ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20231023001052
Original scientific paper

A Text Recognition Algorithm Based on a Dual-Attention Mechanism in Complex Driving Environment

Ling DING, Liyuan WANG*, Yuanfang WANG, Shaohuai YU, Jinsheng XIAO

Abstract: In response to many problems such as complex background of text recognition environment, perspective distortion, shallow handwriting, and mixed Chinese and English characters, we have designed an OCR algorithm framework with features such as landmark extraction and correction, image enhancement, text detection, and text recognition. We have designed a DBNet based on dual attention mechanism and content-aware upsampling. We have also designed a text recognition module incorporating the central loss CRNN + CTC to improve content awareness. Experimental results show that the improved text detection network in this paper has increased accuracy by 5.09%, recall by 2.12%, and *F*-score by 3.46% on the ICDAR2015 dataset. The text recognition network has improved the accuracy of recognizing Chinese and English characters by 1.2%.

Keywords: double attention mechanism; landmark extraction; text detection; text recognition

1 INTRODUCTION

Optical character recognition has various applications, and one of the most important applications is text recognition in natural scenarios. With the rise of autonomous driving technology and the development of the Internet of Everything, the use of OCR technology to replace traditional algorithms for recognition can not only reduce the error rate of information identification but also improve the safety of autonomous driving.

In recent years, researchers have replaced traditional methods of computer vision with deep learning methods in object detection and semantic segmentation, and achieved remarkable results. Liao [1] proposes an improved Textbox++ network for vertical, tilted text detection, increasing decimals in the preselected box aspect ratio and changing the convolution kernel to 3 × 5 to better accommodate the detection of both vertical and tilted text. J. S. Xiao et al. [2] propose an Angle optimization algorithm for seal text detection and recognition, which first stretches the seal through the polar coordinate transformation, and then detects it. Tang et al. [3], for the text bending and dense problem, proposed a module to represent attraction and repulsion between text blocks, and to design the instance-aware loss function so that Seglink++ networks can be trained end-to-end. Wang et al. [4] designed the lightweight FPEM (Feature Pyramid Enhancement Module) and FFM(feature fusion module) to improve the performance of text detection, and adopted a faster pixel clustering method in the post-processing to improve the speed of text detection. Liao et al. [5] found that the segmentation-based algorithm is more time-consuming in the threshold binarization, proposed the network DBNet that can learn the segmentation threshold, and cleverly designed a binary function, which not only improves the detection accuracy, but also greatly improves the detection speed. Zhu and other scholars [6] innovatively expressed the curved text contour in terms of the Fourier transform parameters, while the mathematically superior Fourier coefficient can fit any curve, combined with the designed FCENet network, to improve the accuracy of text detection. Although many word recognition algorithms have obtained relatively accurate detection results, there are still many problems to be solved. For example, the text distribution positions of landmark text lines, character gap differences, resulting in the problem of text detection box adhesion, font handwriting is changeable, the landmark characters including English, punctuation marks and other 6000 characters, plus the differences of various printing fonts, to bring difficulties to text detection [7].

In view of the above problems, this paper designs a text detection network based on dual attention and content-perception upsampling. The dual attention mechanism is used to improve the feature extraction and selection ability of the network. The content-perception upsampling module can increase the sensory field and improve the content perception ability. Combine CRNN (Convolutional Recurrent Neural Network) and CTC, decoding the loss of text recognition by connecting sequence features to CTC, solve the problem of CRNN inconsistent between set time length and real text length, increase the central loss and increase the feature spacing between characters to avoid the problem of "false detection".

2 RELATED WORK

Our work revolves around text detection networks based on dual-attention and content-aware upsampling. Double attention module is added to the feature extraction network of DBNet (Differentiable Binarization Network) text detection network: spatial attention and channel attention, the former adjusts the weight of the key regional information of the feature map, and the latter adjusts the weight of different channels of the feature map. Considering the disadvantages of small nuclear receptive field and no content perception, the upsampling based on content perception is introduced to improve the feature fusion module of DBNet [8]. The DBNet text detection network takes the deformable convolution ResNet network as the backbone network, and the variability convolution improves the feature extraction ability of the text with different shape change and irregular size. A dual attention module (Convolutional Block Attention Module, CBAM), which consists of channel attention and spatial attention. Channel attention adaptively adjusts the weights of

different channels in the feature map, improves the ability to select important features, filters out or weakens the features of interference; spatial attention adaptively adjusts the weight of different positions in the feature map, improves the identification of the text area in the scene, and suppresses the background area of the scene text.

