Industrial Scene Text Detection With Refined Feature-Attentive Network
Tongkun Guan©, Chaochen Gu©, Changsheng Lu©, Jingzheng Tu©, Qi Feng©, Kajjie Wu©, and Xinping Guan©, Fellow, IEEE

Abstract—Detecting the marking characters of industrial metal parts remains challenging due to low visual contrast, uneven illumination, corroded surfaces, and cluttered background of metal part images. Affected by these factors, bounding boxes generated by most existing methods could not locate low-contrast text areas very well. In this paper, we propose a refined feature-attentive network (RFN) to solve the inaccurate localization problem. Specifically, we first design a parallel feature integration mechanism to construct an adaptive feature representation from multi-resolution features, which enhances the perception of multi-scale texts at each scale-specific level to generate a high-quality attention map. Then, an attentive proposal refinement module is developed by the attention map to rectify the location deviation of candidate boxes. Besides, a re-scoring mechanism is designed to select text boxes with the best rectified location. To promote the research towards industrial scene text detection, we contribute two industrial scene text datasets, including a total of 102156 images and 1948809 text instances with various character structures and metal parts. Extensive experiments on our dataset and four public datasets demonstrate that our proposed method achieves the state-of-the-art performance. Both code and dataset are available at: https://github.com/TongkunGuan/RFN.

Index Terms—Text detection, industrial scene, MPSC dataset, SynthMPSC dataset, text recognition.

I. INTRODUCTION

T

HE goal of text detection is to localize the text regions with bounding boxes, which mainly includes horizontal texts, multi-oriented texts, and curved texts in various scenarios. With the advent of laser marking technology, many metal parts are marked with Latin characters and Arabic numerals to record the serial number, production date, and other product information. Detecting these texts plays an increasingly important role in intelligent industrial manufacturing, which is conducive to improving the assembly speed of industrial production lines and the efficiency of logistics transmission in the industrial scene. Compared with the natural scene text detection (e.g., traffic signs, shopping mall trademarks, and billboards), industrial scene text detection has low visual contrast, corroded surfaces, and complex backgrounds. Thus, these characteristics pose greater challenges to industrial text detection. Specifically, the differences between industrial scenes and natural scenes are shown in Fig. 1.

Conventional scene text detection (STD) methods firstly extract regions of interest using shape detectors [1], [2] and then search the text boxes, while the existing ones are mainly based on deep neural networks and consist of three categories: segmentation-based methods, regression-based methods, and a combination of segmentation and regression methods. Most segmentation-based text detection methods [3]–[10] adopt semantic segmentation to perform pixel-level classification (i.e., text/non-text prediction) and group these pixels belonging to text to generate bounding boxes. However, the text edges of metal parts are less clear than natural scene texts. A segmentation network that eliminates background noise causes misclassification of foreground in low-contrast industrial images, which leads to inaccurate localization during post-processing. Regression-based text detection methods [11]–[20] mainly establish geometry metrics on text boxes and calculate regression loss for localizing texts. These methods use one-stage or two-stage detectors to implement text localization. The methods based on a one-stage detector run faster but have
lower accuracy. The methods based on a two-stage detector generate preliminary detection boxes by a region proposal network (RPN) [21] and then select better boxes to feed into a refinement network according to confidence scores. Although the two-stage detectors correct the location of each box, the candidate box quality on metal parts still needs to be improved. Specifically, our method contains more foreground boxes at the same number of candidate boxes compared to RRPN++. Note that a candidate box is defined as foreground box if its center point locates in a groundtruth.

Therefore we propose a refined feature-attentive network (RFN) by utilizing more foreground information to improve localization accuracy. Specifically, we first design a segmentation-based foreground-focus module (SFF) to generate a high-quality attention map with more detailed information of insensitive character areas. The SFF module adaptively integrates multi-resolution features to enhance the perception of multi-scale text features at each scale-specific layer. Then, an attentive proposal refinement module (APR) applies the attention map to construct high-quality foreground boxes, which generates discriminative classification and regression results for localization correction. Finally, experiments demonstrate that our method achieves the state-of-the-art performance on the MPSC dataset and robustly detects horizontal texts, multi-oriented texts, and multi-language texts on the MSRA-TD500 [23], USTB-SV1K [24], ICDAR2013 [25], and ICDAR2017-MLT [26] public datasets.

In addition, we contribute a benchmark dataset on metal parts for industrial text detection. To the best of our knowledge, it is the first industrial text detection benchmark dataset. Specifically, we build a metal part surface character dataset (MPSC) for industrial scenes and synthesize a SynthMPSC dataset based on real-world metal images. The MPSC dataset is the first industrial text detection benchmark dataset.

II. RELATED WORK

In this section, we review the development of text detection methods. Early methods adopt Fourier and Laplacian transform [27], SVM [28], connected components analysis (CCA) [23], [29], [30], maximally stable extremal regions (MSER) [31] and sliding window (SW) based classification [32]–[34] methods to implement text localization tasks. However, the above methods process text components in a bottom-up order, with long and slow pipelines. Later, inspired by general object detection, they utilize deep convolutional neural networks (CNNs) to generate a variety of text geometric metrics, such as bounding boxes, pixel-level masks, contour points, text centerlines, etc. These methods can be roughly divided into three categories: regression-based methods, segmentation-based methods, and combination segmentation and regression methods.

