A Handwritten Text Detection Model Based on Cascade Feature Fusion Network Improved by FCOS

Ruiqi Feng[1], Fujia Zhao[1], Shanxiong Chen[1](✉), Shixue Zhang[2], and Dingwang Wang[1]

[1] Southwest University, Chongqing, China
csxpml@163.com
[2] Guizhou University of Engineering Science, Bijie, Guizhou, China

Abstract. In this paper, we propose a method for detecting handwritten ancient texts. The challenges in detecting this type of data are: the complexity of the layout of handwritten ancient texts, the varying text sizes, mixed arrangement of pictures and texts, the high number of hand-drawn patterns and the high background noise. Unlike general scene text detection tasks (ICDAR, TotalText, etc.), the texts in the images of ancient books are more densely distributed. For the features of the dataset, we propose a detection model based on cascade feature fusion called DFCOS, which aims to improve the fusion of localization information in lower layers. Specifically, bottom-up paths are created to use more localization signals from low-levels, and we incorporate skip connections to better extract information in the backbone, and then improve our model by parallel cascading. We verified the effectiveness of our DFCOS on HWAD (Handwritten Ancient-Books Dataset), a dataset containing four languages - Yi, Chinese, Tibetan and Tangut - provided by the Institute of Yi of Guizhou University of Engineering Science and National Digital Library of China, and its precision, recall and F-measure outperformed most of the existing text detection models.

Keywords: Scene text detection · Handwritten text detection · FCOS · Ancient books

1 Introduction

Scene text detection, which refers to locate the position of text regions in an image, has attracted a lot of attention in the field of computer vision in recent years as the first step in scene text reading. With the rapid development of deep learning and convolutional neural networks, researchers have proposed an increasing number of excellent frameworks for scene text detection. Most of these CNN-based methods are built on top of successful generic object detection frameworks or semantic segmentation technologies. However, at this stage, scene text detection remains challenging even with the support of deep neural

E. H. Barney Smith and U. Pal (Eds.): ICDAR 2021 Workshops, LNCS 12917, pp. 51–65, 2021.
https://doi.org/10.1007/978-3-030-86159-9_4

networks due to various factors such as diverse text fonts and styles, variable text sizes and complex image backgrounds. However, at this stage, scene text detection remains challenging even with the support of deep neural networks, due to various distracting factors such as diverse text fonts and styles, variable text sizes and complex image backgrounds.

As an important carrier of Chinese culture, handwritten ancient books are of great historical value. The digitization of handwritten ancient text images is vital for the preservation of cultural heritage. However, the existing text detection methods do not perform well on handwritten ancient text documents for the following reasons: 1) The layout structure of handwritten ancient texts is complex, which means it is common to find mixed texts and illustrations, arbitrary text sizes, various text arrangements, a large number of hand-painted patterns and noisy background; 2) The text is densely distributed and the overall text target is smaller than the scene text; 3) The backgrounds of different kinds of handwritten ancient texts and writing styles differ. Therefore, we consider proposing a new text detection method for the characteristics of ancient texts.

Anchor-free object detection methods have emerged in recent years. Compared with traditional anchor-based methods, this detection method is more suitable for our research, because the ancient text objects have greater variability in aspect ratio and orientation compared to the objects in object detection tasks, which can be very demanding for the design of anchors and can increase a huge amount of workload, causing the decrease in the overall efficiency of the detection task. Another defect of anchor-based methods is that it does not handle multi-directional and curved texts well. The paper Fully Convolutional One-Stage Object Detection (FCOS) [1] proposed a one-stage object detection algorithm based on FCN [2] and FPN [3], which is an excellent representative of the anchor-free method. Its biggest advantage is that the accuracy of detection is maintained while eliminating anchors. However, the FPN structure used in the FCOS model has an obvious drawback, i.e. it has only top-down paths with horizontal information fusion (add operation), which makes the network obtain mainly information from the high-levels and under-utilize the information from lower layers. This defect makes it less effective on ancient books with small and dense text objects. To overcome these problems, we propose a novel feature fusion network, which we design to integrate into FCOS's framework to improve the detection of textual objects in our ancient books.