In order to avoid the predicted text obtained after getting the CRNN being greater than the actual text label length, and the CTC (Connectionist Temporal Classification) layer is used to complete the text transcript, and to reduce the identification error due to mixed characters in English, the loss module in the text recognition network adds the central loss function, increase the spacing of character feature distribution, and reduce the character recognition error [9]. The overall improved document OCR algorithm includes document extraction, document correction, DB(Differentiable Binarization) text detection network principle and CRNN text recognition network principle. This paper will also focus on the Retinex image brightness enhancement with color recovery, ACE(Automatic Color Enhancement) based image color enhancement, DBNet based on dual-attention and upsampling of content perception, Integrated Convolutional Recurrent Neural Network and Connectionist Temporal Classification text recognition network with center loss.

3 ALGORITHM DESIGN

The roadmap OCR algorithm framework is shown in Fig. 1, which mainly includes three parts: preprocessing, text detection and text recognition. The text detection module adopts the text detection algorithm based on DBNet network to accurately intercept the text line area in the roadmap and retain the text line detection coordinates [10]. The text recognition module adopts a CRNN-based text recognition algorithm to identify the content of the text line images. Simultaneously increase the Retinex image brightness enhancement with color recovery, ACE-based image color enhancement, DBNet based on dual attention and content-perception upsampling, and a CRNN + CTC text recognition network with integrated center loss [11].

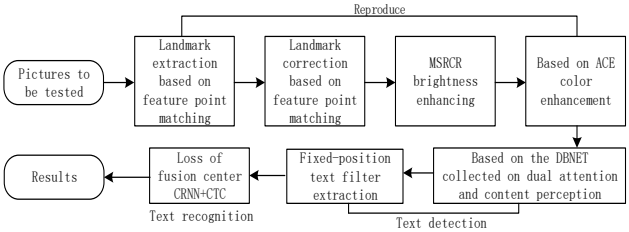


Figure 1 Roadmap OCR algorithm framework

3.1 A DBNET Text Detection Network Based on Dual-Attention and Content-Aware Upsampling

Unlike traditional segmentation networks that binarize the probability graph using a set threshold graph. The DBNet network can generate a threshold graph (Threshold Map). To achieve an adaptive segmentation, generate an approximate binary plot (Approximate Binary Map), text detection results are more accurate, together with a deformable convolution (Deformable Convolution Network, DCN) [12]. The proposed algorithm adds the offset offsets to the convolution kernel of the original standard convolution. For the convolutional networks to adaptively adjust the receptive field, effectively improve the detection ability of DBNet network to detect text lines [13]. Let the input feature graph X, standard convolution of the central points p_0:

$$y(p_0) = \sum_{p_n \in R} W(p_n) \cdot X(p_0 + p_n) \tag{1}$$

In the Eq. (1): R = {(−1, −1), (−1, 0), …, (0, 1), (1, 1)}, the offset is added to the deformable convolution unit Δp_n with the following formula:

$$y(p_0) = \sum_{p_n \in R} W(p_n) \cdot X(p_0 + p_n + \Delta p_n) \tag{2}$$

In Fig. 2, before the input feature graph is used in normal convolution, the offset field with $2N$ channels is calculated by a standard convolution unit, which represents the offset in the x-axis and the y-axis of each pixel of the convolution field of vision, respectively. After the standard convolution kernel adds this offset, the size and position of the convolution kernel can be adaptively adjusted according to the content of the input feature map, so as to better adapt to the irregular and deformed text area [14].

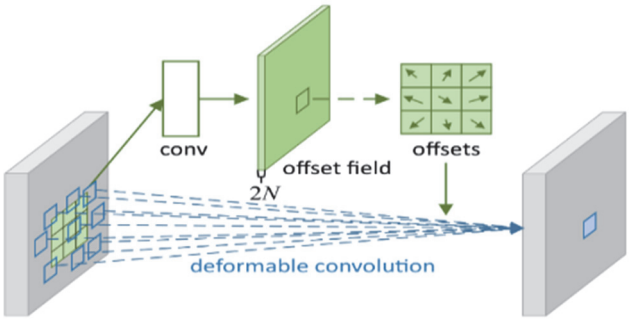


Figure 2 Schematic diagram of the deformable convolution

The feature extraction of the DBNet network adopts the ResNet residual network, and the dual attention module (Convolutional Block Attention Module, CBAM) is added to each residue block of the ResNet, which consists of channel attention (Channel Attention) and spatial attention (Spatial Attention), as shown in Fig. 3. Channel attention adaptively adjusts the weight of different channels of the feature map, improves the selection ability of the important features, filters out or weakens the features of the interference; spatial attention adaptation adjusts the weight of different positions of the feature map, improves the recognition of the text area in the scene, and suppresses the background area of the scene text [15].

