IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 20, 2021, accepted March 1, 2021, date of publication March 9, 2021, date of current version March 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3064564

A Neural Network Model for Text Detection in Chinese Drug Package Insert

HAIWEN WU, RI-GUI ZHOU, (Member, IEEE), AND YAOCHONG LI
College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
Research Center of Intelligent Information Processing and Quantum Intelligent Computing, Shanghai 201306, China

Corresponding author: Ri-Gui Zhou (rgzhou@shmtu.edu.cn)

This work was supported by the Shanghai Science and Technology Project in 2020 under Grant 20040501500.

ABSTRACT The text information in the medical photocopies is of great significance to the construction of medical digital platform. Text region detection, the very first step of extracting medical photocopies information, is functional to detect text area or locate text instance on the sample. Researchers have done a lot works on text area detection in natural scenes, yet few of them in turn pay attention to the medical photocopies scenario which is urgent to be settled. Here, a text line area detection dataset based on Chinese medical photocopies (CMPTD) are created and a fine-grained text line region detection model based on multi-scale feature extraction and fusion are proposed in this paper. The detection model consists of three parts. The first part is feature extraction module. Cspdarknet53 in You Only Look Once version 4 (YOLOv4) is used as the backbone network of our model, and the spatial pyramid pool strategy is used to extract multi-scale features to enhance the robustness of the model. The second part is feature fusion module. By referring to the PANet structure, the three effective feature layers in feature extraction module are fused repeatedly. The last part is prediction module. The network outputs a series of fine-grained text proposals by referring to the CTPN structure, which are connected into text lines by text line construction algorithm. We experimentally demonstrate the effectiveness of the detection model with the precision of 92.46% and the recall of 91.74% in the text detection task of the dataset CMPTD.

INDEX TERMS Chinese medical photocopying, text detection, YOLOv4, text line construction algorithm, convolutional neural network.

I. INTRODUCTION

With the development of artificial intelligence technology, more and more attention has been paid to digital construction, which is one of the key construction projects vigorously promoted by many countries in recent ten years. The pharmaceutical industry is no exception. In the past decade, the digital process of medical services has been introduced into some European countries. Even some have reached a advanced level, such as northern Europe [1]–[3]. In order to promote the use of electronic health records, some developed countries, such as the United States, have also formulated corresponding incentive policies [4], [5]. However, developing countries, such as China, still have some shortcomings in this respect [6]. At the same time, some application softwares have come out one after another. For example, the rational drug use monitoring software of Meikang company, the clinical drug consultation system of Datong company, the drug database of US FDA, etc.

A large number of unstructured medical paper documents produced in the process of long-term historical development are important medical data, which can be transformed into structured data through digital processing technology, so as to create resource database or learning database for medical industry and related personnel [7]. Based on the above background, the original intention of our work is to extract the unstructured data from medical photocopies such as Chinese drug package inserts and convert them into structured data, which mainly involves optical character recognition (OCR) technology.

Although OCR technology performs well in some scenes, such as machine translation and image retrieval [8], [9], text detection and recognition still face many challenges, such as the diversification of text in street scenes and the scanning of low-quality data such as books left in Google Books service [1], [10].

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeeb Dey.

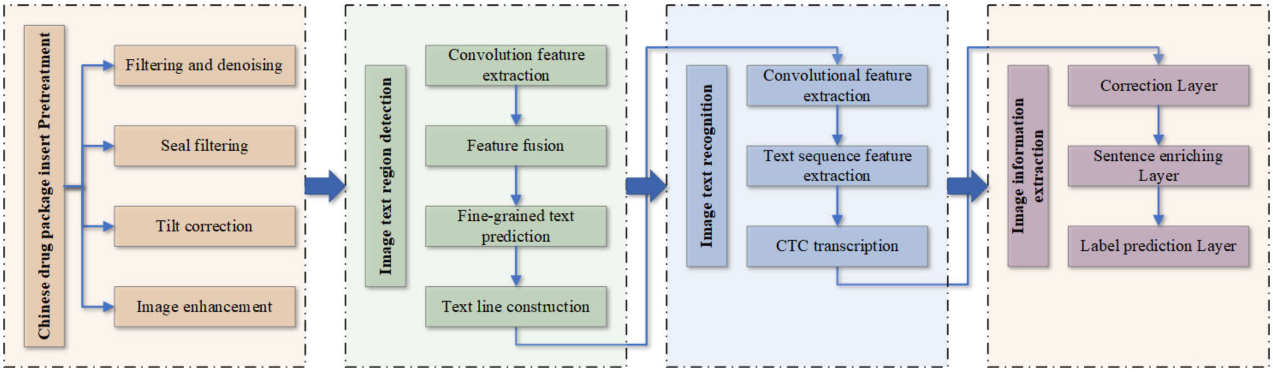

Chinese drug package insert Pretreatment
Filtering and denoising
Seal filtering
Tilt correction
Image enhancement
Image text region detection
Convolution feature extraction
Feature fusion
Fine-grained text prediction
Text line construction
Image text recognition
Convolutional feature extraction
Text sequence feature extraction
CTC transcription
Image information extraction
Correction Layer
Sentence enriching Layer
Label prediction Layer

FIGURE 1. Structured information extraction pipeline for Chinese drug package inserts. The entire process can be divided into four main steps, including preprocessing of drug package inserts, text area detection, text recognition and structured information extraction.

Our work is to extract structured information from the text of drug packaging inserts and store it in the database and finally complete the construction of the medical digital platform. The main process of our project is shown in Figure 1, which can be roughly divided into four parts:

1) Pretreatment of Chinese medicine package inserts: the quality of original medical photocopies may be poor due to acquisition equipment and other human factors, such as inclined text area, red seal interference, etc. In order not to affect the effect of subsequent text detection and recognition, we preprocess some poor quality medical photocopies, including denoising, seal filtering, tilt correction and image enhancement.
2) Image text region detection: text region detection is the main work of this paper, its goal is to detect the text region of the input image, for drug packaging instructions, this paper only focuses on its main content. At this stage, we divide the text detection into four parts: feature extraction, feature fusion, fine-grained text prediction and text line construction, which will be introduced in detail in a later part.
3) Image text recognition: image text recognition is one of the subsequent steps of text region detection. Its purpose is to convert the detected text region into a text string that can be understood by the computer. There are many text recognition algorithms [10]–[12], where CRNN [13] model is suitable for our text detection task. We divide it into three parts: convolution feature extraction, text sequence feature extraction and CTC transcription.
4) Image information extraction: The purpose of this stage is to correct the spelling and grammatical errors of the text information after recognizing the text, and to classify or convert it into structured data. It can be divided into three steps: correction layer, sentence enriching layer, label prediction layer. This is the work we have done [7].

This paper focuses on the text detection of Chinese medicine package inserts. Text detection is the foundation and premise of our whole project. An accurate and effective text detection model will greatly improve the subsequent steps of our project, such as text recognition and text structured information extraction. However, some text detection models that perform well in other text datasets (such as EAST [14] and CRAFT [15]) are not ideal when they are directly applied to the text detection task of drug packaging instructions. It is mainly reflected in two aspects. The first aspect is that the layout of drug packaging instructions is relatively complex, and contains many elements, such as Chinese font, Arabic numerals, text trademark, chemical structure of drugs, etc, which makes the detection difficult. The second aspect is that the text area of drug package inserts usually has a large aspect ratio, and the character spacing of text lines is quite different. The existing text detection model cannot detect the complete text line area well, which leads to a text line area often getting multiple detection boxes or missing some characters, which is not conducive to our subsequent work.