A. Regression-Based Methods

Regression-based text detection methods [11]–[20] predict the offsets from key elements and decode them into bounding boxes. Inspired by SSD [35], methods utilizing the pre-defined anchors (key element) simplify the detection pipeline, which are end-to-end trainable. By adding six text-box layers based on SSD, Liao et al. propose Textboxes++ [12] to predict the offsets from the pre-defined anchors composed of different aspect-ratios and scales. Similarly, Wang et al. [13] first design prior quadrilateral sliding windows for locating multi-oriented texts, which are different from horizontal sliding windows. Ma et al. [14] further add an angle specification into anchor strategy to generate rotating region proposals, which matches text instances of arbitrary orientations. However, single-stage detectors generate increasing failure examples on the cluttered background, which degrades text detection performance. Consequently, different refinement methodologies are adopted to optimize localization results. Similar to two-stage object detection methods, text detectors [14], [15] extract text features from the proposals generated by an RPN-like mechanism and then adopt ROI pooling [21] or RoIAlign [36] to obtain fixed-scale feature maps. The branches of boxes classification and boxes regression are finally utilized to correct the localization results of each proposal. On the basis of two-stage detectors, Zhang et al. [16] and Yang et al. [17]
propose iterative refinement modules to implement text detection correction and improve localization precision. Moreover, Zhou et al. [18] adopt an anchor-free strategy to realize geometry-aware localization. It generates boxes by predicting box edge distances from the current pixel (key element) to the minimal bounding rectangle of its text instances, and then combine the score map to detect the arbitrary-oriented text.

**B. Segmentation-Based Methods**

Inspired by FCN [3], many segmentation-based methods [3]–[10] adopt semantic segmentation and instance segmentation for text detection. He et al. [4] adopt cascaded convolutional networks to implement coarse-to-fine segmentation based on text instances and the centerline of text lines for text localization. Deng et al. [5] segment text/non-text by linking pixels in the same instance and conduct post-processing to extract text bounding boxes without location regression. Wu et al. [6] implement text detection by introducing a border class, and a lightweight FCN is applied to cast each pixel into three categories: text, border, and background. Tian et al. [7] optimize a shape-aware loss to distinguish the pixels among different text instances by embedding a space vector for each pixel. Specifically, it maximizes the Euclidean distances of pixel embedding vectors from different text instances and minimizes those belonging to the same instance. Wang et al. [8] gradually expand the minimal scale kernel size and increase the segmentation area for detecting text instances of arbitrary shapes. Liao et al. [9] develop a differentiable binarization (DB) algorithm for the segmentation network, which performs binarization with an approximate step function and makes the process end-to-end trainable. Instead of setting the fixed thresholds, the segmentation network adds an adaptive threshold map per image to provide a highly robust text feature map.

**C. Combination of Segmentation and Regression Methods**

The combination segmentation and regression methods are proposed to improve the effect of text detection lately. He et al. [37] exploit a regional attention mechanism to predict locations and scores of text boxes. Based on Mask R-CNN [36], Xie et al. [38] merge a text-context module into joint multi-scale pyramid features in order to suppress false alarms and reduce the number of false-positive boxes. Similarly, Huang et al. [39] update feature extraction network derived from Pyramid Attention Network [40], and add a text mask prediction branch that detects curved texts. In addition, Yang et al. [41] and Dai et al. [42] use a fully convolutional instance-aware semantic segmentation (FCIS) [43] method to guide the prediction of three text-related elements: mask, class, and box, by generating an instance-aware segmentation map. Wang et al. [44] propose an Adaptive-RPN with a scale-insensitive metric to accurately generate proposal bounding boxes, and then add contour characteristic of text regions by executing the convolution operation in two orthogonal directions to locate texts with arbitrary shapes.

### III. MPSC & SYNTHMPSC DATASET

The publicly available text detection datasets are mainly taken from natural scenes, and no industrial datasets can be explored and researched in the community. In this section, we establish a benchmark dataset (Metal Part Surface Character Dataset, MPSC) to promote in-depth research on text detection in industrial scenes. Specifically, our dataset spans many challenges affected by four factors (i.e., metal properties, industrial circumstances, scene noise, artificial design) as shown in Fig. 3. For instance, low visual contrast, uneven illumination, corroded surfaces, dirty part surfaces, blurred texts, clutter background, polymorphic characters, and multi-orientations. Moreover, we build an artificial metal part text dataset (Synthesized Metal Part Surface Character Dataset,
Fig. 4. The generation procedure of SynthMPSC dataset. (Top) Four flowcharts generated by the algorithm. The RGB image is first predicted to generate a depth map and a segmentation map, and then the segments suitable for placing texts in the segmentation map are defined as text regions to synthesize characters. (Bottom) Some synthetic images on the SynthMPSC dataset.

SynthMPSC) by synthesizing characters with real-world metal part images.

A. MPSC Dataset

By fully considering different types and styles of characters and metal parts, we collect a metal part surface character (MPSC) dataset. Specifically, 3194 images are constructed into the MPSC dataset, including 2555 training images and 639 testing images.

1) Dataset Construction: We perform industrial data deduplication, data cleaning, and data labelling on the collected images for three months, to promote the industrial application of text detection to a new stage. First, each image needs to be scrutinized that simple images and unqualified images are removed. Refer to ICDAR 2015 incidental text dataset [45], qualified images are then labelled with quadrilaterals at word-level where the four corners must be arranged clockwise. Finally, three rounds of inspections are implemented to reduce manual labelling errors.

2) Dataset Property Analysis: The MPSC dataset provides high-quality ground-truth boxes and text transcriptions. Most of them contain special combination rules that are different from the legal spelling of words, such as “AlSi9Cu3”, “D151C-050506”, and “7M121”. Each label has a specific implication, which embodies the signification of character encoding in industrial scenes. In addition, sufficient statistical results are calculated to show more information about the MPSC dataset. First, the number of characters per text instance is distributed between 1 to 31, with the majority ranging from 2 to 7, and 5.5 is the average. Second, the aspect ratios greater than 1 account for 70.6% of all text instances, while 1 and other aspect ratios account for 19.2% and 10.2%, respectively. Finally, the width of 92.9% of text instances is no more than 40% of the image width, while the height of 97.8% of text instances is less than 20% of the image height. Therefore, the area of most text instances is less than 8% of the image area.