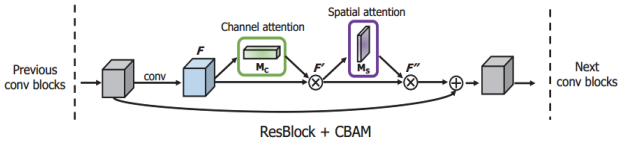


Figure 3 Residual blocks incorporating the dual attention path

Specifically, the channel attention takes the account of the importance of different channels in the feature graph, so serving as coefficients of different channels in the context of the feature graph F by generating vectors of M_c size of $1 \times 1 \times C$. However, the spatial attention considers the importance of the different positions of the feature map, and assigns different coefficients to the different spatial positions of the feature map F' by generating a matrix of sizes $H \times W \times 1$. The following describes the design of the channel attention and spatial attention modules, respectively.

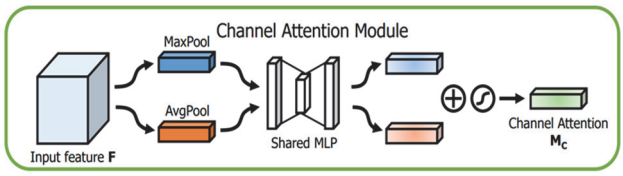


Figure 4 Schematic diagram of the channel attention module

The input feature graph is recorded as F, whose size is $H \times W \times 1$, and the maximum pooling and the average pooling on the spatial dimensions respectively obtain two vectors of size as $1 \times 1 \times C$, where the maximum pooling extracts the detail features, while the average poolin F' g extracts the background features [16]. The two vector features share a multi-layer perceptron (MLP), which add the two output results pixel by pixel and pass through the sigmoid activation layer to obtain the channel attention M_c of the size $1 \times 1 \times C$, as shown in Fig. 4. Finally, the channel coefficient M_c and feature diagram F to multiply the channel purification feature map (Channel-Refined Feature) F'.

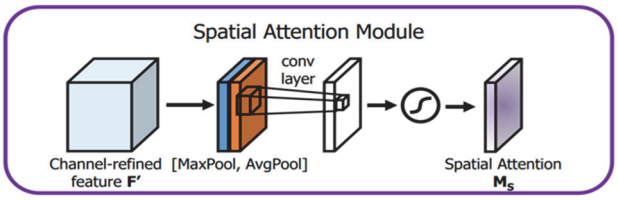


Figure 5 Schematic diagram of the spatial attention module

The channel purification feature map F' as input, unlike channel attention module, spatial attention module is maximum pooling and average pooling on the channel dimension, so that the output size is $H \times W \times 1$, the pooling results together, and then through the convolution layer and sigmoid activation layer, the size $H \times W \times 1$ for spatial attention M_s, as shown in Fig. 5, and as the coefficient, and the channel purification feature map F' multiplied pixel by pixel, get the final purification feature map F''.

The standard convolution of the ResNet part of the residual network was also replaced with the variability convolution shown in Fig. 2 to accommodate the feature extraction of both shape irregularities and font-variable text [17]. In addition, when the FPN (Feature Pyramid Network) network integrates the feature maps of different scales generated by ResNet, Content-Aware Re Assembly of Features (CARFE) is used to fuse the feature maps of different scales, and the operator structure is shown in Fig. 6.

CARFE mainly consists of two parts: the upsampling core prediction module (Kernal Predication Module) and the content perception recombination module (Content-Aware Reassembly Module), the former predicts the upsampling core based on the sample content feature information, and the latter reorganize the content according to the predicted upsampling core. Let the input feature graph be χ, and the size is $H \times W \times C$, assuming that the upsampling multiplier is recorded as σ, then the output upsampling result is χ', and the size is recorded as $\sigma H \times \sigma W \times C$.