In order to solve these problems, in the text detection task, we propose a fine-grained text line region detection model for medical photocopy dataset. The main contributions of this paper are as follows:

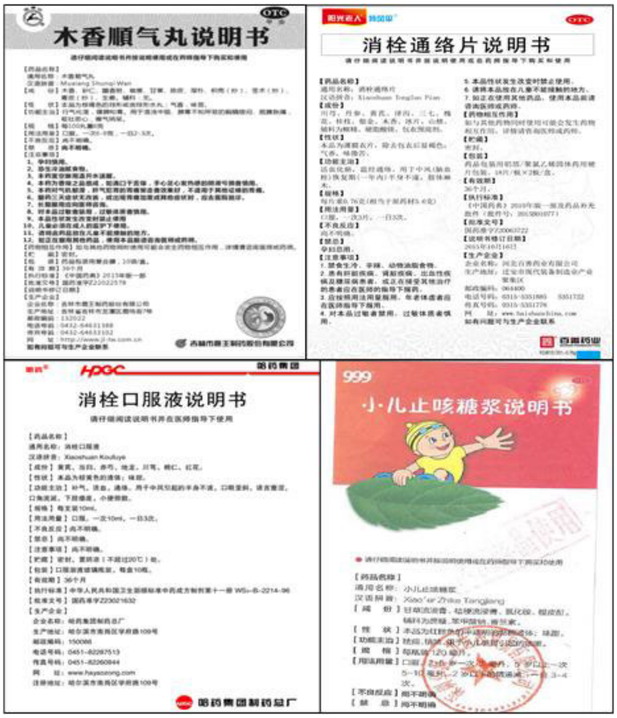
1) Since there are few text detection datasets in this field, we created a Chinese medical photocopies text detection dataset with reference to ICDAR2015 [16] text detection dataset, and named it CMPTD.
2) A fine-grained text region detection model based on YOLOv4 [17] and CTPN [18] is proposed, which can effectively detect the text region of drug packaging instructions.
3) An improved text line construction algorithm is used, the output prediction box can wrap the whole text line area completely and accurately.
4) Module experiments and comparative experiments are completed, which show the superiority of the proposed text detection model in CMPTD dataset.

II. RELATED WORK

This paper focuses on the text region detection of medical packaging instructions. There is a lot of recent work on text region detection, but it can be roughly divided into two kinds, one is the traditional text detection method based on manually designed features, the other is the text detection method based on deep learning. In traditional text detection, researchers mainly use bottom-up text detection method, and use artificial features [19]–[21] to detect strokes or characters in the image. For example, the texture based method [19] extracts stroke features of Chinese characters through a filter, and then uses support vector machine for texture classification to detect text regions. There are also region based methods [20], [21] which use some text features (such as stroke width, number of holes and other morphological information) to find candidate regions of text from complex background, and then get text lines according to some filtering rules.

The performance of traditional text detection methods largely depends on artificial features, and is not robust in different detection scenarios. With the rise of deep learning, researchers apply deep learning method to text detection task, which not only makes researchers get rid of the tedious manual design work, but also greatly improves the effect of text region detection. Recent text detection methods based on deep learning are mainly inspired by object detection [22]–[24] and semantic segmentation [25]–[27]. For example, CRAFT [15] uses the idea of segmentation to output text lines by detecting single characters and the connection relationship between characters. CTPN [18] improves Faster R-CNN [22], uses CNN and BLSTM to obtain text sequence features, and obtains complete text boxes by merging small text boxes belonging to the same line. The fine-grained text detection model in this paper is based on the same idea. The network predicts a series of small text boxes, and the text line construction algorithm connects the small text boxes belonging to the same line, and then outputs a series of complete text line regions.

The text detection method based on deep learning can be divided into three directions: method based on bounding box regression, method based on segmentation and hybrid method. Most of the methods based on bounding box regression [18], [28], [29] use convolutional neural network to extract features and directly predict the location, size and text score of bounding box. Segmentation based text detection method [15], [30], [31] attempts to classify or segment text regions directly from complex background at pixel level, and get the final text bounding box according to the classification or segmentation results. The hybrid method [32]–[34] may use the idea of segmentation and bounding box regression at the same time. Although it can achieve better text detection effect, the prediction speed of the network is likely to be reduced due to the tedious steps. In this paper, we use the text detection method based on bounding box regression, which is also the most common text detection method among the three methods. Our method predicts the regression of the bounding box by correcting the center coordinates, height and width of the anchor point, and then obtains the text area of the Chinese medicine packaging instructions.


木香顺气丸说明书

消栓通络片说明书

消栓口服液说明书

小儿止咳糖浆说明书

FIGURE 2. Some Chinese drug package inserts in the data set CMPTD.

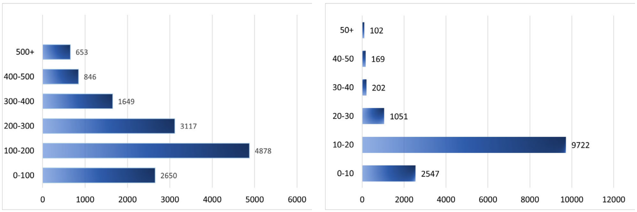
III. TEXT DETECTION DATA SET FOR CHINESE DRUG PACKAGE INSERTS

Due to the lack of text detection data sets about Chinese medical photocopies, in order to promote the research in this field, we refer to the data format of ICDAR2015 [16] to label the Chinese drug packaging instructions and create the Chinese medical photocopies text detection data set (CMPTD). The image resources of the data set come from China National Pharmaceutical Group Corp,[1] but due to the acquisition equipment or human factors, the quality of the original photocopies is very poor, such as low definition due to high exposure, text area tilt, red seal interference and so on. In order not to affect the text area detection and subsequent copy recognition, we preprocess some poor quality photocopies, such as seal filtering, tilt correction and image enhancement. Some Chinese drug packaging instructions in CMPTD dataset are shown in Figure 2.

After statistics, CMPTD data set has a total of 13793 text line instances, including 9531 text instances in training set and 4262 text instances in test set, accounting for about 31%.

In order to make our text detection model have better performance, we need some prior information of CMPTD dataset. Therefore, we calculate the minimum circumscribed rectangle of each text instance, obtain their length and height, and make mathematical statistics. Figure 3(a) represents the length distribution of text instances. It shows that in CMTD dataset, the length of text instances is evenly distributed in the range of 0-500px; Figure 3(b) shows the height distribution

[1] http://www.sinopharm.com/56.html



(a) (b)

FIGURE 3. (a) Length distribution of text line instances in cmptd dataset; (b) Height distribution of text line instances in cmptd dataset.