The main contributions of this paper are as follows: 1) We constructed a Handwritten Ancient-Books Dataset (HWAD) containing 8k images in four languages [4], which laid the data foundation for the subsequent research on digitization of handwritten ancient books; 2) For the case of smaller and denser objects of ancient books, we are inspired by FPN and propose a bottom-up information fusion path for its tendency to miss low-level localization information, and at the same time, we add an additional skip connection path between the backbone and the bottom-up outputs; 3) In order to more fully reuse the obtained feature information, we form a feature cascade fusion network by overlaying multiple bottom-up and skip-connected structures in a parallel cascade manner; 4) Since the fixed sampling points in standard convolutional layers will restrict the receptive field to a fixed position, which is not conducive to the detection of ancient texts, we introduce a deformable convolution [5] for adaptive sampling; 5) After an in-depth study of the advantages and disadvantages of the FCOS, we introduce Gaussian weighted Soft-NMS [6] instead of NMS in the post-processing part, and replace the loss function in regression branch with a more suitable CIoU [7]. Experiments have proved that our DFCOS performs well on HWAD, surpassing most existing text detection methods.

2 Related Work

2.1 Scene Text Detection

In recent years, scene text detection methods can be roughly divided into two categories: regression-based methods and segmentation-based methods.

Regression-based methods are often improved from commonly used object detection frameworks, such as Faster R-CNN [8] and SSD [9]. TextBoxes [10] modified the scale of the anchors and the convolution kernels for text detection on the basis of SSD. EAST [11] used FCN to directly predict the score map, rotation angle and b-box for each pixel. LOMO [12] proposed an iterative optimization module and introduced an instance-level shape expression module to solve the problem of detecting scene texts of arbitrary shapes.

Regression-based methods often achieve certain results, but are highly dependent on anchor sizes which are manually set, and many regression-based text detectors are designed for specific text detection scenarios, resulting in low robustness.

According to segmentation-based methods, each pixel in the original image is segmented into text or non-text to determine the approximate text region. It has become the mainstream method for detecting multi-directional and arbitrary-shaped texts. PixelLink [13] predicted the connections between pixels and localized text regions by separating links belonging to different text instances. PSENet [14] adopted progressive scale expansion network using ground-truths to generate a series of masks of different sizes, ultimately improving the detection of irregular text.

Although PSENet works well, the post-processing process is quite time-consuming, leading to slow inference, which is also a common problem of segmentation-based detectors. To address this phenomenon, Liao et al. [15] proposed DBNet, which incorporates the binarization process into training and removes it during inference, resulting in a speedup in model inference.

2.2 Text Detection of Ancient Books

Over the past few years, a number of studies have been conducted on the detection of texts in ancient Chinese and minority languages. Su et al. [16] first binarized the Mongolian ancient texts by OTSU, then used the vertical projection information to locate the text columns, and finally got the individual Mongolian characters by analyzing connected components. However, the datasets involved in the study are neatly arranged, with the images are less polluted and noisy. Yang et al. [17] proposed a single-character detection framework for Chinese Buddhist scriptures using recognition results to guide detection. Shi et al. [18] detected and segmented oracle bones written on bone fragments by a connected component-based approach. Han et al. [19] presented a binarization algorithm based on the combination of Lab color space channels as well as local processing, starting from the different colors of Tibetan ancient documents.

The detection of ancient texts in China started relatively late, and most of the current studies are completed under the condition of standardized printed characters or well-arranged layout with less noise, which does not fit the characteristics of most extant ancient texts. As a result, the models obtained do not have good generalization. Our images are of different styles of handwritten texts with noise due to age and poor preservation, etc. A study on this dataset would be more meaningful.

3 Methodology

In Fig. 1 shows the architecture of our proposed network. It is based on FCOS, but to better avoid the problem of gradient disappearance and gradient explosion caused by too many layers, Dense Network (DesNet) [20] was chosen to build our backbone, i.e. each layer is connected to all previous layers by concatenate. This structure of DesNet is also capable of multiplexing the low-level information while ensuring the complete transmission of them. We propose a new feature fusion network that is independent of the feature extraction process, at the same time, we improve the feature extraction method, the loss function of the regression part and the post-processing algorithm, while keeping the other parts of the FCOS framework. Below, we will separately introduce the specific improvement methods of each part.

3.1 Improved Multi-scale Feature Fusion Network

FPN, compared with using a single feature map for prediction, makes use of the features of neural networks at each stage and can handle the multi-scale variation in object detection (i.e. the fusion of low-resolution feature maps with strong semantic information and high-resolution feature maps with weak semantic information but rich spatial information) with a small increase in computational effort. This is why FCOS chose FPN as their feature fusion network. However, as the targets in our research are smaller and more densely arranged than those in object detection and scene text detection, we need to focus more on the low-level spatial localization signals, while the top-down fusion of FPN is rich in information but many detailed features will be lost after layer-by-layer pooling, which makes the fused feature maps focus more on the abstract information from higher layers and less on low-level spatial information, which is not exactly suitable for our images.

