# Industrial Scene Text Detection with Refined Feature-attentive Network

Tongkun Guan[1], Chaochen Gu[1], Changsheng Lu[2], Jingzheng Tu[1], Qi Feng[1], Kaijie Wu[1], Xinping Guan[1]

[1] Shanghai Jiao Tong University, [2] The Australian National University

IEEE Transactions on Circuits and Systems for Video Technology — Journal

## Problem Definition and Contribution

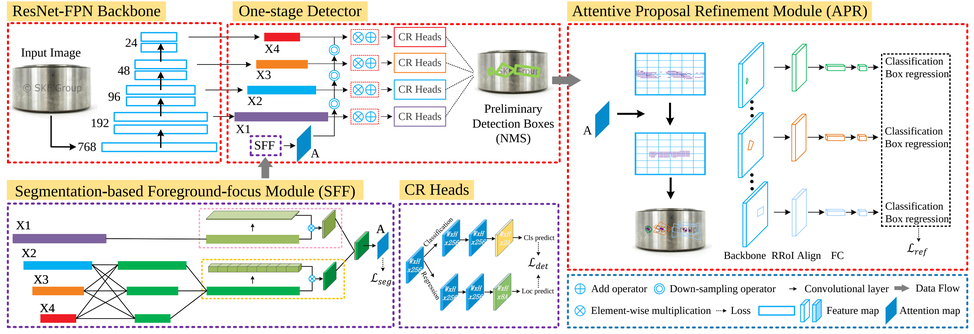
**Goal:** Detecting the multi-oriented texts on the surface of metal parts in industrial scenarios.

**Key Contributions:** A flexible deep learning framework for industrial text detection that

- focuses on more detailed text foreground features to weaken the influence of other factors (*e.g.*, low visual contrast, corroded surfaces, and complex backgrounds).
- excavates high-quality foreground boxes for localization correction.
- achieves state-of-the-art performance on the industrial dataset and robustly detects horizontal texts, multi-oriented texts, and multi-language texts of natural scenes.
- establishes the first benchmark dataset (Metal Part Surface Character Dataset, MPSC) to promote in-depth research on industrial text detection.

## Method

**Network Architecture:** RFN consists of three components, namely a ResNet-FPN backbone, a one-stage detector, and an attentive proposal refinement module.




**Post-processing:** We combine the instance score $S_I$ and classification score $S_c$ to generate a new confidence score. And then, they are fed into the NMS algorithm to get the best prediction boxes:

$$S_I = \frac{\Sigma_{j=1}^{N}\rho_j}{N}, S' = e^{S_c}(1+\mu\frac{e^{S_I}}{e^{1-S_I}}) \tag{1}$$

## Optimization

**Assumption:** $\omega_i$ and $\omega_i^*$ are the confidence score of pixel $i$ in $S_{gt}$ and the attention map $A$.

**Loss function:**

$$\omega_d = \omega_i - \omega_i^*, \mathcal{D}_a = \frac{\Sigma_{i=1}^{N}1_{[\omega_d \geqslant \frac{1}{2}]}(1-\omega_i^*)}{\Sigma_{i=1}^{N}(\omega_i^*)}, \mathcal{D}_b = \frac{\Sigma_{i=1}^{N}1_{[-\omega_d \geqslant \frac{1}{2}]}\omega_i^*}{\Sigma_{i=1}^{N}(\omega_i^*)},$$

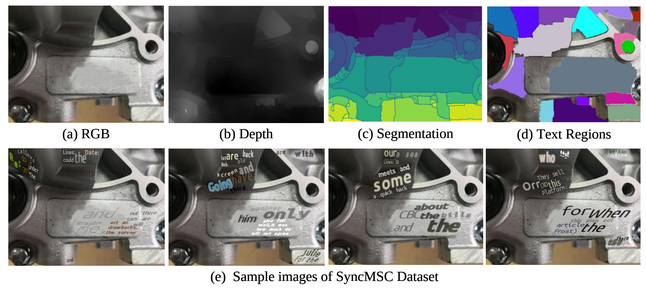
$$\mathcal{L}_g = \begin{cases} \mathcal{D}_a & \text{if } \mathcal{D}_b < \Delta, \\ \mathcal{D}_a + \mathcal{D}_b - \Delta & \text{if } \mathcal{D}_b \geqslant \Delta \end{cases}, \mathcal{L}_{seg} = \mathcal{L}_d + e^{-1*\mathcal{L}_d*\gamma} * \mathcal{L}_g, \tag{2}$$

where $\mathcal{L}_d$ denotes the dice loss, N is the number of pixels in the attention map $A$.

**Main Idea:** We allow certain false-positive classification results in exchange for detecting more textual features in low-contrast and indistinguishable regions when the $\mathcal{L}_d$ stabilizes.