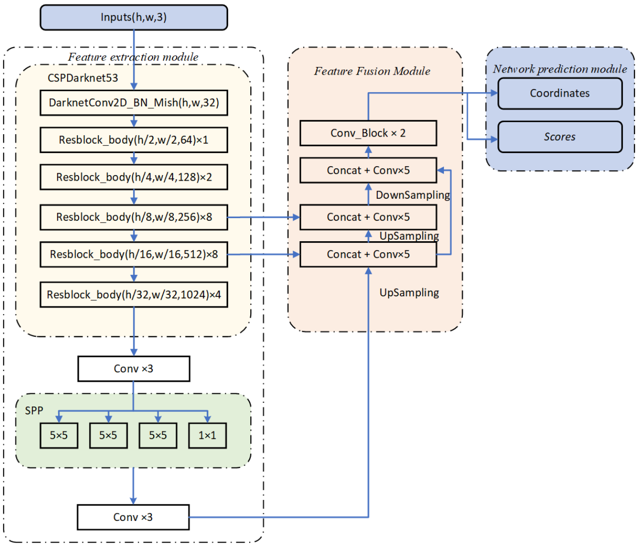
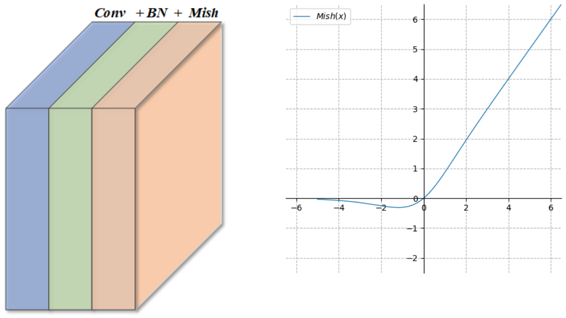

Inputs(h,w,3)
Feature extraction module
CSPDarknet53
DarknetConv2D_BN_Mish(h,w,32)
Resblock_body(h/2,w/2,64)×1
Resblock_body(h/4,w/4,128)×2
Resblock_body(h/8,w/8,256)×8
Resblock_body(h/16,w/16,512)×8
Resblock_body(h/32,w/32,1024)×4
Conv ×3
SPP
5×5
5×5
5×5
1×1
Conv ×3
Feature Fusion Module
Conv_Block × 2
Concat + Conv×5
DownSampling
Concat + Conv×5
UpSampling
Concat + Conv×5
UpSampling
Network prediction module
Coordinates
Scores

FIGURE 4. The fine-grained text detection model proposed In this paper. It can be divided into three modules: feature extraction module, feature fusion module and network prediction module.



(a) (b)

FIGURE 5. (a) The network architecture of convolution block DarknetConv2D_BN_Mish; (b) The change curve of Mish activation function.

of text instances, which shows that the height distribution of text instances is extremely uneven, and 96.9% of the text instances are below 30px. Some common target detection algorithms using anchor mechanism of fixed size, such as Faster R-CNN [22], SSD [23], may not be suitable for this situation. Therefore, this paper establishes a fine-grained text region detection model based on YOLOv4 [17] by referring to the idea of CTPN [18]. The network predicts a series of small-scale text boxes, and then is connected to obtain the text line region through the text line construction algorithm.

IV. THE PROPOSED METHOD

In this section, we will describe our proposed method in detail. Firstly, the fine-grained text detection model can be divided into three modules, namely feature extraction module, feature fusion module and network prediction module; secondly, according to the distribution of the length and height of CMPTD data set. We have improved the anchor mechanism, which will be introduced in detail in the following chapters; finally, this paper also introduces the multi task loss function used in the model and the improved text line construction algorithm will be introduced.

A. NETWORK STRUCTURE

Due to the excellent performance of YOLOv4 [17] in the task of target detection, we design a fine-grained text detection model based on Chinese medical packaging instructions with reference to the network architecture of YOLOv4 [17]. Figure 4 shows our proposed network architecture. It mainly includes three modules: feature extraction module, feature fusion module and network prediction module. These modules are described in detail below.

1) FEATURE EXTRACTION MODULE

In our proposed text detection network architecture, the feature extraction module is composed of CSPDarknet53 and SPP structure. As the backbone network of the model, CSPDarknet53 is used to extract the main features in Chinese medical packaging instructions. It is composed of the convolution block DarknetConv2D_BN_Mish and a series of residual network structure blocks Resblock_body. The network structure of the convolution block DarknetConv2D_BN_Mish is shown in Figure 5 (a). As shown, it is composed of a standard convolutional layer, a standardized layer and a Mish activation function layer. Mish activation function is a smooth non-monotonic activation function, which can help neural networks get better performance, and has been applied in many computer vision tasks [35]. The mathematical form of Mish activation function is shown in Eq (1), and its graph is shown in Figure 5(b).

$$Mish(x) = x * \tanh\left(\log\left(1 + e^{x}\right)\right). \tag{1}$$

The structure of the residual network structure block Resblock_body is shown in Figure 6, which is improved from the convolution block resblock_body of YOLOv3 [35]. It uses the CSPnet structure [36]. Its structure is not complicated. The principle is to split the original residual block stack into two branches, the main branch continues the original residual block (ResBlock) stacking. The other branch performs a little bit of processing first(Conv2D_BN_Mish), and then merges with the processing results of the main branch on the channel, Finally, the merged results are employed as the input of the

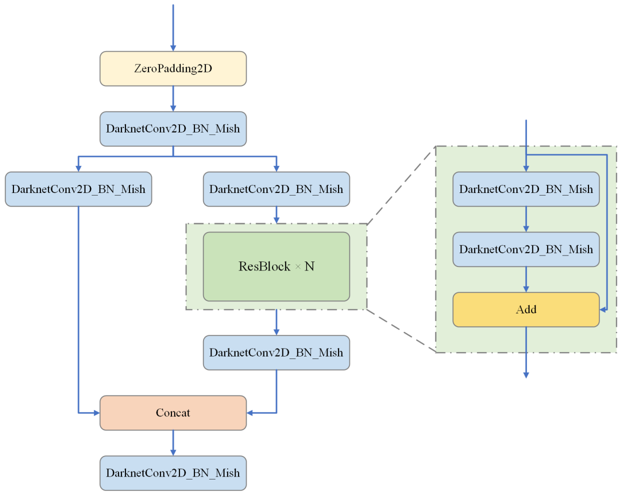


FIGURE 6. The structure of the residual network structure block Resblock_body, which is divided into two branches, the main branch and the other branch.

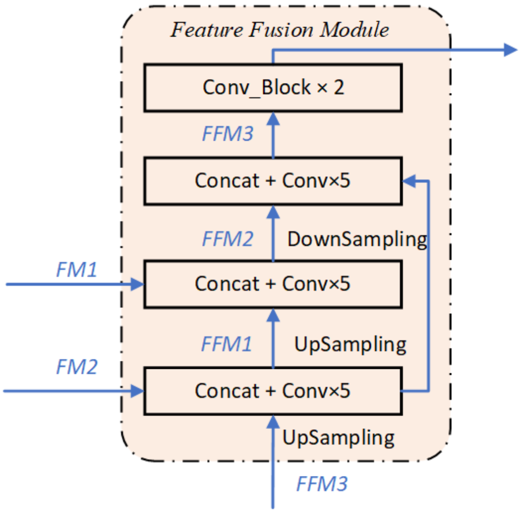


FIGURE 7. The network architecture of feature fusion module.

transition layer directly. CSPNet [36] has three main advantages: (1) while maintaining the accuracy of CNN network, it reduces the amount of network calculation and improves the learning ability of CNN network; (2) it reduces the calculation bottleneck; (3) the lightweight network structure reduces the memory cost [36]. There are 5 ResBlock_body convolutional blocks in the backbone network CSPDarknet53, and the stacking numbers(N) of residual blocks(Resblock) on the main branches are 1, 2, 8, 8 and 4 respectively.