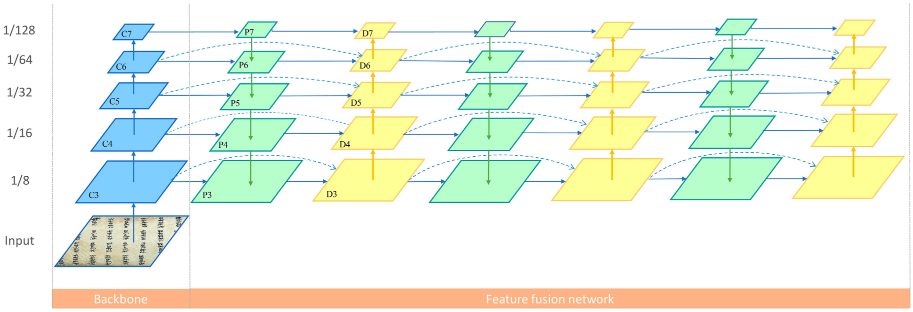


Fig. 1. Architecture of our model, we omit the structure of the subsequent forecast part because we adopted the same structure as FCOS. {C3, C4, C5, C6, C7} represent the features extracted using CNN, {P3, P4, P5, P6, P7} represent all levels of Feature maps generated by FPN, among which P6 and P7 are directly upsampled by P5; {D3, D4, D5, D6, D7} represent the feature maps generated by the bottom-up and skip connection structure.

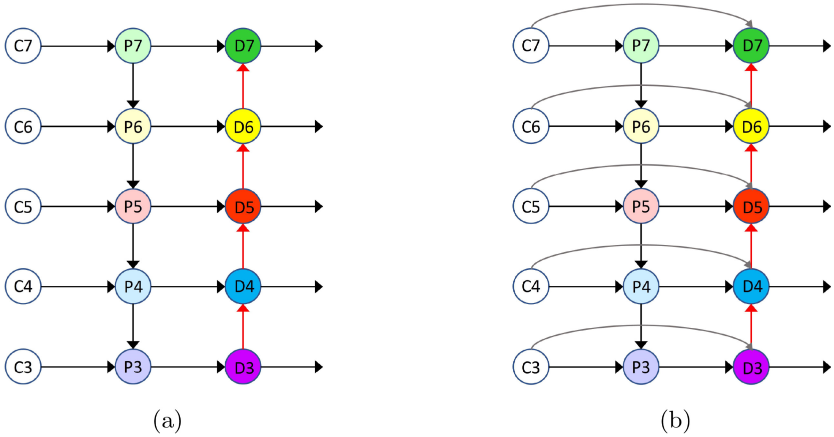


Fig. 2. (a) Bottom-up feature fusion method; (b) Feature fusion structure with skip connection.

Bottom-Up Paths. The top-down structure pays more attention to the abstract information of higher layers, while the bottom-up network contains more low-level localization information. Thus, we consider combining the two to improve the performance of the model. As shown in Fig. 2a, bottom-up feature fusion paths are created to incorporate more localization signals from lower layers. Similar to FPN, the method of feature combination still uses add operation with less computation, that is, to keeping the image dimension unchanged and adding elements correspondingly. Similar to FPN, the size of feature map of each layer among the bottom-up paths corresponds to the same layer of the previous level. The following shows how to calculate the bottom-up fusion:

$$\begin{cases} D_3 = Conv\left(P_3\right) \\ D_4 = Conv\left(P_4 + Resize\left(D_3\right)\right) \\ \cdots \\ D_7 = Conv\left(P_7 + Resize\left(D_6\right)\right) \end{cases} \tag{1}$$

where *Resize* is an upsampling operation, which is used to make the size of Feature maps consistent; *Conv* represents a convolution operation for feature processing.