B. SynthMPSC Dataset

1) Motivation: The self-built dataset fully considers the possibility of character structures and metal parts, whereas some attributes, e.g., character types, aspect ratios, area ratios, and directions, may exist uneven distribution as other public datasets, which limits the capability of sophisticated methods in real-world scenarios. Though current transfer neural networks [46]–[48] are promising to reduce domain shift to real texts, we use a simple syntheText algorithm [49] to generate large batches of synthetic images with rich text attributes.

2) Data Synthesis: We collect 1,153 metal images without characters to synthesize the SynthMPSC dataset containing 98,962 images and 1933234 text instances. Fig. 4 describes the visualization process of the SynthMPSC dataset. Specifically, the generation process of the SynthMPSC dataset starts with sampling images and texts. First, we select an image without characters that accord with the metal background characteristics, and predict its dense depth map. Then, we segment the image into multiple regions by colour and texture cues [50], and the areas suitable for text placement are marked with random colouring. Finally, we extract the corpus from the Newsgroup20 dataset [51] and pick the appropriate fonts and colours to synthesize images of metal parts with texts.

C. Comparison With Other Public Datasets

Public datasets ICDAR2013 [25], ICDAR2015 [45], MSRA-TD500 [23], and USTB-SV1K [24] are widely used in text localization tasks. They are taken from natural scenes, including traffic signs, shopping mall trademarks, billboards, etc. These texts have relatively clear texts with variable styles and colours against a chaotic background. For example, merchants want their trademarks to be more colorful and distinctive for attracting customers’ attention. Advertisers use clear and bright texts to let readers understand product value straightforwardly.

1) ICDAR 2013: It is widely used to implement text detection and recognition tasks. With its text instances almost horizontal, images are annotated with rectangular boxes.

2) ICDAR 2015: It is taken from street-viewed scenes. The text regions are annotated by 4 vertices of the quadrangle.

3) MSRA-TD500: It uses line-level bounding boxes instead of word-level bounding boxes to annotate labels so that the entire dataset contains many large-scale and long text instances. These images contain both Chinese texts and English texts.

4) USTB-SV1K: It contains many low-resolution and blur images. These images are artificially blurred to a certain extent, and text instances are characterized by multiple orientations, views, and perspective distortion.
The quantitative and statistic comparison results between the proposed dataset and other benchmarks are summarized in Table I. The number of images in our MPSC is 6.91 times (i.e., 3, 194 vs. 462), 6.388 times (i.e., 3, 194 vs. 500), 2.129 times (i.e., 3, 194 vs. 1000) that in ICDAR 2013, MSRA-TD500, ICDAR 2015 and USTB-SV1K, respectively. Our MPSC dataset is the first industrial text detection benchmark dataset. It covers many challenges from the natural scene e.g., text detection benchmark dataset. It covers many challenges e.g., corroded surfaces, low contrast, and uneven illumination, in the industrial scene to the state-of-the-art methods.

IV. REFINED FEATURE-ATTENTIVE NETWORK

In this section, we propose an industrial text detection method to locate text robustly and effectively. First, we summarize the overall structure of our proposed method, then illustrate the details of the SFF, APR, and Re-scoring modules, and finally, introduce the loss function of our method.

A. Overall Pipeline

Our network mainly includes four parts: a ResNet-FPN backbone for extracting multi-scale features, a detection branch of classification and regression tasks, a semantic segmentation branch for highlighting foreground features, and an attentive proposal refinement module. In RFN, we first employ ResNet [52] with 50 layers as a backbone of FPN [53], which is used to extract the multi-scale feature maps. Then, we introduce a segmentation-based foreground-focus module to highlight and retain more text area features. They are fed into two structure-sharing and parameter-separating sub-networks for obtaining preliminary detection boxes. Next, an attentive proposal refinement module corrects the location deviation of candidate boxes in which high-quality preliminary detection boxes are selected by combining multi-scale attention maps. Finally, we establish a re-scoring mechanism to assess the quality of the correction boxes by combining instance and classification scores. We illustrate the specific details of the text detection network RFN in Fig. 5.

B. Segmentation-Based Foreground-Focus Module

The surface of metal parts has a complicated visual context, with similar texture, uneven illumination, and varying character structures. Thus a feature extraction network should provide robust multi-scale text features on complex metal surfaces. However, layers with different resolutions have perception differences in multi-scale text features. While a high-scale feature map represents more details to capture small objects, the low-scale feature map with more decisive semantic information is usually more suitable for large objects. Existing methods mainly adopt the bottom-up integration approach to learn text features, which weaken the perception of multi-scale text features at each scale-specific layer and may not provide robust feature representations on complex metal surfaces. To exchange the information across multi-scale representations, we design the following network to enhance the ability to fusion and complementarity between feature layers with different scales.

1) Network Design: We design a novel feature extraction module from complex metal backgrounds, adaptively fusing multi-resolution features to enhance the perception of multi-scale text features at each scale-specific layer. First, we employ the ResNet-FPN backbone [53] to extract multi-scale features, which are defined as \( X_i \), \( X_2 \), \( X_3 \), and \( X_4 \). With different resolutions \( s_i = (h_i, w_i) \), \( i = 1, 2, 3, 4 \), they are divided into low- (i.e., \( X_1 \)) and high- (i.e., \( X_2 \), \( X_3 \), and \( X_4 \)) levels to enhance text feature representations in different ways. It compensates for the deficiencies of the scale-specific layer and enhances the adaptability to complex variations.

For low-level input, we highlight text information to the semantic features. Specifically, the input \( X_1 \in R^{h_1 \times w_1 \times c} \) is first fed into several convolutional layers with BatchNorm and ReLU activation function. Followed by
average-pooling operation along the channel axis, the foreground response value naturally gets a high accumulation. Instead of the sigmoid function, the low-level attention map is activated by the exponential operation to expand the difference between the response weights of foreground and background due to similar texture. Finally, the attention map is broadcasted and element-wise multiplicated with the input images $X_1$ to get low-level response maps $L$.