After the main features are extracted from the main network CSPDarknet53, the text picture also needs to go through the SPP module. The SPP module enables the CNN network to extract feature maps on multiple scales, and while enhancing the ability of the neural network to extract features, it can also be applied to target detection tasks, making the target detection network adapt to image input of any size. In our network structure, after performing three convolutions on the last feature layer of CSPdarknet53, the SPP module uses the maximum pooling of different scales to process it. The maximum pooling core size is 13 × 13, 9 × 9, 5 × 5, 1 × 1. Finally, the different pooling results are merged in the channel dimension and then subjected to three convolutions to complete our text image feature extraction.

2) FEATURE FUSION MODULE

In our proposed text detection network architecture, the feature fusion module applies the idea of PANet [37]structure to get better text area features, which is part of the feature fusion network in YOLOV4, as shown in Figure 7. In the feature extraction module, we extract three effective feature maps and name them respectively FM_1, FM_2, FM_3. These three effective feature maps are respectively the feature map output by the third Resblock_body in the backbone network CSPDarknet, the feature map output by the fourth Resblock_body, and the last feature map output by the feature extraction module. The size of the FM_3 feature map is 1/32 of the input image, so the FM_3 is first upsampled by 2 times, then is concatenated with the FM_2 feature map whose size is 1/16 of the input image on the channel dimension, and then through 5 convolutions For feature fusion, we name the fused feature map FFM_1. In the same way, the feature layer FM_1 is fused with FFM_1 after upsampling twice to obtain FFM_2. FFM_2 and FFM_1 are concatenated on the channel and convolved 5 times to obtain FFM_3. Finally, the fused feature layer FFM_3 is convoluted twice. So far, the work of the feature fusion module has been completed, which is expressed mathematically as follows.

$$\begin{aligned} FFM_1 &= Conv5(Concat(Upsamp_2(FM_3), FM_2)) \\ FFM_2 &= Conv5(Concat(Upsamp_2(FFM_1), FM_1)) \\ FFM_3 &= Conv5(Concat(Downsamp_2(FFM_2), FFM_1)) \end{aligned} \quad (2)$$

In the above formula, $Upsamp2()$ represents the 2 times upsampling operation; $Downsamp2()$ represents the 2 times downsampling operation; $Concat()$ represents the operation of concatenating the two feature maps on the channel dimension; $Conv5()$ represents the feature map Perform five convolution operations.

3) NETWORK PREDICTION MODULE

After feature extraction and feature fusion, our fine-grained text detection network comes to the network prediction module. Inspired by the RPN structure [22], our network prediction module is divided into two branches, namely the prediction text box coordinate regression branch and the prediction text box text score branch. Their size is respectively $(h/16, w/16, 4K)$, $(h/16, w/16, 2K)$, where $(h/16, w/16)$ is inherited from the size of the final output feature map of the feature fusion module; 4 represents the offset between the center coordinates, length and height of the network prediction box and anchor box, which can be expressed as $(\Delta x, \Delta y, \Delta h, \Delta w)$; 2 represents the text and non-text scores of the small prediction box, K represents the number of

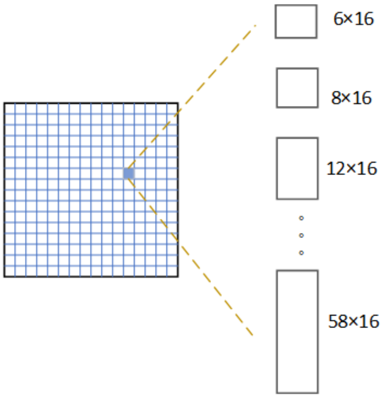


FIGURE 8. The specific anchor mechanism, each feature point on the feature graph corresponds to 10 small-scale anchors with width of 16 and length of 6, 8, 12, 16, 18, 24, 28, 38, 48, 58 respectively.

anchor boxs we set in advance, we will introduce the Anchor mechanism in detail below.

B. SPECIFIC ANCHOR MECHANISM

In our text detection project, the object of text area detection is Chinese medical packaging instructions. From Figure 2, it can be seen that the length distribution of the text instances in the CMPTD dataset is relatively balanced in each interval, but the height distribution of the text instances is not balanced, most of them are distributed in the interval below 30px, which indicates that the length distribution of the text instances in the CMPTD dataset is not clustered in a certain interval, the number in each length interval is not much different and the height of the text line belongs to the same dataset, so the text scale is similar and most of them are distributed in the interval below 30px. Anchor-based general target detection, such as FasterRCNN [22], SSD [23], YOLO [24] etc. Its specific size and specific aspect ratio anchor box mechanism performs well in some target detection tasks with little scale changes, such as face detection, vehicle detection, cat and dog detection, etc., but it is difficult to accurately predict the text line area in such situations. Therefore, inspired by CTPN [18], our anchor box mechanism consists of 10 small-scale anchor boxes with fixed width but not fixed length. In our project, the width of the anchor box is fixed to 16px, and the length set is set to 6, 8, 12, 16, 18, 24, 28, 38, 48, 58 according to the height distribution of text instances in the CMPTD dataset, as shown in Figure 8. This anchor box mechanism can not only regress the bounding box more accurately, but also predict the text line area we need more effectively.

C. MULTITASK LOSS FUNCTION

In the fine-grained text detection network in this article, the network needs to predict the text and non-text scores of each small text box and the location information of each small text box. The location information here includes the center coordinate point, height and width of the text box. Therefore, the multi task loss function is used as the objective function of our training network, which includes text classification loss and text box regression loss, as shown in Eq (3).

$$\begin{aligned} Loss\left(s_i, \delta v_j\right) = & \frac{1}{N_{cls}} \sum_i L_{cls}\left(s_i, s_i^*\right) \\ & + \frac{\lambda}{N_{reg}} \sum_j L_{reg}\left(\delta v_j, \delta v_j^*\right). \end{aligned} \tag{3}$$

In the above formula, because the mini-batch strategy is used to train the text detection network, N_{cls} represents the number of anchors that need to be classified in the mini-batch and i represents the index of the classified anchor in the mini-batch; s_i represents the probability that the network prediction box is a text box, and s_i^* represents whether the anchor is ground truth, and the value is 0 or 1. Similarly, N_{reg} represents the number of anchors that need to be regression prediction in mini-batch. Inspired by CTPN [18], this article only considers the anchors that are judged as positive samples. Here, anchors with ground truth IOU (Intersection over Union) ≥ 0.7 are considered as positive samples, and anchors with $IOU \leq 0.3$ are considered as negative samples. Finally, δv_j represents the position offset between the network prediction box and the j-th positive sample in the mini-batch, δv_j^* represents the position offset between the ground truth and the j-th positive sample in the mini-batch, and λ represents the balance Parameter of multi-task training, which is set to 1 in this paper. L_{cls} represents the classification loss function. Here we use the commonly used two-class softmax loss function. The specific mathematical form is as follows,

$$L_{cls}\left(s_i, s_i^*\right) = -\left[s_i^* \log s_i + \left(1 - s_i^*\right) \log\left(1 - s_i\right)\right]. \tag{4}$$

s_i represents the probability that the network prediction box is a text box, and s_i^* represents whether the anchor is positive sample, and the value is 0 or 1. L_{reg} represents the bounding box regression loss. In CTPN, the Bounding box regression loss uses the smoothed-L1 loss function proposed in Fast RCNN [38]. Smoothed-L1 loss has better robustness than L1 loss and L2 loss. When the difference between the model prediction box and the ground truth is large, the gradient value is limited; when the difference between the model prediction box and the ground truth is small, the gradient value remains small enough. Therefore, in this paper, the bounding box regression loss still uses the smoothed-L1 loss function. Its mathematical form is as follows.