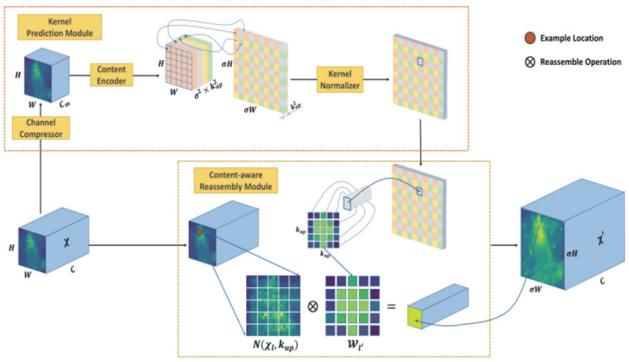


Figure 6 Upsampling module structure based on content sensing

Upsampling kernel prediction module: The input feature graph channel is first compressed to C_m first through the channel compression module (channel compressor) composed of the convolution of 1×1 to reduce the subsequent computation [18]. Assuming that the upsampling kernel size is $k_{up} \times k_{up}$, there is a corresponding upsampling kernel for any element $l' \in \chi'$ in the output result, so the predicted upsampling kernel size should be $\sigma H \times \sigma W \times k_{up} \times k_{up}$. Therefore, $\sigma^2 k_{up}$ convolutional kernels of size $k_{encoder} \times k_{encoder}$ are used for content coding (content encoder), with the output size $H \times W \times \sigma^2 k_{up}$, and expanded in the channel dimension, so that the size is $\sigma H \times \sigma W \times k_{up} \times k_{up}$ upsampled convolution kernel, and finally after kernel normalization processing (kernel normalizer), so that the sum of the elements of the upsampled kernel is equal to 1 [19].

Content perception reorganization module: For any element χ_l in the input feature map, the pixel $N(\chi_l, k_{up})$ of the $k_{up} \times k_{up}$ receptive field range at the center is removed, and the element l' of the corresponding position is removed from the upsampling core, and expanded in the channel dimension to obtain the $k_{up} \times k_{up}$ size upcore $W_{l'}$ at the point, and the upsampling value of the position is obtained through the point product [20].

Thus, the input feature map of different locations of the pixel corresponding upsampling core is different, the upper sampling core is determined by the input feature graph content, do the content of the "perception", in addition to the sampling feeling field size is $k_{up} \times k_{up}$, compared to the nearest neighbor upsampling or double linear interpolation sampling, do the larger feeling field, effectively make up for the defects of the original sampling [21].

3.2 CRNN+CTC

Text recognition network is mainly divided into Seq2Seq + Attention and CRNN + CTC, the former limited Seq2Seq serial mechanism, resulting in unsatisfactory long text recognition, and Attention mechanism brings huge

additional parameters to the network, while the latter in short text recognition, and thanks to CTC forward-backward recurrence, can maintain low computational complexity, without bringing additional parameters to the network. However, CRNN + CTC is only applicable to the text lines of one-dimensional shape rules. For deformation and irregular text, the recognition effect is poor, but considering that the text lines in the roadmap are regular text lines, so this paper adopts CRNN + CTC. This paper, based on CRNN + CTC, tries to add the center loss to further improve the character recognition accuracy [22].

As can be seen from the above, CTC loss in CRNN + CTC text recognition network first finds the probability of each predicted character in the sequence through softmax, and then solves the text sequence alignment problem through CTC transcription. Therefore, in essence, text recognition is the classification problem of character images, and character classification error is determined by character characteristics [23]. Therefore, the central loss function (Center Loss) for increasing the spacing of the sample feature distribution is added. The principle is as follows.

For a fully connected classification network, assuming the input feature vector $x_i \in R^d$, the network matrix parameter is $W \in R^{d\times n}$, the prediction class is y_i, and the total class is m, the softmax classification loss is as follows:

$$L_S = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{3}$$

$$L_S = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{4}$$

The Center Loss function is defined as follows:

$$L_c = \frac{1}{2}\sum_{i=1}^{m} \left\| x_i - c_{y_i} \right\|^2 \tag{5}$$

In the equation, C_{y_i} represents the center of the feature distribution of the y_i category, x_i represents the features of the input into the fully connected classification layer, and m represents the number of samples per batch in the training [24]. According to the above equation, Center Loss is to calculate the distance between the sample feature and the feature center. We hope that the distance between the feature of the input sample in a batch and the feature center of the class, the smaller the better. Center Loss L_c and Feature Center C_{y_i} update updated as follows:

$$\frac{\partial L_c}{\partial x_i} = x_i - c_{y_i} \tag{6}$$

$$\Delta c_j = \frac{\sum_{i=1}^{m} \delta(y_i = j)\cdot(c_j - x_i)}{1 + \sum_{i=1}^{m} \delta(y_i = j)} \tag{7}$$

Where $\delta(yi = j)$ means 1 when the classification category is j; otherwise, 0. It follows that the feature center C_j of the class is only updated when the predicted category y_i is equal to the real label j. Add Center Loss to the classification loss function with a certain weight λ, as follows:

$$L_{\text{total}} = L_s + \lambda L_c \tag{8}$$

The literature takes the classification of handwritten digit recognition as an example to show the changes of the feature distribution of handwritten digit in different λ situations, as shown in Fig. 7.

