ICDAR2015.

Algorithm	R	P	F	FPS
baseline1†	0.541	0.751	0.62	**16.1**
baseline2†	0.67	0.847	0.748	15.4
DSRN†	**0.712**	**0.876**	**0.785**	13.3
baseline1*	0.612	0.70	0.657	**12.1**
baseline2*	0.777	0.826	0.80	10.8
DSRN*	**0.796**	**0.832**	**0.814**	8.8

Table 3: Performance gain of Scale Relationship module. † and * means experiments on MSRA-TD500 and ICDAR2015 respectively.

Scale Relationship module. To explore the gain of our Scale Relationship module, we train baseline networks without Scale Relationship module (baseline1) and with unidirectional convolution operation (baseline2). All baseline models contain the identical backbone, prediction module and training settings as DSRN. With slight time consumption, DSRN boosts the performance greatly.

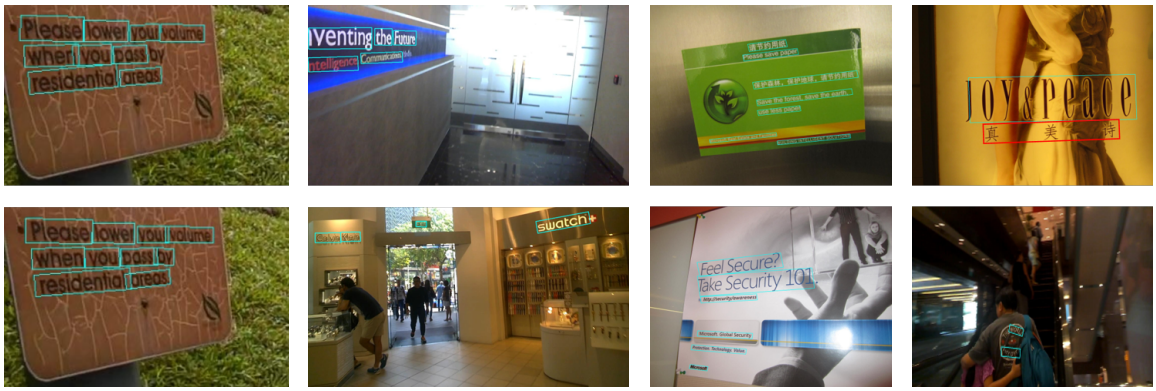
5.4 Comparison with existing methods

We evaluate our model on **ICDAR2015** incidental scene text to test its performance of arbitrarily oriented text detection. 229 images from ICDAR2013 are also used for training. We choose θ and β to 1.7 and 0.3 respectively in training stage and resize images to 768×1280 in test stage .

We compare the proposed method with state-of-art algorithms in Table 4. DSRN works better in direct regression based methods [He *et al.*, 2017b; Zhou *et al.*, 2017a], and outperforms them by a large margin (0.814 vs 0.70 and 0.782). We attribute our high performance to the aggregation of multi-scale information in Scale Relationship module, which essentially improves the quality of feature maps. When tested at single scale, the proposed method surpasses most state-of-art methods [Lyu *et al.*, 2018; Wang *et al.*, 2018; Liu *et al.*, 2018] with the fastest speed. Although [Liao *et al.*, 2018] is slightly better than ours (0.822 vs 0.814), which contains a corner grouping process resulting a lot of time

Algorithm	R	P	F	FPS
[He *et al.*, 2017b]*	0.62	0.82	0.70	-
[Zhou *et al.*, 2017a]*	0.735	0.836	0.782	-
[Jiang *et al.*, 2018]*	0.743	0.764	0.753	2.2
[Liu *et al.*, 2018]	**0.80**	0.72	0.76	-
[Wang *et al.*, 2018]	0.741	0.857	0.795	-
[Lyu *et al.*, 2018]*	0.707	**0.941**	0.807	3.6
[Liao *et al.*, 2018]*	0.79	0.856	**0.822**	6.5
DSRN*	0.796	0.832	0.814	**8.8**

Table 4: Results on ICDAR2015. *means single scale.