For high-level input, we design a parallel structure to fuse multi-resolution feature maps by exchanging information mutually. Each subnet of the high-level input adaptively learns the features of adjacent subnets, which enriches the spatial features and retains more multi-scale text features. Taking $X_i \in R^{h_i \times w_i \times c}$, $i = \{2, 3, 4\}$, the parallel structure is summarized based on:

$$Y_k = \sum_{i=2}^{4} F(X_i, s_k),$$

where $F(X_i, s_k)$ means $X_i$ is upsampled or down-sampled operation from resolution $s_i$ to $s_k$. Specifically, upsampling operation refers to $1 \times 1$ convolutional layers followed by the bilinear sampler, while $3 \times 3$ convolutional layers with the stride of 2 are implemented for down-sampling. If $s_i = s_k$, $F(\cdot)$ represents $3 \times 3$ convolutional layers without sampling layer. We fuse multi-resolution $Y_k$ to generate the final high-level response map $H$ in the following:

$$H = \phi(T(Y_2, s_1) \parallel T(Y_3, s_1)) \parallel T(Y_4, s_1),$$

where $T(\cdot)$ refers to upsampling $Y_k$ from resolution $s_k$ to $s_1$, “$\parallel$” represents the concatenation along the channel axis, and $\phi(\cdot)$ means adopting the channel-wise attention mechanism assigns large responses for foreground features.

Multi-level text features (i.e., Low-level response map $L$, High-level response map $H$) are then fused to generate an attention map $A$, which provides rich and discriminative semantic information and endows higher foreground response values. It guides sub-networks of all levels to focus on text features. We implement the specific details as follows:

$$\hat{X}_i = X_i \odot (1 + e^{H(\Delta, s_i)}),$$

where $H(\cdot)$ refers to down-sampling the attention map $A$ from resolution $s_1$ to $s_i$.

Subsequently, each subnet $\hat{X}_i$ is fed into regression and classification branches respectively. They adopt the common structure with separate parameters, consisting of four 3x3 convolutional layers, followed by a 3x5 convolutional layer for oriented texts. Based on pre-defined anchors from the generation strategy of Fig. 6, each subnet $\hat{X}_i$ generates a total of $h_i \times w_i \times 8$ preliminary detection boxes denoted as $B_i$, and $h_i \times w_i \times 8$ confidence scores denoted as $S_i$.

2) Segmentation Loss: We propose a novel loss function to boost the segmentation result of the supervised attention map. Unlike the natural scene text, the text edge of metal parts is unclear, and the visual contrast is low, which brings challenges to the accurate distinction between foreground and background areas. Thereby we pay more attention to the foreground and take two original intentions for establishing the attention map’s loss mechanism in the order of priority. a) Include as many text features as possible. b) Minimize the amount of false detection. Specifically, assuming that the groundtruth foreground mask is $S_{gt}$, which can be constructed from the quadrilateral bounding boxes. We first use the Dice [54] as the auxiliary loss function to deal with the extremely unbalanced positive and negative samples, since the text areas of interest occupies only a very small region of the image.

$$L_d = 1 - \frac{2 \sum_{i=1}^{N} (\omega_i \omega_i^*)}{\sum_{i=1}^{N} (\omega_i) + \sum_{i=1}^{N} (\omega_i^*)},$$

where $N$ is the number of pixels in the attention map $A$, $\omega_i$ and $\omega_i^*$ are the confidence score of pixel $i$ in $S_{gt}$ and $A$, respectively. Then the coefficient of false negative and false positive can be calculated as $D_a$ and $D_b$, respectively.

$$\omega_d = \omega_i - \omega_i^*,$$

$$D_a = \frac{\sum_{i=1}^{N} I_{[\omega_i < \frac{1}{2}]}(1 - \omega_i^*)}{\sum_{i=1}^{N} (\omega_i^*)},$$

$$D_b = \frac{\sum_{i=1}^{N} I_{[-\omega_i > \gamma]}(\omega_i^*)}{\sum_{i=1}^{N} (\omega_i^*)}$$

Finally the loss function are designed as follows:

$$L_g = \begin{cases} D_a & \text{if } D_b < \Delta, \\ D_a + D_b - \Delta & \text{if } D_b \geq \Delta \end{cases}$$

$$L_{seg} = L_d + e^{-1\times\gamma\cdot L_d} \ast L_g,$$

where $\gamma$ means a balance parameter to adjust the ratio of $L_g$ and $L_d$, and $\Delta$ represents the threshold of allowable false-positive classification results in exchange for detecting more text features in the low-contrast and indistinguishable area.

C. Attentive Proposal Refinement Module

Most preliminary detection boxes cover the text instances incompletely, especially those with oriented rectangle shape in the industrial scene. To achieve better location accuracy, we propose a novel box selection algorithm by applying attention maps to mine high-quality candidate boxes attached to the foreground. More foreground boxes are extracted in a high-priority order and fed into the refinement network.

1) Box Selection Algorithm: Given a set of the prediction boxes with scores $D = \{(B_i, S_i)\mid i = 1, \ldots, l\}$, our goal is to select top $\beta$ boxes and apply them into the refinement network. First, we binarize the supervised attention map $A$ to obtain the mask map $F$. Then, the $F$ will be scaled into...
the map \( F_i \) at each resolution \( s_i \) to filter out the invalid anchor points and only keep those anchors falling on the predicted foreground regions. Specifically, the points with the pixel value of 1 in \( F_i \) can be gathered and form the set of \( V = \{ V_i | i = 1, \ldots, I \} \). Each point in \( V \) corresponds to 8 candidate boxes with different aspect ratios, and the optimal box is selected according to the confidence score. Therefore, we effectively filter background boxes and obtain a multi-scale candidate box set \( \bar{V} \). Finally, the foreground boxes with the top \( \beta \) confidence scores are selected from \( \bar{V} \).