## Experiments & Results

### Synthetic Datasets (SynthMPSC, 98962 images):

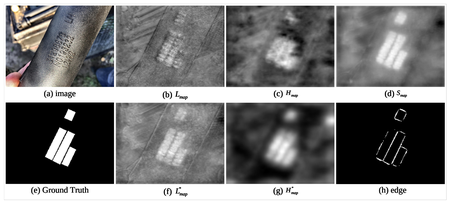


(a) RGB (b) Depth (c) Segmentation (d) Text Regions

(e) Sample images of SyncMSC Dataset

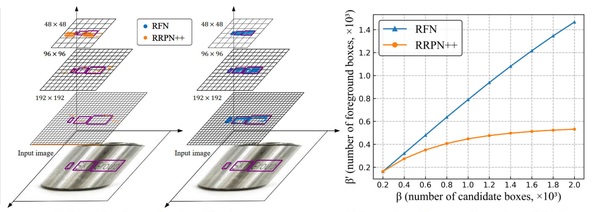
### Self-built Benchmark Dataset (MPSC, 3194 images):


### Feature Visualization (SFF):



(a) image (b) $L_{map}$ (c) $H_{map}$ (d) $S_{map}$ (e) Ground Truth (f) $L'_{map}$ (g) $H'_{map}$ (h) edge

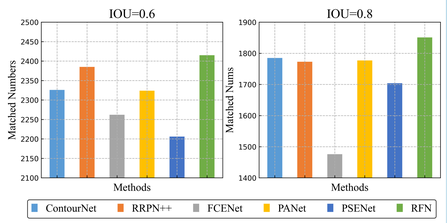
### Central location results of candidate boxes:




(a) Central location results of candidate boxes

(b) the number of foreground boxes in candidate boxes

### Location Correction Results (APR):




**Analysis of SFF and APR Models:** 1) Scale-sensitive feature fusion. It enhances the feature complementarity and generates robust feature representations with multi-scale texts. 2) Box selection algorithm by multi-scale masks. Background boxes have a low opportunity to be re-corrected, which reduces FP. More foreground boxes (several times than RPN) at each ground-truth box are obtained, which improves the quality of candidates. It promotes positive learning of the refined network and obtains more high-IOU boxes, increasing TP.

### Qualitative Results on the MPSC Dataset and Publicly Available Natural Text Datasets:

| MPSC | | | | MSRA-TD500 | | | | ICDAR2013 | | | | ICDAR2017-MLT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Algorithms | Precision (%) | Recall (%) | F-measure (%) | Algorithms | Precision (%) | Recall (%) | F-measure (%) | Algorithms | Precision (%) | Recall (%) | F-measure (%) | Algorithms | Precision (%) | Recall (%) | F-measure (%) |
| EAST | 76.33 | 73.04 | 74.65 | TextSnake† | 83.2 | 73.9 | 78.3 | SegLink* | 92 | 84.4 | 88.1 | Sensetime OCR | 56.9 | 69.4 | 62.6 |
| Mask R-CNN | 85.28 | 79.25 | 82.15 | PixelLink* | 83 | 73.2 | 77.8 | SSTD | 89 | 86 | 88 | FOTS* | 81.9 | 62.3 | 67.3 |
| RRPN | 81.98 | 78.91 | 80.42 | RRPN | 82 | 68 | 74 | TextBoxes++* | 92 | 86 | 89 | LOMO | 78.8 | 60.6 | 68.5 |
| PSENet | 85.42 | 78.4 | 81.76 | RRD† | 87 | 73 | 79 | RRD* | 92 | 86 | 89 | LOMO* | 80.2 | 67.2 | 73.1 |
| PAN | 87.07 | 81.6 | 84.24 | Lyu et al. | 87.6 | 76.2 | 81.5 | PixelLink* | 88.6 | 87.5 | 88.1 | PSENet | 73.8 | 68.2 | 70.9 |
| BDN | 86.6 | 77.49 | 81.79 | AS-RPN | 84.7 | 80.4 | 82.5 | RRPN | 84 | 77 | 80 | PSENet‡ | 75.4 | 69.2 | 72.1 |
| ContourNet | 87.79 | 81.02 | 84.27 | CRAFT | 88.2 | 78.2 | 82.9 | Melinda et al. | 93.9 | 91.5 | 92.6 | CharNet | 77.1 | 70.1 | 73.4 |
| RRPN++ | 86.73 | 83.9 | 85.3 | ATRR | 85.2 | 82.1 | 83.6 | FTPN | 93.2 | 91.9 | 92.5 | CRAFT | 80.6 | 68.2 | 73.9 |
| FCENet | 87.13 | 81.63 | 84.29 | PAN‡ | 84.4 | 83.8 | 84.1 | Liu et al. | 90.2 | 86.3 | 88.2 | Unrealtext | 82.2 | 67.4 | 74.1 |
| RFN (ours) | 89.3 | 83.33 | 86.21 | RFN | 88.4 | 80 | 84 | Wei et al. | 93.7 | 87.4 | 90.4 | ISNet | 78 | 67.4 | 72.3 |
| RFN* (ours) | 89.82 | 84.45 | 87.05 | RFN‡ | 88.4 | 87.8 | 88.1 | RFN (ours) | 92.5 | 90.7 | 91.6 | RFN (ours) | 79.4 | 67.6 | 73 |

### Qualitative Results on MPSC Dataset:

### Qualitative Results on Natural Scene Datasets:

(a) MSRA-TD500

(b) USTB-SV1K

(c) ICDAR2013

(d) ICDAR2017-MLT