$$L_{reg}\left(\delta v_j, \delta v_j^*\right) = \sum_{j \in \{x,y,h,w\}} smooth_{L_1}\left(\delta v_j - \delta v_j^*\right) \tag{5}$$

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{6}$$

δv_j and δv_j^* represents the output value of the text detection model and the value of bounding box to be regressed, which includes four parts: the abscissa and ordinate of the center of the detection box or the real box, and the corresponding height and width. Among them, the specific mathematical form of $\delta v_j^* (j \in x, y, h, w)$ is as follows

$$\delta v_x^* = \left(c_x^* - c_x^a\right)/w^a \quad \delta v_y^* = \left(c_y^* - c_y^a\right)/h^a \tag{7}$$

$$\delta v_h^* = \log\left(h^*/h^a\right) \quad \delta v_w^* = \log\left(w^*/w^a\right) \tag{8}$$

(c_x^*, c_y^*, h^*, w^*) represents the coordinates of the center point, length and height of the ground truth box; (c_x^a, c_y^a, h^a, w^a) represents the coordinates of center point, length and height of the anchor corresponding to ground truth box; Anchor height has been mentioned in the specific anchor mechanism section above. There are 10 candidate values for anchor height, which is determined according to the distribution of text line length in CMPTD dataset, while h^a represents the height of the anchor that has the highest IOU value with the ground truth box; w^a has only one candidate value, which is set to 16px in this paper. δv_j is the output value of the text detection model, but it also needs to carry out the inverse operation of Eq (7) and Eq (8) to convert it into the center point coordinates, height and width of the prediction box.

D. TEXT LINE CONSTRUCTION

Inspired by CTPN and [39], the text line construction algorithm is shown in algorithm1. The fine-grained text prediction model proposed in this paper infers a series of small-scale prediction boxs pb after feature extraction and fusion of medical packaging instructions. After transformation, the upper left corner coordinates (x_1^i, y_1^i), lower right corner coordinates (x_2^i, y_2^i) and the corresponding text prediction score $score_i$ of each small-scale prediction box are given. First, the prediction text box set is sorted from small to large in the x_1 direction (line 1); Then, the following operation is performed for each prediction box pb^i. In the prediction box set $lookforward(pb^i, gap)$ whose horizontal forward distance from pb^i is less than gap (gap in this paper is set to 35 through experiments), the pb^j whose $Overlap_v(pb^i, pb^j)$ is greater than the threshold th_o (th_o in this paper is set to 0.7) and $Similarity(pb^i, pb^j)$ is greater than the threshold th_s (th_s in this paper is set to 0.7) is selected and added to the candidate set of pb_{cdd_1} (lines 4-8); In the algorithm, $Overlap_v(pb^i, pb^j)$ and $Similarity(pb^i, pb^j)$ functions are defined as follows:

$$overlap_v(pb^i, pb^j) = \frac{\min\left(y_2^i, y_2^j\right) - \max\left(y_1^i, y_1^j\right)}{\min\left(y_2^i - y_1^i, y_2^j - y_1^j\right)} \tag{9}$$

$$Similarity(pb^i, pb^j) = \frac{\min\left(y_2^i - y_1^i, y_2^j - y_1^j\right)}{\max\left(y_2^i - y_1^i, y_2^j - y_1^j\right)}. \tag{10}$$

Finally, the text prediction box pb_k with the largest score value is found in the candidate set (line 9). Similarly, the prediction box score $max_{score}(pb_{cdd_2})$ satisfying the above conditions is obtained by horizontal reverse searching for pb^k. If the prediction box pb^i score is greater than $max_{score}(pb_{cdd_2})$, then the prediction boxs pb^i and pb^k is the longest connection, and $Graph(i, k) = True$ is set (lines 10-17). After traversing the prediction box set pb, a series of small-scale prediction boxs belonging to the same line of text area are connected into a line, and the final algorithm outputs the text line proposal set tlp (lines 19-21).

Algorithm 1 Text Line Construction

```
Input: Predicted box set:
    $pb\{pb^1, pb^2, \cdots, pb^n\}; pb^i = (x_1^i, y_1^i, x_2^i, y_2^i, score^i)$.
Output: Output: Text line proposal set: $tlp$.
1: $pb \leftarrow Sort_{x_1}(pb)$;
2: for $pb^i$ in $pb$ do do
3:   $pb_{cdd_1} \leftarrow [\,]$, $pb_{cdd_2} \leftarrow [\,]$;
4:   for $pb^j$ in $lookforward(pb^i, gap)$ do do
5:     if $Overlap_v(pb^i, pb^j) \geq th_o$
           $\& \; Similarity(pb^i, pb^j) \geq th_s$ then
6:       $pb_{cdd_1} \leftarrow pb_{cdd_1} \cup pb^j$;
7:     end if
8:   end for
9:   $pb^k \leftarrow argmax_{score}(pb_{cdd_1})$;
10:  for $pb^l$ in $lookbackward(pb^k, gap)$ do do
11:    if $Overlap_v(pb^k, pb^l) \geq th_o$
           $\& \; Similarity(pb^k, pb^l) \geq th_s$ then
12:      $pb_{cdd_2} \leftarrow pb_{cdd_2} \cup pb^l$;
13:    end if
14:  end for
15:  if $pb^i[score] \geq max_{score}(pb_{cdd_2})$ then
16:    $Graph(i, k) = True$;
17:  end if
18: end for
19: $tlp\_group \leftarrow graph\_connect(Graph)$;
20: $tlp \leftarrow tlp\_connect(tlp\_group)$;
21: return $tlp$;
```

V. EXPERIMENTS AND ANALYSIS

A. EXPERIMENTAL DETAILS

Since Cspdarknet53 in YOLOv4 is used as the backbone network of the fine-grained text line region detection model in this paper, we use the model trained on COCO data set [40] as the pre-training model of our backbone network, and then use our own CMPTD data set to train the model. In order to enhance the robustness of our text detection model, we adopt a multi-scale training strategy. When the length width ratio of the input image remains unchanged, the long side size is set to (512,576,640,736,800) in turn, and the corresponding batch size is set to (10,8,6,4,4) in turn. Each size is trained with 50 epochs. For example, the first input image size of the model is 512. After 50 epochs data sets training, the input size of the model is increased to 576, and the training continues. The model trains 360 epochs in total, and the long side size of the image is increased to 800. The network optimizes the model by SGD and momentum method, where the momentum is set to 0.99 and the weight decay is set to 5×10^{-4}. The initial learning rate of the model is 0.001, and the cosine annealing strategy [41] is used to adjust the learning rate, which helps the model to jump out of the local minimum and reach the global optimal solution. In the i-th run, the learning

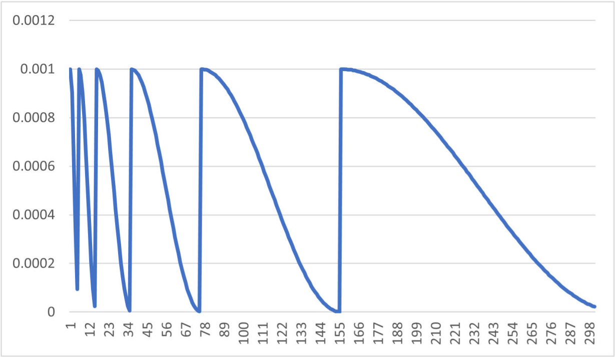


FIGURE 9. The learning rate change curve of using the cosine annealing decay strategy, the learning rate repeatedly decreases and rises in the early and mid-term of training. At the end of training, the learning rate no longer rises and gradually decreases.

rate decay formula is shown in Eq (11).