2) Refinement Network: Inspired by [14], the selected boxes are utilized to extract regions of interest (ROIs) as feature patches from the first four levels in the ResNet-FPN backbone. These feature batches are flattened and fed into a fully connected layer to form high-dimensional feature vectors, and then two fully connected layers are implemented to predict the classification and regression outputs for each box, respectively.

D. Re-scoring Mechanism

For standard post-processing such as Faster R-CNN [21], Mask R-CNN [36], the non-maximum suppression (NMS) process is implemented to retain the prediction boxes with the highest score for different objects. The confidence scores \( S_i \) are predicted by the classification branch for each proposal. However, this approach may ignore some prediction boxes without the highest classification scores but more accurate locations. We thus add an instance score \( S_i \) to each prediction box as follows:

\[
S_i = \frac{\sum_{j=1}^{N} \rho_j}{N}, \tag{10}
\]

where \( \rho_j \) represents the pixel value of the attention map \( A \). Compared to directly using weighted instance score \( S_i \) and confidence score \( S_i \), we adopt the below numeric formulation to form an overall score \( S' \), which has a higher gradient value under the same classification score:

\[
S' = e^{S_i} (1 + \mu \frac{e^{S_i}}{e^{S_i} - S_i}), \tag{11}
\]

where \( \mu \) is the trade-off coefficient. Finally, \( S' \) is taken as the new confidence score and fed into the NMS algorithm to get the best prediction boxes.

E. Loss Function

The overall loss function of RFN consists of \( L_{seg} \), \( L_{det} \), and \( L_{ref} \). Firstly, the output attention map of the SFF, APR, and Re-scoring modules on the MPSC dataset. We then test RFN on other public scene text datasets and compare them with state-of-the-art methods to demonstrate its robustness. Finally, we conduct an ablation study of the SFF, APR, and Re-scoring modules on the MPSC dataset.

A. Implementation Details

1) Evaluation Metrics: To fairly compare with other methods, we evaluate the proposed method on the MPSC, MSRA-TD500, USTB-SV1K, ICDAR2013 and ICDAR2017-MLT datasets, using the standard evaluation protocol proposed in [24]–[26], [45], [56], respectively. All experiments are implemented on a server with an NVidia Tesla V100 (32G) GPU.

2) Parameters Settings: Our proposed method is optimized by stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 1×10-4. The image size is set to 768×768, and the batch size is set to 12. A multi-step learning rate strategy is adopted to update weights, in which the initial learning rate is set to 0.001 and halved every 50 epochs. In the experiment, \( \Delta \) is set to 0.01 × \( S_{gt} \), \( \gamma \) is set to 0.1, \( \mu \) is set to 0.5, and \( \beta \) is set to 1000.
B. Performance Evaluation on MPSC Dataset

We implement comparative experiments to verify the effectiveness of the proposed method on the MPSC dataset compared with the state-of-the-art text detection methods [8], [14], [18], [19], [22], [36], [40], [44], [58]. These methods design novel feature representations and achieve excellent performance in multi-oriented scene text detection. As shown in Table II, our method achieves the best performance with an F-measure of 86.21% on the MPSC dataset and outperforms the currently best method [22] among the nine methods by 1.51% in the precision metric. Moreover, the recall metric of our method ranks only second to RRPN++, which adds an extra recognition branch to recall low-score prediction boxes with high recognition scores. Note that RFN can also deploy the recognition branch to improve detection performance.

To promote the proposed method, we pre-train the RFN network on SynthMPSC dataset and then fine-tune it on the MPSC dataset. Specifically, the last row of Table II represents the final result, reaching 87.05% in F-measure. Compared to training only on the MPSC dataset, the performance of RFN* improves by 0.84%, verifying that the artificially synthesized samples enhance the ability to detect characters of metal parts. Some qualitative results on the MPSC dataset are shown in Fig. 7.

Considering the practical application of deep learning, the quality of text detection results directly determines the end-to-end text recognition rate. The accurate bounding boxes provide rich information for the text recognition network. Thus different from general evaluation metrics, we change the fixed IOU threshold to calculate the number of matched boxes from the best model of each method. Specifically, the predicted bounding box is defined as a matched box if the IOU value between it and one of the ground-truth boxes is greater than the artificially set threshold (we set it to 0.6 and 0.8, respectively). The number of the matched bounding boxes obtained by different text detection methods is shown in Fig. 8. As represented by green bars, RFN generates the matched bounding boxes with the largest number. It implies that the APR module significantly corrects deviations in text localization and generates more high-quality bounding boxes.

In general, the statistical results demonstrate the effectiveness of our proposed method from another perspective and provide substantial improvements for the subsequent text recognition task, which is beneficial to the tracking of metal parts in industrial scenes.

Affected by low visual contrast, corroded surfaces, complex backgrounds, etc., the texts are not salient and have low visibility in the industrial scene image, further limiting the ability of text detection algorithms to be deployed in real-world scenarios. Thus we focus on the text foreground feature...
to weaken the influence of other factors. Specifically, SFF guides the framework to obtain more foreground information of metal parts by learning adaptive feature representations. The detailed discussions about SFF are given in the following: 1) Scale-sensitive feature fusion. The multi-resolution features are divided into low- and high- levels to enhance the text perception. The low-level aggregates foreground features along the channel axis, and the exponential operation is adopted to activate the foreground feature response. The high-level adaptively learns the features of adjacent subnets by exchanging information mutually, which enriches the spatial features and retains more multi-scale text features. 2) Foreground feature optimization. We establish a loss mechanism that imposes large weights to focus on foreground prediction results. And a threshold of allowable false-positive classification results is set in exchange for detecting more text features in the low-contrast and indistinguishable area. Moreover, the foreground prediction result generated by the SFF module is applied to the entire subsequent localization process, including the one-stage detector and refinement network. For the detailed discussions, APR is developed to construct high-quality foreground boxes. These boxes attached to the foreground are extracted in a high-priority order, and many boxes belonging to the background have a low opportunity to be re-corrected, which promotes the number of region-of-interest features for bounding box correction. Some examples of comparison results with other methods are shown in the Fig. 9.