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

Figure 7: Some detection samples on ICDAR2015 and MSRA-TD500. (a) illustrates the performance of training without Recall Loss (top) and training with Recall Loss (down). The miss-detected word 'by' is able to be located by training with Recall Loss. Some failure cases are presented in (d), where red boxes are ground truths while green boxes are predicted results

Algorithm	R	P	F	FPS
[He *et al.*, 2017b]	0.70	0.77	0.74	-
[Zhou *et al.*, 2017a]	0.674	0.873	0.761	13.2
[Long *et al.*, 2018]	0.739	0.832	0.783	1.1
[Deng *et al.*, 2018]	0.732	0.83	0.778	3
[Liao *et al.*, 2018]	0.73	0.87	0.79	10
[Lyu *et al.*, 2018]	**0.762**	**0.876**	**0.815**	5.7
DSRN	0.712	**0.876**	0.785	**13.3**

Table 5: Results on MSRA-TD500.

consumption, our method achieves 35% faster speed (8.8 FPS vs 6.5 FPS) due to a lighter architecture with simpler post processing (NMS).

On **MSRA-TD500**, we evaluate the capacity of our algorithm for detecting long and multi-lingual text lines. We pre-train our model on ICDAR2015 and then finetune it until convergence. In test stage, we resize images to 672×672.

With comparisons to other representative results in Table 5, our method achieves the state-of-art performance in recall, precision and F-measure (0.712, 0.876 and 0.785), outperforming the direct regression based methods [He *et al.*, 2017b; Zhou *et al.*, 2017a] (0.785 vs 0.74 and 0.761) and other representative methods [Long *et al.*, 2018; Deng *et al.*, 2018]. Furthermore, our method achieves the identical state-of-art performance in recall to [Lyu *et al.*, 2018] with 142% faster speed (13.3 FPS vs 5.7 FPS).

Particularly, [Lyu *et al.*, 2018] and [Liao *et al.*, 2018] use SynthText [Gupta *et al.*, 2016] (800000 images in [Lyu *et al.*, 2018]) to pre-train their models and then finetune them on corresponding tasks (both ICDAR2015 and MSRA-TD500). The proposed DSRN can obtain better or competitive results with much less training datasets, which proves that our DSRN is a more general model for scene text detection task.

Some detection samples of DSRN are visualized in Figure 7. The proposed method can perform well in most situations, handling the large variance of text scale and arbitrary orientation in scene text detection. However, it fails to detect text lines with large character spacing (top of Figure 7 (d)). Furthermore, our method performs not well in curved text detection (bottom of Figure 7(d)), as few curved samples are in training set.

5.5 Rationality of High Performance and Fast Speed

DSRN is proposed to detect multi-size texts automatically. The huge increase in accuracy and efficiency is mainly due to three aspects. 1) We develop a Scale Relationship module to learn the contextual scale information and to obtain uniform activation of multi-size text instances, which essentially improves the quality of convolutional feature maps. 2) We propose a new loss function called Recall Loss, which effectively reduces miss-detected text instances by up-weighting the weight of poor-classified examples. 3) The direct regression approach and light network make our model efficient.

6 Conclusion

In this paper, we have presented a novel end-to-end method for scene text detection. The main idea is mapping multi-scale features onto a scale invariant space, which obtains uniform activation of multi-size text instances and effectively handles the problem of large variance of text scale in scene text detection without setting complex hand-crafted hyperparameters. Another improvement is that we propose a new loss function called Recall Loss, which reduces miss-detected text instances by up-weighting the poor-classified examples. The experiments on benchmarks reveal that our model achieves the state-of-art performance with impressive speed. Furthermore, we also analyze the reasons of our high performance and fast speed. As for future work, we will further improve the performance by combining our detection framework with a recognition branch.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2017YFC0820600), National Defense Science and Technology Fund for Distinguished Young Scholars (2017-JCJQ-ZQ-022), the National Nature Science Foundation of China (61525206, 61771468), the Youth Innovation Promotion Association Chinese Academy of Sciences (2017209).

References

[Berman *et al.*, 2018] Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, pages 4413–4421, 2018.

[Deng *et al.*, 2018] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. Pixellink: Detecting scene text via instance segmentation. In *AAAI*, pages 6773–6780, 2018.