Skip Connection Paths. We connect the inputs and outputs in a skip connection manner in both the top-down and bottom-up paths, as shown in Fig. 2b, which has the advantage of fusing more features without introducing additional parameters. Just as the original intention of ResNet [21], skip connection can solve the problem of gradient disappearance and gradient explosion, while helping with backpropagation and speeding up the training process. By transferring the feature maps of convolution layer to the bottom-up stage, we can get more details from images and improve the final detection accuracy. Since the information in the backbone is the most important during the entire feature fusion process, we hope to combine the information in the backbone with the top-down features. The concrete calculation process of combining bottom-up and skip connection structure is shown in (2):

$$\begin{cases} D_3 = Conv\left(P_3 + C_3\right) \\ D_4 = Conv\left(P_4 + Resize\left(D_3\right) + C_4\right) \\ \cdots \\ D_7 = Conv\left(P_7 + Resize\left(D_6\right) + C_7\right) \end{cases} \tag{2}$$

Cascade Feature Fusion Structure. With the addition of the bottom-up and skip connection structures, there is a visible improvement in the performance of the model, but to more fully reuse the feature information, we introduce the idea of cascade (i.e. multiple identical structures are connected in series, and the output of the previous part is used as the input of the next stage). Compared with the single-layer feature fusion network, it can make full use of the information in backbone. However, it is worth mentioning that with the deepening of cascade layers, it will inevitably lead to more computation and a higher model complexity. How to balance the model performance and complexity is also an issue of great concern (see the experiments in the following parts of this paper).

3.2 Deformable Convolution

There are many types of handwritten ancient books with complex and variable text sizes, resulting in different aspect ratios of the texts. Since the geometrical structure of standard convolution (CNNs) is fixed (that is, the convolutional unit samples a fixed position of the input feature map), considering that different kinds and sizes of texts require different sizes of perceptual fields, modulated deformable convolutions are applied in all the convolutional layers to extract features, which enables the convolutional layers to adapt to the input images, thus enhancing the overall network's ability to model geometric transformations.

3.3 Other Improvements

GIoU [22] was used as the loss function for the regression part in FCOS. Through our research, we found that GIoU tends to make a larger intersection area between the bounding boxes and ground-truth boxes by increasing the size of the anchor boxes during the regression process, which leads to a slow decrease in the average loss and more iterations for training. In this paper, we introduce CIoU as an alternative, which not only converges faster, but also takes into account the aspect ratio of the texts. We retain the loss functions for other parts of FCOS, so that the total loss function can be expressed as (3):

$$\begin{gathered} L\left(\left\{P_{x,y}\right\},\left\{t_{x,y}\right\}\right)=\frac{1}{N_{pos}}\sum\nolimits_{x,y}L_{cls}\left(P_{x,y},c_{x,y}^{*}\right) \\ +\frac{\lambda}{N_{pos}}\sum\nolimits_{x,y}1\left\{c_{x,y}>0\right\}L_{re.g.}\left(t_{x,y},t_{x,y}^{*}\right) \\ +BCE(Centerness) \end{gathered} \tag{3}$$

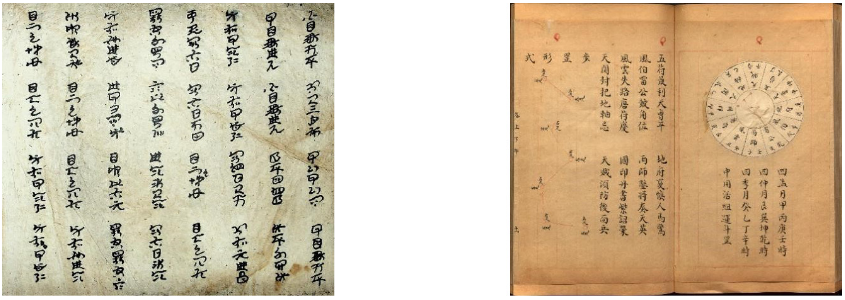
$$L_{reg}=Centerness*L_{CIoU} \tag{4}$$

Equation (3) comes from FCOS, and the part we change (i.e. L_{cls}, represents the regression loss) is shown in (4).

Due to the "one-size-fits-all approach" of NMS algorithm according to scores, the texts tend to be overcut (a single character is divided into multiple characters) and undercut (the bounding box fails to cover the entire character) during prediction process, thus Soft-NMS with Gaussian weighting is chosen to optimize the post-processing module.

4 Dataset

The dataset contains the scanned pictures of ancient books provided by Research Institute of Yi Nationality Studies of Guizhou University of Engineering Science and National Digital Library of China, including the four languages of Yi, Chinese, Tibetan and Tangut. After manual organization and screening, 600 Yi scriptures, 500 Chinese scriptures, 300 Tangut scriptures and 200 Tibetan scriptures were obtained. These images were divided into simple layout (as shown in Fig. 3a) and complex layout (as shown in Fig. 3b) according to whether the layout structure was neat. Among them, 1,250 (78%) of the antiquities were in simple layout and 350 (22%) were in complex layout. Referring to ICDAR2015 [23], we annotated them in the following way: starting from the top left corner, annotate four points clockwise to obtain the result.