C. Performance Comparison on Public Datasets

Our method is evaluated on these typical benchmark datasets to demonstrate its robustness. Similar to most scene text detection methods, we also pre-train our method on the SynthText dataset. Although SFF and APR modules are designed for text detection of the metal parts, the below experimental results still illustrate it gets comparable performance for other scene text detection. We report the best results of the comparison methods, each of which was reported in the original paper.

1) Detecting Multi-Oriented Text: MSRA-TD500 has become one of the most challenging multi-oriented text datasets with very few training samples and super-long large text instances. Covering the text area more completely and appropriately is the biggest challenge. Hence we evaluate our method on the MSRA-TD500 benchmark dataset to verify its ability to detect multi-oriented texts. Table III reports our results and compares them with the state-of-the-art methods. Our method obtains the best precision and F-measure among all comparison methods. Specifically, RFN achieves a precision of 88.4%, recall of 80.0%, and F-measure of 84.0% without extra data training. Compared to PAN with ignoring difficult samples to evaluate, our method further achieves 88.1% in F-measure.

2) Detecting Horizontal Text: To verify the robustness to the horizontal texts, we select the popular ICDAR2013 dataset to evaluate the performance of RFN. Table IV reports our results and compares them with the state-of-the-art methods. It is observed that our result reaches the precision of 92.5%, recall of 90.7%, and F-measure of 91.6%, outperforming most existing methods even if some of them are tested in multi-scales. It is worth noting that RFN only adopts single-scale images for training and testing, and no other tricks are employed to improve performance. It proves that RFN is robust for text detection with horizontal texts in natural scenes.
3) Detecting Multi-Language Text: The ICDAR2017-MLT dataset has a wide range of resolutions and includes many small and dense texts from 9 languages. To fully mine small samples, we use the high resolution to test ICDAR2017-MLT and set the number of candidate boxes \( \beta \) to 2500. In addition, the ratio aspects of the default boxes are set to \{1, 2, 3, 5, 7.5, 1/2, 1/4, 1/6\} to adapt to the diversity of text instances. Table V reports our results and compares them with the state-of-the-art methods. The proposed method achieves competitive performance with an F-measure of 73.0%. Although each language has its specificity, RFN can successfully locate the text instances with multiple languages, demonstrating the robustness in multilingual scenes.

4) Detecting Low-Resolution Text: To further evaluate the generalization ability of RFN, we select the challenging USTB-SV1K dataset with many low resolution and blurred images. Table VI reports our results and compares them with the state-of-the-art methods. Without pre-training on the SynthText dataset, our method still achieves the best results, outperforming Wang et al.’s method (the best-reported result currently) by 3.2% in the F-measure metric. Overall, expensive experiments illustrated that our method robustly detects multi-oriented texts, horizontal texts, multi-language texts, low-resolution texts, and can be deployed in more complex scenes.

### Table IV
Comparison Results on the ICDAR2013 Dataset

| Algorithms   | Precision (%) | Recall (%) | F-measure (%) |
|--------------|---------------|------------|---------------|
| SegLink* [59] | 92.0          | 84.4       | 88.1          |
| SSTD [37]    | 89.0          | 86.0       | 88.0          |
| TextBoxes++ [12] | 92.0      | 86.0       | 89.0          |
| FOTS [67]    | -             | -          | 87.3          |
| RRD* [61]    | 92.0          | 86.0       | 89.0          |
| PixelLink* [5] | 88.6        | 87.5       | 88.1          |
| RPRN [14]    | 84.0          | 77.0       | 80.0          |
| Melinda et al. [68] | 93.9    | 91.5       | 92.6          |
| FTPN [69]    | 93.2          | 91.9       | 92.5          |
| Liu et al. [70] | 90.2        | 86.3       | 88.2          |
| Wei et al. [71] | 93.7        | 87.4       | 90.4          |
| RFN (ours)   | 92.5          | 90.7       | 91.6          |

* means training with multiple scales.

### Table V
Comparison Results on the ICDAR2017-MLT Dataset

| Algorithms   | Precision (%) | Recall (%) | F-measure (%) |
|--------------|---------------|------------|---------------|
| SenseTime OCR [26] | 56.9        | 69.4       | 62.6          |
| FOTS [67]    | 81.0          | 57.5       | 67.3          |
| FOTS* [67]   | 81.9          | 62.3       | 67.3          |
| LOMO [16]    | 78.8          | 60.6       | 68.5          |
| LOMO* [16]   | 80.2          | 67.2       | 73.1          |
| PSENet [8]   | 73.8          | 68.2       | 70.9          |
| PSENet* [8]  | 75.4          | 69.2       | 72.1          |
| CharNet [72] | 77.1          | 70.1       | 73.4          |
| CRAFT [64]   | 80.6          | 68.2       | 73.9          |
| UrealText [73] | 82.2        | 67.4       | 74.1          |
| ISNet [74]   | 78.0          | 67.4       | 72.3          |
| RFN (ours)   | 79.4          | 67.6       | 73.0          |

* means training with multiple scales.

† means the model uses ResNet152 as the backbone.