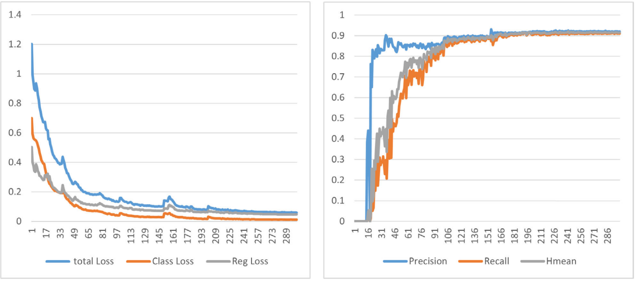
$$\eta_t = \eta_{\min}^i + \frac{1}{2}\left(\eta_{\max}^i - \eta_{\min}^i\right)\left(1 + \cos\frac{T_{\text{cur}}}{T_i}\pi\right). \tag{11}$$

$\eta_{\min}^i$ and $\eta_{\max}^i$ are the minimum and maximum learning rates respectively; T_{cur} represents the number of epochs that have been iterated since the last restart, and T_i is a fixed value, usually half of the cosine period. The learning rate curve of this paper is shown in Figure 9. In addition, in order to enhance the generalization ability of the model, some common data enhancement operations are used in this paper, such as randomly changing the brightness, contrast and saturation of the input image. All the experiments are based on the PyTorch deep learning framework. Our model runs on the Ubuntu 20.04 system and two GTX2080Ti graphics cards.

B. EVALUATION METHOD

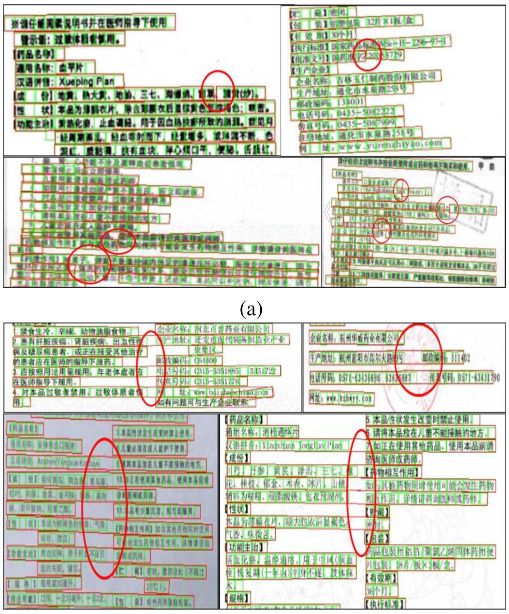
In order to objectively evaluate the text line detection effect of our model on CMPTD dataset, we use the evaluation method for text detection based on DetEval [42]. DetEval, as a criterion of ICDAR2013 competition, is often used by researchers to evaluate the text detection performance of the model on ICDAR2013 dataset. This evaluation method has three indexes, including Precision, Recall, and Hmean. Precision reflects the correct ratio of the model text prediction, while Recall reflects the proportion of the real text area in the dataset correctly predicted. As for Hmean, it is a comprehensive index of Precision and Recall. They can effectively evaluate the matching relationship between the model detection boxs and the ground truth boxs, including one-to-one, one-to-many, and many-to-one matching. The mathematical form of these three indicators is as follows:

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+FN} \\ Hmean &= 2 \times \frac{Pecision \times Recall}{Precision + Recall}. \end{aligned} \tag{12}$$



(a) (b)

FIGURE 10. (a) The declining curve of the loss function when we train the model, including regression loss, classification loss, and total loss; (b) The change curve of the training model on the test set, including accuracy, recall, hmean.



(a)

(b)

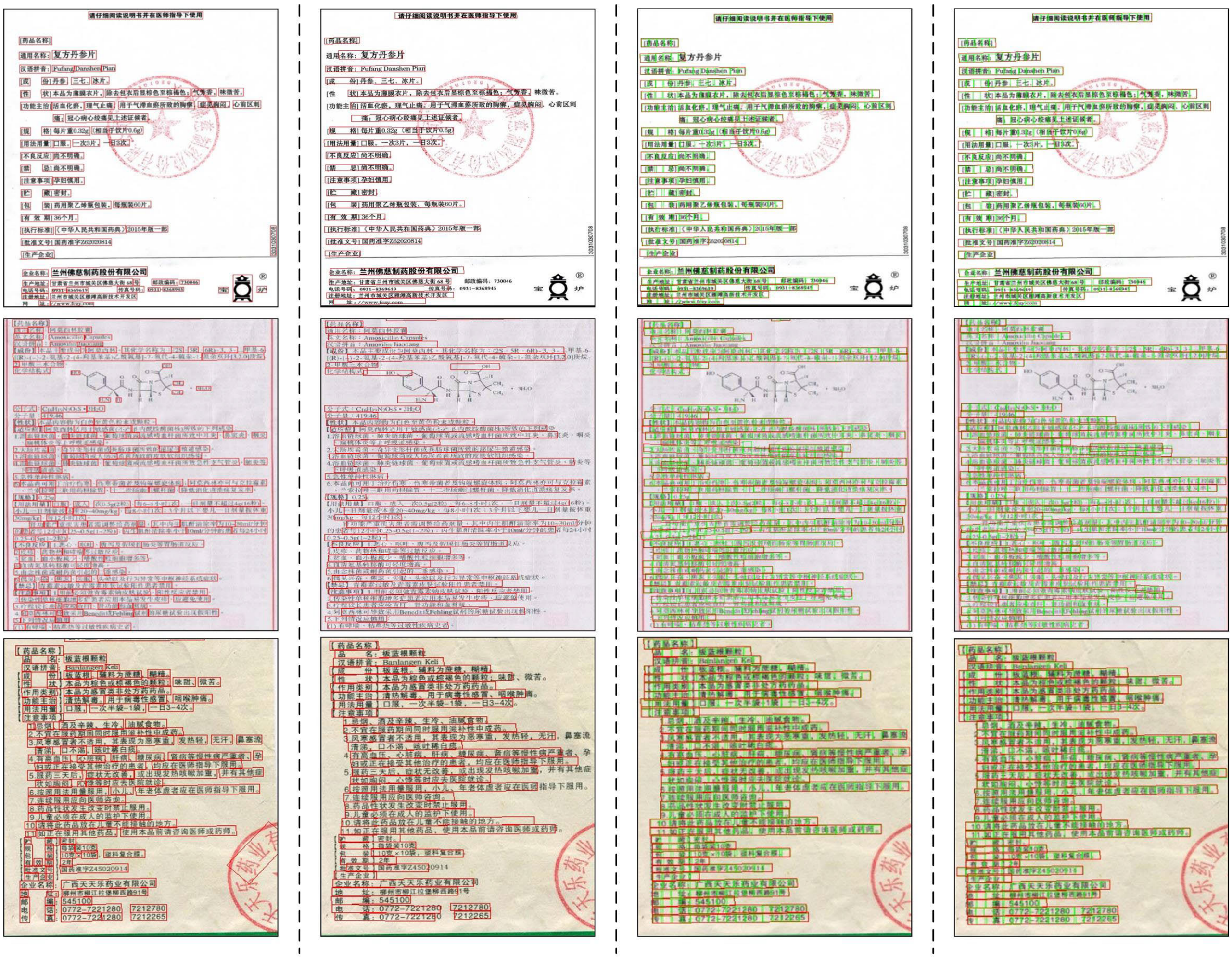
FIGURE 11. (a) The parameter *gap* is set too small, and the text line is broken; (b) The parameter *gap* is set too large, and the prediction boxes belonging to two text lines are assigned to the same text line.

where, TP, FP and FN are the number of correct detection, error detection and missing detection respectively. In this paper, precision, recall and hmean are used to evaluate the detection effect of our fine-grained text detection model on CMPTD test set.