(a)

(b)

Fig. 3. (a) Images in simple layout; (b) Images in complex layout.

As our images are not quantitatively sufficient, data augmentation is introduced in this paper. In addition to regular methods of panning, rotation, flipping, scaling, colour dithering and adding noise, we propose a stitching-based data augmentation referring to CutMix [24], which is explained in detail below:

For samples with simple layout, we directly divided them into four equal parts by centre point through a script; for those with complex layout, we manually cropped them, keeping only the geometrically shaped text layouts, hand-drawn graphics and irregular parts of them. The cut images were brought together and four of them were taken at a time, each with conventional data enhancement before being stitched together (the gaps in the background are filled with grey), as shown in Fig. 4.

The 8k images obtained after data augmentation are our entire dataset. There are 4000 simple pages and 4000 complex pages, which are named HWAD-s and HWAD-c respectively for convenience of writing. We randomly selected 70% of these images as the training set and the rest as the test set.

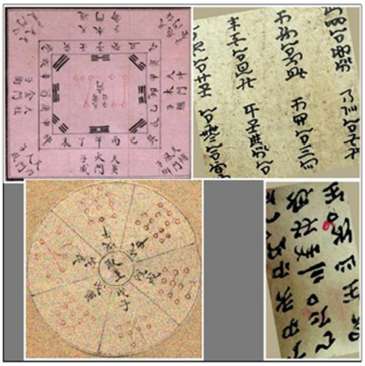


Fig. 4. Image stitching method.

5 Experiments

5.1 Implementation Details

The proposed model is implemented in PyTorch [25] framework using the open source toolkit for object detection, Mmdetection [26]. We conduct experiments on a workstation with 3.6 GHz CPU, RTX 2080 Super 8G GPU and Ubuntu 64-bit OS. SGD is chosen as the optimizer with momentum set to 0.9 and weight decay of 0.0001. We also use a warm up strategy, making the learning rate increase from 0 to 0.0025, so that the model can be stabilized quickly, and to approach this point by reducing the learning rate through cosine annealing as the Loss approaches a global minimum.

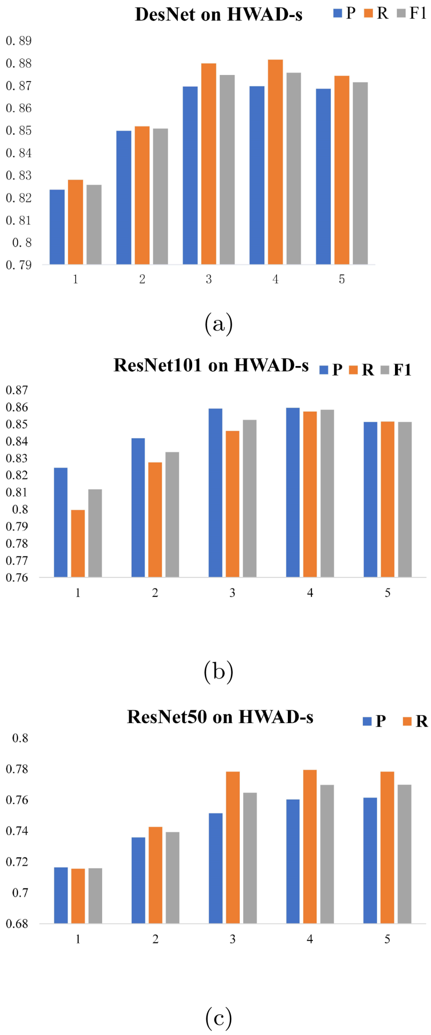
5.2 Ablation Study

Bottom-Up and Skip Connection Fusion. In this part, we applied the bottom-up and skip connection paths to the feature fusion module of FCOS. Considering that the effect will not be obvious if separate experiments are conducted for these two structures, we combined them for testing. In Table 1, we can see that our proposed structure significantly improves the performance for FCOS with ResNet-50, ResNet-101 and DesNet. For the ResNet-50 backbone, performance gain in terms of F-measure of 2.9% and 11.3% is achieved on HWAD-s and HWAD-c respectively. For the ResNet-101 backbone, it brings 1.9% (on HWAD-s) and 3.4% (on HWAD-c) improvements. For the DesNet backbone, the module results in 3.3% (on HWAD-s) and 6.9% (on HWAD-c) improvements.