### Table VI
Comparison Results on the USTB-SV1K Dataset

| Algorithms   | Precision (%) | Recall (%) | F-measure (%) |
|--------------|---------------|------------|---------------|
| Liu et al.† [70] | 72.3          | 50.3       | 59.3          |
| FTPN [69]    | 61.4          | 63.8       | 62.6          |
| Wang et al.† [75] | 73.0        | 67.0       | 70.0          |
| RFN (ours)   | 80.3          | 67.2       | 73.2          |

† indicates that the method adds extra data to train.

Fig. 10. Precision-recall curves. We evaluate the robustness of the two combinations (“SFF” vs “SFF+APR”, “RFN” vs “RFN*”), separately.

### Table VII
Evaluate the Effectiveness of the MPSC Dataset in the Proposed Modules of SFF, APR, Re-Score

| SFF | APR | Re-score | Precision (%) | Recall (%) | F-measure (%) | \( \Delta F \) (%) |
|-----|-----|---------|---------------|------------|---------------|-----------------|
| ✔   | ✔   | ✔       | 82.41         | 79.22      | 80.78         | -               |
| ✔   | ✔   | ✔       | 85.19         | 82.65      | 83.90         | 3.12%           |
| ✔   | ✔   | ✔       | 85.44         | 83.09      | 84.25         | 3.47%           |
| ✔   | ✔   | ✔       | 89.18         | 83.19      | 86.08         | 5.30%           |
| ✔   | ✔   | ✔       | 89.30         | 83.33      | 86.21         | 5.43%           |

\( \Delta F \) is the improvement of F-measure relative to baseline.

### D. Ablation Study
We implement an ablation experiment for our method on the MPSC dataset. Specifically, we analyze the influence of the model structure on performance for the MPSC dataset. The model is split into five combinations and trained separately to verify the effectiveness of the proposed method. The comparative test results are shown in Table VII.

1) Segmentation-Based Foreground-Focus Module (SFF): The SFF module improves the accuracy rate and recall rate by 2.78% and 3.43%, respectively. It proves that the foreground-focus mask branch is suitable for the MPSC dataset and effectively highlights the text features in the complex background of metal parts. Compared to the baseline, the mask branch integrates the high-quality attention map into the regression network to generate discriminative feature representations, which optimizes the geolocation parameters and provides more competitive prediction boxes. The attention map is the primary factor in improving text detection performance in the MPSC dataset.

2) Attentive Proposal Refinement Module (APR): The combination of SFF and APR further improves the text detection performance, reaching 86.08% in F-measure. It shows that APR significantly improves the precision metric of detection.
performance, which first applies the attention map to mine candidate boxes in complex backgrounds of metal parts. On the one hand, these boxes attached to the foreground are extracted in a high-priority order and fed into the refinement network. It means that many boxes belonging to the background have a low opportunity to be re-corrected, reducing the number of false positives (FP). On the other hand, multi-scale detection boxes with high classification scores in the first stage are selected to improve the quality of candidate boxes. Ideally, a good candidate box can accurately cover the text area, so there is no need to re-correct the location deviation. Therefore, high-quality candidate boxes promote the refinement network to generate more accurate prediction boxes with high IOU values to ground-truth boxes. It increases the number of true positives (TP). As shown in Fig. 8, RFN generates more prediction bounding boxes with high IOU. The increase in TP and the decrease in FP improve the precision metric.  

3) Re-Scoring Mechanism: The re-scoring module responds to more accurate prediction boxes by adding instance scores and positively affects all combinations in Table VII. Some prediction boxes with low classification scores and high instance scores are kept as final bounding boxes to increases the number of true positives (TP). Compared to SFF+APR, the re-scoring module has a more significant impact on SFF, side reflecting that APR improves the overall location accuracy of the prediction boxes. More importantly, the precise location can improve text recognition performance and facilitate metal parts tracking while maintaining measurement indicators.

Finally, we draw the precision-recall curve of text detection on the MPSC dataset as shown in Fig. 10 to illustrate the whole performance of the model.

4) Recognition Head: To prove the effectiveness of the RFN method, we add two groups of control experiments. First, we cancel the recognition head of RRPN++. As shown in the second and fourth rows of Table VIII, the recall rate is 1.36% lower than that of RFN, which demonstrates that our method achieves the best performance among the comparative methods above. Then the recognition branch is added to our proposed method to improve the detection performance. As shown in the third and fifth rows of Table VIII, our recall rate is 0.52% higher than RRPN++ and exceeds its accuracy metric by 2.73%.

E. Limitation

Some failure examples cause performance reduction. Specifically, one class of failed examples happened at extremely low-resolution industrial images where our RFN method exists false negatives in the low-salient text regions as shown in the first row of Fig. 11. Another class listed in the second row is mislocated sentence-level and word-level texts affected by the spacing between words due to a non-unified standard for the spacing between sentences, words, and characters among the various labels.

VI. CONCLUSION

In this paper, we propose an effective text detection method, RFN, to locate text instances of metal parts in industrial scenes. By designing the SFF, APR, and re-scoring modules, RFN is robust to tackle location deviation problems in the complex background. Experiments demonstrate that our method achieves the state-of-the-art performance on the MPSC dataset. Second, our method effectively detects multi-oriented, horizontal, and multi-language texts and gets competitive performance on public benchmark datasets, indicating the generalization ability to be deployed in more complex scenes. Third, we contribute two benchmark datasets of metal parts (MPSC and SynthMPSC dataset) dedicated to industrial text detection research. To the best of our knowledge, these are the first industrial text datasets. In the future, we will apply text detection in metal parts tracking, involving a text recognition task to record metal parts information in industrial production lines.