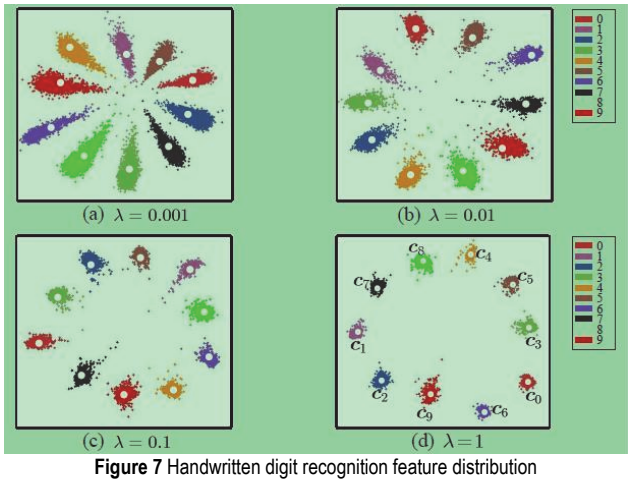


Figure 7 Handwritten digit recognition feature distribution

According to the Fig. 7, with the λ increasing number, the gap between class and class increases more and more, and the feature distribution boundary of each class becomes more and more clear, so that the classification error can be reduced.

For the landmark text identification, the text is mixed with English characters, numbers and other characters, a total of more than 6000 categories, there are a large number of English characters, so compared with the pure English text identification, the character classification difficulty in this article is greatly improved. In addition, the characteristics of the printed font and the amount of noise, which are more likely to have similar characters, such as the letters "l" and the number "1". This article tries to add the center loss to the CTC loss function and set λ to 0.1. It is hoped to increase the gap between the character feature distribution and reduce the situation of character misrecognition, so as to improve the accuracy of text recognition.

4 EXPERIMENTAL RESULTS AND THE ANALYSIS

The extracted background of the landmark is simpler than that of a natural scene, and the background type of the landmark image is relatively fixed. In order to better

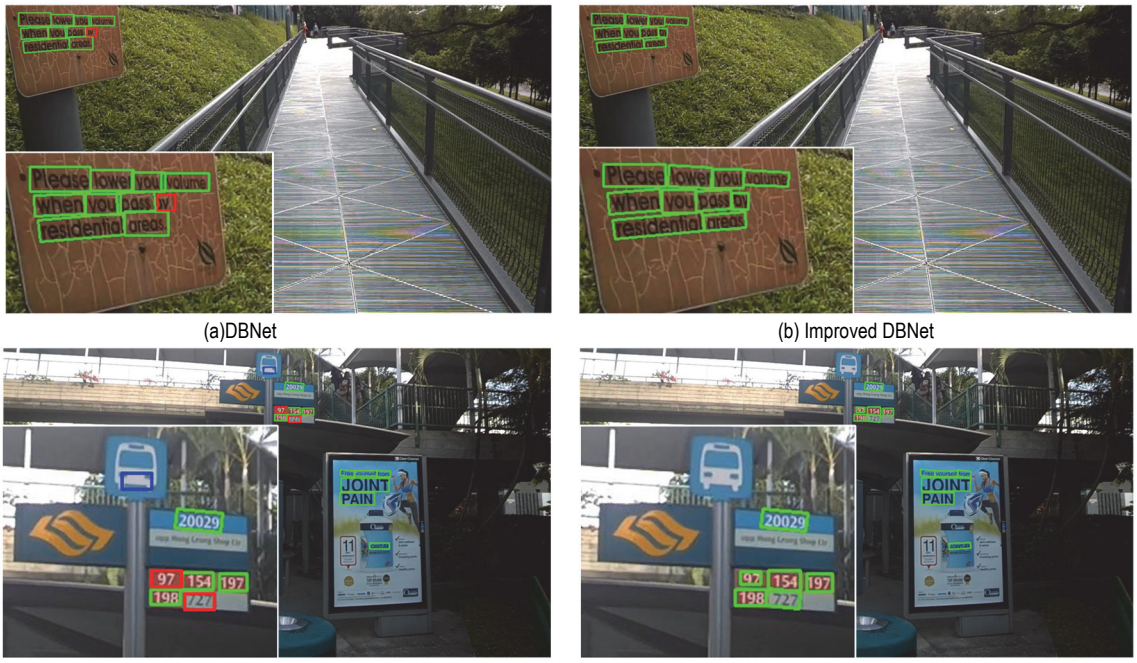
measure the performance of the scene text detection network, a more complex ICDAR2015 scene text detection data set is used for comparison experiments [25].