[Fang *et al.*, 2018] Shancheng Fang, Hongtao Xie, Zheng-Jun Zha, Nannan Sun, Jianlong Tan, and Yongdong Zhang. Attention and language ensemble for scene text recognition with convolutional sequence modeling. In *2018 ACM MM*, pages 248–256. ACM, 2018.

[Gupta *et al.*, 2016] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, pages 2315–2324, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2017a] Pan He, Weilin Huang, Tong He, Qile Zhu, Yu Qiao, and Xiaolin Li. Single shot text detector with regional attention. In *ICCV*, pages 3066–3074, 2017.

[He *et al.*, 2017b] Wenhao He, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Deep direct regression for multi-oriented scene text detection. In *ICCV*, pages 745–753, 2017.

[Jiang *et al.*, 2018] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo. R 2 cnn: Rotational region cnn for arbitrarily-oriented scene text detection. In *2018 24th ICPR*, pages 3610–3615. IEEE, 2018.

[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 *2015 13th ICDAR*, pages 1156–1160. IEEE, 2015.

[Kim *et al.*, 2018] Yonghyun Kim, Bong-Nam Kang, and Daijin Kim. SAN: learning relationship between convolutional features for multi-scale object detection. In *ECCV*, pages 328–343. Springer, 2018.

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[Liao *et al.*, 2018] Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-Song Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. In *CVPR*, pages 5909–5918, 2018.

[Lin *et al.*, 2017a] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944, 2017.

[Lin *et al.*, 2017b] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2999–3007, 2017.

[Liu *et al.*, 2016] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

[Liu *et al.*, 2018] Zichuan Liu, Guosheng Lin, Sheng Yang, Jiashi Feng, Weisi Lin, and Wang Ling Goh. Learning markov clustering networks for scene text detection. In *CVPR*, pages 6936–6944, 2018.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[Long *et al.*, 2018] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *ECCV*, pages 19–35. Springer, 2018.

[Lyu *et al.*, 2018] Pengyuan Lyu, Cong Yao, Wenhao Wu, Shuicheng Yan, and Xiang Bai. Multi-oriented scene text detection via corner localization and region segmentation. In *CVPR*, pages 7553–7563, 2018.

[Min and Chen, 2018] Shaobo Min and Xuejin Chen. A robust deep attention network to noisy labels in semi-supervised biomedical segmentation. *arXiv preprint arXiv:1807.11719*, 2018.

[Wang *et al.*, 2017] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, pages 6450–6458, 2017.

[Wang *et al.*, 2018] Fangfang Wang, Liming Zhao, Xi Li, Xinchao Wang, and Dacheng Tao. Geometry-aware scene text detection with instance transformation network. In *CVPR*, pages 1381–1389, 2018.

[Xie *et al.*, 2019] Hongtao Xie, Shancheng Fang, Zheng-Jun Zha, Yating Yang, Yan Li, and Yongdong Zhang. Convolutional attention networks for scene text recognition. *ACM TOMM*, 15(1s):3, 2019.

[Yao *et al.*, 2012] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *CVPR*, pages 1083–1090, 2012.

[Yao *et al.*, 2014] Cong Yao, Xiang Bai, and Wenyu Liu. A unified framework for multioriented text detection and recognition. *IEEE Trans. Image Processing*, 23(11):4737–4749, 2014.

[Zeng *et al.*, 2016] Xingyu Zeng, Wanli Ouyang, Bin Yang, Junjie Yan, and Xiaogang Wang. Gated bi-directional CNN for object detection. In *ECCV*, pages 354–369, 2016.

[Zhou *et al.*, 2017a] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. EAST: an efficient and accurate scene text detector. In *CVPR*, pages 2642–2651, 2017.

[Zhou *et al.*, 2017b] Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Oriented response networks. In *CVPR*, pages 4961–4970, 2017.

[Zhou *et al.*, 2018] Peng Zhou, Bingbing Ni, Cong Geng, Jianguo Hu, and Yi Xu. Scale-transferrable object detection. In *CVPR*, pages 528–537, 2018.