REFERENCES

[1] C. Lu, S. Xia, M. Shao, and Y. Fu, “Arc-support line segments revisited: An efficient high-quality ellipse detection,” IEEE Trans. Image Process., vol. 29, pp. 768–781, 2020.
[2] C. Lu, S. Xia, W. Huang, M. Shao, and Y. Fu, “Circle detection by arc-support line segments,” in Proc. ICIP, 2017, pp. 76–80.
[3] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 640–651, Apr. 2017.
[4] T. He, W. Huang, Y. Qiao, and J. Yao, “Accurate text localization in natural image with cascaded convolutional text network,” 2016, arXiv:1603.09423.
[5] D. Deng, H. Liu, X. Li, and D. Cai, “PixelLink: Detecting scene text via instance segmentation,” in Proc. AAAI, 2018, pp. 6773–6780.
[6] Y. Wu and P. Natarajan, “Self-organized text detection with minimal post-processing via border learning,” in Proc. ICCV, 2017, pp. 5010–5019.
[7] Z. Tian et al., “Learning shape-aware embedding for scene text detection,” in Proc. CVPR, 2019, pp. 4229–4238.
[8] W. Wang et al., “Shape robust text detection with progressive scale expansion network,” in Proc. CVPR, 2019, pp. 9336–9345.
[66] C. Yao, X. Bai, and W. Liu, “A unified framework for multi-oriented text detection and recognition,” *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4737–4749, Nov. 2014.

[67] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, “FOTS: Fast oriented text spotting with a unified network,” in *Proc. CVPR*, 2018, pp. 5676–5685.

[68] L. Melinda and C. Bhagvati, “Parameter-free table detection method,” in *Proc. ICDAI*, 2019, pp. 454–460.

[69] F. Liu, C. Chen, D. Gu, and J. Zheng, “FTPN: Scene text detection with feature pyramid based text proposal network,” *IEEE Access*, vol. 7, pp. 44219–44228, 2019.

[70] Z. Liu, W. Zhou, and H. Li, “Scene text detection with fully convolutional neural networks,” *Multimedia Tools Appl.*, vol. 78, no. 13, pp. 18205–18227, Jul. 2019.

[71] G. Wei, W. Rong, Y. Liang, X. Xiao, and X. Liu, “Toward arbitrary-shaped text spotting based on end-to-end,” *IEEE Access*, vol. 8, pp. 155906–155914, 2020.

[72] L. Xing, Z. Tian, W. Huang, and M. R. Scott, “Convolutional character networks,” in *Proc. ICCV*, 2019, pp. 9126–9136.

[73] S. Long and C. Yao, “UnrealText: Synthesizing realistic scene text images from the unreal world,” 2020, arXiv:2003.10608.

[74] P. Yang et al., “Instance segmentation network with self-distillation for scene text detection,” *IEEE Access*, vol. 8, pp. 45825–45836, 2020.

[75] X. Wang, X. Feng, and Z. Xia, “Scene video text tracking based on hybrid deep text detection and layout constraint,” *Neurocomputing*, vol. 363, pp. 223–235, Mar. 2019.

**Tongkun Guan** received the B.S. degree in electrical engineering and automation from Hunan University, Changsha, China, in 2020. He is currently pursuing the M.S. degree with the Key Laboratory of System Control and Information Processing, Shanghai Jiao Tong University. He has wide research interests mainly including computer vision, text detection, image processing, and text recognition.

**Chaochen Gu** received the bachelor’s degree from Shandong University, Jinan, China, in 2007, and the Ph.D. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2013. He is currently an Associate Professor at the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University. His current research interests include industry robotics, machine vision, and man–machine interfaces.

**Changsheng Lu** received the B.S. degree in automation from Southeast University, Nanjing, China, in 2017, and the M.S. degree in control science and engineering from Shanghai Jiao Tong University, Shanghai, China, in 2020. He is currently pursuing the Ph.D. degree with the College of Engineering and Computer Science, Australian National University. He has wide research interests mainly including computer vision, few-shot learning, transfer learning, image processing, pattern recognition, and robotics. Particularly, he is interested in the theories and algorithms that empower robot to see, think, and conduct more like a human. Previously, he was awarded the national scholarship and the outstanding graduate student at SEU and SJTU. He has served as a Reviewer of *IJCV*, *IEEE TRANSACTIONS ON IMAGE PROCESSING*, *IEEE Computational Intelligence Magazine*, *PR*, *IEEE ROBOTICS AND AUTOMATION LETTERS*, and *IJCVIR*.

**Jingcheng Tu** received the B.Eng. degree from the College of Automation, Xi’an Jiaotong University, Xi’an, China, in 2018. She is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. Her current research interest includes intelligent video analytics of edge-enabled Industrial Internet-of-Things.

**Qi Feng** received the B.S. degree in automation from Nanjing University, China, in 2010. He is currently pursuing the Ph.D. degree with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University. His research interests include computer vision, machine learning, and 3D scene understanding.

**Kaijie Wu** received the Ph.D. degree in biomedical engineering from Tianjin University, Tianjin, China, in 2006. He is currently an Associate Professor at the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University. His current research explores biomedical optical imaging, medical information processing, and pattern recognition.

**Xinping Guan** (Fellow, IEEE) received the B.Sc. degree in mathematics from Harbin Normal University, China, in 1986, and the Ph.D. degree in control and systems from the Harbin Institute of Technology, China, in 1999. He is currently a Chair Professor with Shanghai Jiao Tong University, Shanghai, China, where he is the Deputy Director of the University Research Management Office and the Director of the Key Laboratory of Systems Control and Information Processing, Ministry of Education of China. His current research interests include industrial cyber-physical systems and wireless networking and applications in smart factory. He is also the Leader of the prestigious Innovative Research Team, National Natural Science Foundation of China (NSFC). He is an Executive Committee Member of the Chinese Automation Association Council and the Chinese Artificial Intelligence Association Council. He was a recipient of the First Prize of Natural Science Award from the Ministry of Education of China in 2006 and 2016 and the Second Prize of the National Natural Science Award of China in 2008. He is a “National Outstanding Youth” honored by NSFC and the “Changjiang Scholar” by the Ministry of Education of China.