The backbone networks used variability convolution ResNet50 with an initial learning rate of 0.007 and decreased exponentially with the iterations iter. The formula is $\left(1-\frac{\text{iter}}{\max \text{iter}}\right)^{\text{power}}$. A total of 1200 epoch were trained, and to prevent overfitting, an l2 regularization was applied, with the weight set to 0.001. None of the experimental models completed the pre-training with the huge synthetic text detection dataset, and the remaining network parameters remained consistent, and the experimental results are shown in Tab. 1 [26].

Table 1 Text test Results

Detection algorithm	Precision	Recall	*F*-score
DBNet	86.73%	76.79%	81.46%
Integrating the dual-attention mechanism of DBnet	88.69%	75.47%	81.54%
Content-based upsampling DBNet	88.17%	77.19%	82.31%
The improved DBNet of this paper	91.82%	78.91%	84.88%

In this paper, we first experiment with the original DBNet network and the improved DBNet, while for the comparative analysis of the optimization measures, the DBNet integrated into the dual-attention mechanism and the DBNet based on content-perception upsampling are trained separately. From Tab. 1, the text detection accuracy by 5.09%, recall by 2.12% and *F* score by 3.42%. The DBNet shows that the dual attention mechanism improves the detection accuracy, and the DBNet comparison improves the detection accuracy and recall rate based on the content perception.



(a)DBNet (b) Improved DBNet

c) DBNet (d) Improved DBNet

Figure 8 Comparison of the text test results

According to the comparison of Fig. 8, compared with the improved DBNet, the improved DBNet can also well detect smaller text areas, and the "omission" (red box) is significantly reduced, thus improving the recall rate of text detection. Compared with (*c*) and (*d*), the original DBNet misdetected the logo plate (blue box) into the text area, and the improved "misdetection" situation of DBNet was also improved to a certain extent, thus improving the accuracy of text detection [27].

The text recognition dataset used in this experiment consists of 2000 landmark text line images, 200000 OCR document text line images and the ICDAR2013 text recognition dataset, including 80% for the training set and 20% for the test set. The CRNN + CTC text recognition network was trained, the backbone network adopts ResNet34, the RNN sequence length *T* is set to 25, the input text image size is 32032, the optimizer uses Adam, beta1 is set to 0.9, and beta2 is set to 0.999. The learning rate uses the cosine annealing algorithm, the initial value learning rate is set to 0.001, uses the L2 regularization, and the weight is set to [28]. As a contrast, this experiment also uses the pure English text dataset ICDAR2013 for training and testing. The text recognition results are shown in Tab. 2.

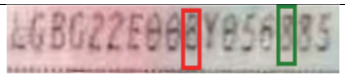
Table 2 Scene text recognition experiment results

Recognition algorithm	This paper presents the Sino-English datasets	ICDAR2013
CRNN+CTC	91.2%	89.6%
Seq2Seq+Attention	89.8%	89.9%
CRNN+CTC+C_Loss	92.4%	90.2%

According to the experimental results, CRNN + CTC and Seq2Seq + Attention recognition networks are used respectively. For short text data sets and ICDAR2013, Seq2Seq + Attention recognition effect is slightly better, but CRNN and CRNN + CTC have more advantages, which verifies the unsatisfactory interpretation of Seq2Seq + Attention for long text recognition. At the same time, on the basis of the original CTC loss, Center Loss was added to improve the recognition accuracy of English

character data sets by 1.2%, and the ICDAR2013 English recognition dataset was also improved by 0.6%. The test cases are shown in Tab. 3.

Table 3 A sample of the text identification results

Test Case	CRNN+CTC	CRNN+CTC+Center_Loss
	LGBG22E008Y056885	LGBG22E00BY056885
	JOINT	J0INT
	198727	19B727

For CRNN + CTC, there will occasionally be character identification error, such as the fuzzy "B" identified "8", "O" identified "0", because some characters are too similar, coupled with fuzzy, image noise and other factors, it is easy to misidentification. However, Center Loss can increase the distance between the character feature distributions and reduce the classification error due to similar character features, thus improving the character recognition accuracy.

5 CONCLUSION

In this paper, we study the scene text detection algorithm, and adopt and improve the DBNet text detection network with both performance and speed. The FPN network module (CBAM) is added to the DBNet attention module (CBAM to improve the ability of the network to select important features of the text region in the scene and inhibit the interference caused by other background regions. The upsampling operator (CARAFE) with content perception is adopted to overcome the disadvantages of the traditional sampled small receptive field and no content perception, and apply it to the feature graph fusion of different scales to further improve the text detection performance of the network. The scene text recognition algorithm CRNN + CTC is also improved to add the Center Loss to increase the character feature spacing, to reduce the misidentification caused by the character appearance in English. At the same time, a large number of landmark images in complex scenes were collected and made into a scene text data set and a scene text recognition data set. Combined with the public dataset to train the improved DBNet network and the improved CRNN + CTC network, the experimental results show that the detection and recognition performance of the network has improved ideal.