C. EXPERIMENTAL RESULTS AND ANALYSIS

1) MODULE EXPERIMENT

Inspired by CTPN and YOLOv4, we do experimental research on SPP module, feature fusion module(FF) and BGRU module. In order to find out which combination of these three modules has the most positive impact on fine-grained text detection network, this section conducts addition and subtraction experiments on these three modules on CMPTD dataset. In order to ensure the effectiveness of the experiment, the experimental conditions are basically the same except for the three module variables. The experimental results are shown in Table 1.



(a) CRAFT (b) EAST (c) CTPN (d) Our Method

FIGURE 12. Experimental results of CRAFT, EAST, CTPN and our method on CMPTD dataset.

TABLE 1. Addition and subtraction combination experiment of three modules: SSP module, the FF(Feature fusion) module and BGRU module. "P" means Precision, "R" means Recall, "H" means Hmean.

our model	SSP	BGRU	FF	P(%)	R(%)	H(%)
baseline				91.90	91.37	91.63
			✓	91.90	91.25	91.50
		✓		92.29	90.71	91.49
	✓		✓	**92.46**	91.74	**92.10**
		✓	✓	91.70	90.71	91.20
	✓	✓	✓	91.45	**92.07**	91.76

The benchmark model in Table 1 means that we directly use Cspdarknet53 as the feature extraction module of the model, and then directly connect with the network prediction module. At this time, the hmean value of the model reaches 91.63%, and the effect is not bad. This may be due to the fact that the number of images in the CMPTD dataset is not very large. Secondly, in the experiment in Table 1, it can be found that when the SSP module and the FF module are used in combination, the accuracy and hmean value of the model are the highest, respectively 92.46% and 92.10%, which indicates that the text detection model has the best accuracy and comprehensive performance in text region prediction. Therefore, the combination of SSP module and FF module is selected as a part of the fine-grained text detection model proposed in this paper. On this basis, we continue to increase the BGRU module(The BGRU module is added after the fusion feature layer FFM3 in the manner of adding BLSTM in CTPN). At this time, the recall rate of the model is the highest, reaching 92.07%, but the precision rate and hmean value are reduced. This tells us that the text detection model at this time can predict the most text regions, but the accuracy and overall performance of the text region prediction are decreased.

In the section of text line construction algorithm, the super parameter *gap* represents the maximum horizontal distance that the prediction box is looking for. Theoretically, the larger the gap is set, the larger the spacing of small prediction boxes in the same text line can be. In other words, for the small

TABLE 2. Experimental data of superparametric *gap*.

Gap	P(%)	R(%)	H(%)
20	88.64	91.18	89.89
35	92.46	**91.74**	**92.10**
50	**92.51**	91.55	92.03
65	92.33	91.18	91.75

prediction box set with larger spacing, the model tends to attribute it to the same text line. When the *gap* is set too small, the prediction box belonging to the same text line will be divided into different text lines, and the text detection of the model will break, As shown in Figure 11(a); When the *gap* is set too large, the prediction box that belongs to two text lines will be assigned to the same text line, which will lead to a longer text line and more white space, which is not conducive to the subsequent work of text recognition. As shown in Figure 11(b); Table 2 shows the detection effect of the fine-grained text detection model in the case of different super parameter horizontal *gap* values in the CMPTD dataset. It can be found that when the horizontal gap is set to 35, the effect of the model is the best.

2) COMPARISON EXPERIMENT

In this section, in order to objectively evaluate the text line detection ability of our proposed fine-grained text detection model on the CMPTD dataset, we compare it with the current text detection model (CRAFT [15], EAST [14], CTPN [18]) that performs well in text detection tasks on the CMPTD dataset. CRAFT belongs to segmentation based text detection algorithm, but it is not pixel level segmentation. It detects a single character and the connection relationship between characters and then connects characters according to the connection relationship, finally forms a text line. CRAFT has strong generalization ability, and can process text in any direction, even curve text and distorted text. However, the detection effect in CMPTD dataset is not ideal, as shown in Figure 12(a). For characters with large interval, it is easy to divide them into two separate text lines, and there are some false detection, in other words, it performs well in recall, but the precision of text detection results is not ideal, which is not conducive to the subsequent text recognition work. EAST is a two-stage text detection method and its detection shape is a rotating rectangle and a quadrilateral, which can detect both words and text lines. However, limited by the receptive field, the detection of both ends of long text is often inaccurate, that is, its precision of text detection on the CMPTD dataset is poor, as shown in Figure 12(b); CTPN is improved from Faster R-CNN [22], a detection algorithm that performs well in target detection tasks. CTPN combines convolutional neural network (CNN) and bidirectional long-term and short-term memory network (BLSTM) and has good detection effect for horizontal text and the detection effect for other directions is general. For most of the CMPTD data sets with horizontal text lines, it has good detection effect, but there are still some missed detections, as shown in Figure 12(c). Our method not only can better extract text features, but also can accurately and completely detect text regions in any direction by setting a specific anchor mechanism and using an improved text line construction algorithm, In other words, it has achieved excellent performance in terms of precision and recall, as shown in Figure 12(d).

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a fine-grained text detection network for the text detection task of Chinese packaging instructions and built a dataset CMPTD to train and test the model. On the basis of YOLOv4 and CTPN, through the specific anchor mechanism and using the improved text line construction algorithm, in the experiment of module addition and subtraction combination, our model achieves 92.46% precision and 91.74% recall on the CMPTD dataset, and the effect is excellent. In the future, in order to structurally extract the text information from medical photocopies, we will continue to study the text recognition and information extraction of Chinese medical photocopies after the improvement of text region detection. Finally, we will form a complete information extraction pipeline of Chinese medical photocopies, which will be conducive to the construction of medical digital platform in developing countries.

REFERENCES

[1] W. Xue, Q. Li, and Q. Xue, "Text detection and recognition for images of medical laboratory reports with a deep learning approach," *IEEE Access*, vol. 8, pp. 407–416, 2020, doi: 10.1109/ACCESS.2019.2961964.

[2] C. Rossignoli, A. Zardini, and P. Benetollo, "The process of digitalisation in radiology as a lever for organisational change: The case of the academic integrated hospital of verona," in *DSS 2.0—Supporting Decision Making With New Technologies*, vol. 261. Amsterdam, The Netherlands: IOS Press, 2014, pp. 24–35.

[3] S. Bonomi, "The electronic health record: A comparison of some European countries," in *Information and Communication Technologies in Organizations and Society*. Cham, Switzerland: Springer, Jan. 2016, pp. 33–50.

[4] M. B. Buntin, M. F. Burke, M. C. Hoaglin, and D. Blumenthal, "The benefits of health information technology: A review of the recent literature shows predominantly positive results," *Health Affairs*, vol. 30, no. 3, pp. 464–471, Mar. 2011.

[5] A. K. Jha, C. M. DesRoches, P. D. Kralovec, and M. S. Joshi, "A progressreport on electronic health records in US hospitals," *Health Affairs*, vol. 29, no. 10, pp. 1951–1957, 2010.