Acknowledgment

The authors would like to thank CCCC Second Highway Consultants Co.,Ltd for helpful discussions on topics related to this work. This work is supported by the Technology Research and Development Projects of China Communications Construction Group Corporation (Project No: 2019-ZJKJ-ZDZX02), Hubei Provincial Scientific research Platform project of Hubei Second Normal University(Project No: ESRC20220043) and the Guiding project of scientific research plan of Hubei Provincial Department of Education(Project No: B2021261).

6 REFERENCES

[1] Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M. H. (2022). Restormer: efficient transformer for high-resolution image restoration. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5728-5739. https://doi.org/10.1109/CVPR52688.2022.00564

[2] Xiao, J. S., Zhao, T., Xiong, W. X., Yang, T., & Yao, W. Q. (2021). Seal text detection and recognition algorithm with angle optimization network. *Journal of Electronics & Information Technology*, *43*(11), 3327-3334.

[3] Tang, J., Yang, Z. B., Wang, Y. P., Zheng, Q., Xu, Y. C., & Bai, X. (2019). Seglink plus plus: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping. *Pattern Recognition*, *96*, 106954. https://doi.org/10.1016/j.patcog.2019.06.020

[4] Tanmay, J., Palaiahnakote, S., Pal, U., & Liu, C. L. (2021). Deformable scene text detection using harmonic features and modified pixel aggregation network. *Pattern Recognition Letters*, *152*, 135-142. https://doi.org/10.1016/j.patrec.2021.10.006

[5] Chopra, A., Sharma, D. K., Jha, A., & Ghosh, U. (2023). A Framework for Online Hate Speech Detection on Code-mixed Hindi-English Text and Hindi Text in Devanagari Chopra. *ACM Transactions on Asian and Low Resource Language Information Processing*, *22*(5). https://doi.org/10.1145/3568673

[6] Liao, M., Wan, Z., Yao, C., Chen, K., & Bai, X. (2020). Real-time scene text detection with differentiable binarization. *Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI*, *34*(7), 11474-11481. https://doi.org/10.1609/aaai.v34i07.6812

[7] Zhu, Y. Q., Chen, J. Y., Liang, L. Y., Kuang, Z. H., Jin, L. W., & Wayne, Z. (2021). Fourier contour embedding for arbitrary-shaped text detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE*, 3123-3131. https://doi.org/10.1109/CVPR46437.2021.00314

[8] Wan, Q., Ji, H., & Shen, L. (2021). Self-attention based text knowledge mining for text detection. *Computer Vision and Pattern Recognition IEEE*, 5979-5988. https://doi.org/10.1109/CVPR46437.2021.00592

[9] Shivakumara, P., Banerjee, A., Pal, U., Nandanwar, L., Lu, T., & Liu, C. L. (2023). A New Language-Independent Deep CNN for Scene Text Detection and Style Transfer in Social Media Images. *IEEE Transactions on Image Processing*, *32*, 3552-3566. https://doi.org/10.1109/tip.2023.3287038

[10] Liu, Z., Lin, G., Yang, S., Feng, J., Lin, W., & Wang, L. G. (2018). Learning Markov clustering networks for scene text detection. *Proc. CVPR*, 6936-6944. https://doi.org/10.1109/CVPR.2018.00725

[11] Liu, Y., Jin, L., Zhang, S., Luo, C., & Zhang, S. (2019). Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition*, *90*, 337-345. https://doi.org/10.1016/J.PATCOG.2019.02.002

[12] Liao, M., Lyu, P., He, M., Yao, C., Wu, W., & Xiang, B. (2019). Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. *IEEE Trans. Pattern Anal*. https://doi.org/10.1109/TPAMI.2019.2937086

[13] Liu, Z., Lin, G., Yang, S., Liu, F., Lin, W., & Wang, L. G. (2019). Towards robust curve text detection with conditional spatial expansion. *Proc. CVPR*, 7269-7278. https://doi.org/10.1109/CVPR.2019.00744

[14] Xu, Y., Wang, Y., Zhou, W., Wang, Y., Yang, Z., & Xiang, B. (2019). Textfield: Learning a deep direction field for irregular scene text detection. *IEEE Trans. Image Processing*, *28*(11), 5566-5579. https://doi.org/10.1109/TIP.2019.2900589

[15] Xie, E., Zang, Y., Shao, S., Yu, G., Yao, C., & Guangyao, L. (2019). Scene text detection with supervised pyramid context network. *Proc. AAAI*, *33*, 9038-9045. https://doi.org/10.1609/aaai.v33i01.33019038

[16] Tian, Z., Shu, M., Lyu, P., Li, R., Zhou, C., Xiaoyong, S., & Jiaya, J. (2019). Learning shape-aware embedding for scene text detection. *Proc. CVPR*, 4234-4243. https://doi.org/10.1109/CVPR.2019.00436

[17] Huang, M., Liu, L., Peng, Z., Liu, C., Lin, D., Zhu, S., Yuan, N., Ding, K., & Jin, L. (2022). SwinTextSpotter: Scene text spotting via better synergy between text detection and text recognition. *CVPR*, 2022. https://doi.org/10.1109/CVPR52688.2022.00455

[18] Ch'ng, C. K., Chan, C. S., & Liu, C. L. (2020). Total-text: toward orientation robustness in scene text detection. *International Journal on Document Analysis and Recognition (IJDAR)*, *23*(1), 31-52. https://doi.org/10.1007/s10032-019-00334-z

[19] Alshammari, M., Al-Smadi, M., Arqub, O. A., Hashim, I., & Alias, M. A. (2020). Residual series representation algorithm for solving fuzzy duffing oscillator equations. *Symmetry*, *12*(4). https://doi.org/10.3390/sym12040572

[20] Arqub, O. A. (2017). Adaptation of reproducing kernel algorithm for solving fuzzy Fredholm-Volterra integrodifferential equations. *Neural Comput & Applic*, *28*, 1591-1610. https://doi.org/10.1007/s00521-015-2110-x

[21] Fang, S. C., Xie, H. T., Wang, Y. X., Mao, Z. D., & Zhang, Y. D. (2021). Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7098-7107. https://doi.org/10.1109/CVPR46437.2021.00702

[22] Liao, M., Lyu, P., He, M., Yao, C., Wu, W., & Bai, X. (2021). Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(2), 532-548. https://doi.org/10.1109/TPAMI.2019.2937086

[23] Liu, Y. L., Chen, H., Shen, C. H., He, T., Jin, L. W., & Liangwei, W. (2020). Abcnet: Real-time scene text spotting with adaptive bezier-curve network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9809-9818. https://doi.org/10.1109/CVPR42600.2020.00983

[24] Ze, L., Yutong, L.,Yue, C., Han, H., Yixuan, W., Zheng, Z., Stephen, L., & Baining, G. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/ICCV48922.2021.00986

[25] Nguyen, N., Thu, N., Vinh, T., Minh-Triet, T., Thanh Duc, N., Thien Huu, N., & Minh, H. (2021). Dictionary-guided scene text recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR46437.2021.00730

[26] Qiao, L., Chen, Y., Cheng, Z., Xu, Y., Niu, Y., Pu, S., & Wu, F. (2021). Mango: A mask attention guided one-stage scene text spotter. *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, 2467-2476. https://doi.org/10.1609/aaai.v35i3.16348

[27] Zhang, C. S., Tao, Y. F., Du, K., Ding, W. P., Wang, B., Liu, J., & Wang, W. (2021). Character-level street view text spotting based on deep multi-segmentation network for smarter autonomous driving. *IEEE Transactions on Artificial Intelligence*, 1-1. https://doi.org/10.1109/TAI.2021.3116216

[28] Arqub, O. A., Singh, J., Maayah, B., & Alhodaly, m. (2021). Reproducing kernel approach for numerical solutions of fuzzy fractional initial value problems under the Mittag-Leffler kernel differential operator. *Role of Fractional Operators and Mathematical Modelling in Applied Sciences*, *46*(7), 7965-7986. https://doi.org/10.1002/mma.7305

[29] Arqub, O. A., Singh, J., & Alhodaly, M. (2021). Adaptation of kernel functions-based approach with Atangana-Baleanu-Caputo distributed order derivative for solutions of fuzzy fractional Volterra and Fredholm integrodifferential equations. *Role of Fractional Operators and Mathematical Modelling in Applied Sciences*, *46*(7), 7807-7834. https://doi.org/10.1002/mma.7228

Contact information:

Ling DING
School of Computer Science,
Hubei University of Education,
Wuhan, China

Liyuan WANG
(Corresponding author)
CCCC Second Highway Consultants Co., Ltd,
Wuhan, China
E-mail: wangliyuan_wh@163.com

Yuanfang WANG
School of Electronic Information,
Wuhan University,
Wuhan, China

Shaohuai YU
CCCC Second Highway Consultants Co., Ltd,
Wuhan, China

Jinsheng XIAO
School of Electronic Information,
Wuhan University,
Wuhan, China