[6] T. Shu, H. Liu, F. R. Goss, W. Yang, L. Zhou, D. W. Bates, and M. Liang, "EHR adoption across China's tertiary hospitals: A cross-sectional observational study," *Int. J. Med. Informat.*, vol. 83, no. 2, pp. 113–121, Feb. 2014.

[7] R.-G. Zhou, S. Chang, and Y. Li, "A neural network architecture for information extraction in Chinese drug package insert," *IEEE Access*, vol. 8, pp. 51256–51264, 2020, doi: 10.1109/ACCESS.2020.2978079.

[8] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," *Frontiers Comput. Sci.*, vol. 10, no. 1, pp. 19–36, Feb. 2016, doi: 10.1007/s11704-015-4488-0.

[9] S. Karaoglu, R. Tao, T. Gevers, and A. W. M. Smeulders, "Words matter: Scene text for image classification and retrieval," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1063–1076, May 2017, doi: 10.1109/tmm.2016.2638622.

[10] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1480–1500, Jul. 2015.

[11] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 161–184, 2020.

[12] X. Chen, L. Jin, Y. Zhu, C. Luo, and T. Wang, ''Text recognition in the wild: A survey,'' 2020, *arXiv:2005.03492*. [Online]. Available: http://arxiv.org/abs/2005.03492

[13] B. Shi, X. Bai, and C. Yao, ''An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017.

[14] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, ''EAST: An efficient and accurate scene text detector,'' in *Proc. CVPR*, Jul. 2017, pp. 2642–2651.

[15] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, ''Character region awareness for text detection,'' in *Proc. CVPR*, Jun. 2019, pp. 9365–9374.

[16] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, ''ICDAR 2015 competition on robust reading,'' in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1156–1160.

[17] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, ''YOLOv4: Optimal speed and accuracy of object detection,'' 2020, *arXiv:2004.10934*. [Online]. Available: http://arxiv.org/abs/2004.10934

[18] Z. Tian, W. Huang, T. HePan, and H. Qiao, ''Detecting text in natural image with connectionist text proposal network,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 56–72.

[19] J. Yan, J. Li, and X. Gao, ''Chinese text location under complex background using Gabor filter and SVM,'' *Neurocomputing*, vol. 74, no. 17, pp. 2998–3008, Oct. 2011, doi: 10.1016/j.neucom.2011.04.031.

[20] L. Neumann and J. Matas, ''A method for text localization and recognition in real-world images,'' in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 770–783.

[21] W. Huang, Z. Lin, J. Yang, and J. Wang, ''Text localization in natural images using stroke feature transform and text covariance descriptors,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1241–1248.

[22] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and C. A. Berg, ''SSD: Single shot multibox detector,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ''You only look once: Unified, real-time object detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[25] K. He, G. Gkioxari, P. Dollár, and R. Girshick, ''Mask R-CNN,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[26] J. Long, E. Shelhamer, and T. Darrell, ''Fully convolutional networks for semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[27] O. Ronneberger, P. Fischer, and T. Brox, ''U-net: Convolutional networks for biomedical image segmentation,'' 2015, *arXiv:1505.04597*. [Online]. Available: http://arxiv.org/abs/1505.04597

[28] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, ''TextBoxes: A fast text detector with a single deep neural network,'' in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 416–4161.

[29] M. Liao, B. Shi, and X. Bai, ''TextBoxes++: A single-shot oriented scene text detector,'' *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018.

[30] D. Deng, H. Liu, X. Li, and D. Cai, ''PixelLink: Detecting scene text via instance segmentation,'' in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 6773–6780.

[31] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, ''TextSnake: A flexible representation for detecting text of arbitrary shapes,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 20–36.

[32] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, ''Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 67–83.

[33] Y. Tang and X. Wu, ''Scene text detection using superpixel-based stroke feature transform and deep learning based region classification,'' *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2276–2288, Sep. 2018.

[34] Y. Liu, S. Zhang, L. Jin, L. Xie, Y. Wu, and Z. Wang, ''Omnidirectional scene text detection with sequential-free box discretization,'' 2019, *arXiv:1906.02371*. [Online]. Available: https://arxiv.org/abs/1906.02371

[35] J. Redmon and A. Farhadi, ''YOLOv3: An incremental improvement,'' 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[36] C.-Y. Wang, H.-Y. Mark Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, ''CSPNet: A new backbone that can enhance learning capability of CNN,'' 2019, *arXiv:1911.11929*. [Online]. Available: http://arxiv.org/abs/1911.11929

[37] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, ''Path aggregation network for instance segmentation,'' 2018, *arXiv:1803.01534*. [Online]. Available: http://arxiv.org/abs/1803.01534

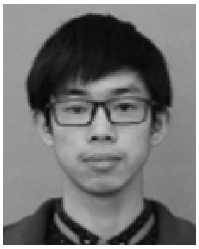
[38] R. Girshick, ''Fast R-CNN,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, CA, USA, Dec. 2015, pp. 1440–1448.

[39] L. Cao, H. Li, R. Xie, and J. Zhu, ''A text detection algorithm for image of student exercises based on CTPN and enhanced YOLOv3,'' *IEEE Access*, vol. 8, pp. 176924–176934, 2020, doi: 10.1109/ACCESS.2020.3025221.

[40] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, ''Microsoft COCO: Common objects in context,'' 2014, *arXiv:1405.0312*. [Online]. Available: https://arxiv.org/abs/1405.0312

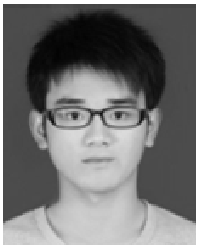
[41] I. Loshchilov and F. Hutter, ''Decoupled weight decay regularization,'' 2017, *arXiv:1711.05101*. [Online]. Available: http://arxiv.org/abs/1711.05101

[42] C. Wolf and J.-M. Jolion, ''Object count/area graphs for the evaluation of object detection and segmentation algorithms,'' *Int. J. Document Anal. Recognit.*, vol. 8, no. 4, pp. 280–296, 2006, doi: 10.1007/s10032-006-0014-0.



HAIWEN WU was born in 1997. He is currently pursuing the M.Sc. degree with Shanghai Maritime University. His research interests include computer vision, text detection, and text recognition.

RI-GUI ZHOU (Member, IEEE) was born in March 1973. He received the B.S. degree from Shandong University, China, the M.S. degree from the Department of Computer Science and Technology, Nanchang Hangkong University, China, in 2003, and the Ph.D. degree from the Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronauts, China, in 2007. From 2008 to 2010, he was a Postdoctoral Fellow with Tsinghua University, China. From 2010 to 2011, he held a postdoctoral position at Carleton University, Ottawa, ON, Canada. From 2014 to 2015, he was a Visiting Scholar with North Carolina State University, Raleigh, NC, USA. He is currently a Professor with the College of Information Engineering, Shanghai Maritime University, China. His main research interests include quantum image processing, quantum reversible logic, and quantum genetic algorithm. He is also a Senior Member of the China Computer Federation (CCF). He was a recipient of the New Century Excellent Talents Program, Ministry of Education of China, in 2013.



YAOCHONG LI received the B.S. degree from Henan Polytechnic University, China, in 2017. He is currently pursuing the joint master's and Ph.D. degree with Shanghai Maritime University. His research interests include deep learning, bioinformatic analysis, and quantum neural networks.

• • •