Table 1. Effect of bottom-up and skip connection fusion structure on HWAD-s and HWAD-c

Method	HWAD-s			HWAD-c		
	P	R	F1	P	R	F1
DesNet	0.801	0.785	0.793	0.640	0.595	0.617
DesNet+skip	0.824	**0.828**	**0.826**	**0.679**	**0.694**	**0.686**
ResNet50	0.698	0.675	0.687	0.432	0.365	0.395
ResNet50+skip	0.716	0.716	0.716	0.498	0.519	0.508
ResNet101	0.814	0.773	0.793	0.513	0.432	0.469
ResNet101+skip	**0.824**	0.800	0.812	0.547	0.465	0.503

Cascade Feature Fusion Network. Based on the experiment above, we further verify the effectiveness of cascade feature fusion network, and explore whether there is saturation in cascade operations. Because the experiment will become very time-consuming as the cascade network deepens, we only conducted the experiment on HWAD-s. As shown in Fig. 5, with the increase of cascade layers, the values of the three evaluation metrics P, R and F1 rise first, reach a peak, and then start to decline instead. Take the DesNet for example, when the number of cascade layers reaches 3, the value of P reaches its maximum. Even though the recall increases when the cascade layers are 4, the F-measure is almost unchanged. The improvement in recall is important, but every additional layer of cascade increases the complexity of the model, and the weak increase in one metric alone is of little significance. When the cascade layers reach 5, the performance even starts to degrade. According to the experimental results, we decided to set the cascade layers as 3 to balance the performance and complexity of the model.



(a)

(b)

(c)

Fig. 5. Experimental results of cascade feature fusion on different backbones: (a) DesNet; (b) ResNet-101; (c) ResNet-50.

Deformable Convolution. We tested the effects of deformable convolution network (DCN) on the DesNet and ResNet. As shown in Table 2, DCN brings a relatively limited improvement on HWAD-s, but on HWAD-c, it generates significant improvements of 9.9% (ResNet-50), 3.1% (ResNet-101) and 6.6% (DesNet) respectively.

Table 2. Performance grows with DCN

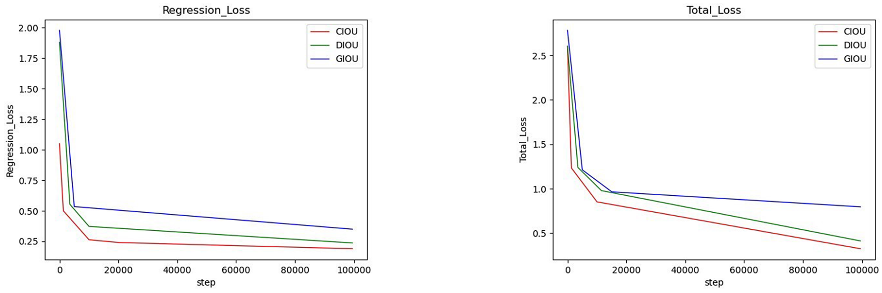
Backbone	DCN	HWAD-s			HWAD-c		
		P	R	F1	P	R	F1
DesNet	×	0.801	0.785	0.793	0.640	0.595	0.617
DesNet	✓	0.818	**0.802**	0.810	**0.679**	**0.687**	**0.683**
ResNet50	×	0.698	0.675	0.687	0.432	0.365	0.395
ResNet50	✓	0.723	0.743	0.733	0.512	0.477	0.494
ResNet101	×	0.814	0.773	0.793	0.513	0.432	0.469
ResNet101	✓	**0.828**	0.801	**0.814**	0.525	0.459	0.490

Comprehensive Experiment. Combining the modules from the previous three experiments makes up our final feature fusion network. We verified the performance of DFCOS and FCOS incorporating this feature fusion network on HWAD-s and HWAD-c. As we can see in Table 3, DFCOS outperforms FCOS on both datasets, and has a substantial lead on the HWAD-c dataset (ResNet-50 18.9%, ResNet-101 18.8%).

Table 3. Comparison of DFCOS and FCOS on HWAD-s and HWAD-c

Method	HWAD-s			HWAD-c		
	P	R	F1	P	R	F1
DFCOS (DesNet)	**0.870**	**0.880**	**0.875**	**0.736**	**0.762**	**0.749**
FCOS (ResNet50)	0.751	0.778	0.765	0.553	0.568	0.560
FCOS (ResNet101)	0.859	0.846	0.853	0.593	0.532	0.561

Loss Function. As shown in Fig. 6, we compared the performance of GIoU, DIoU [7] and CIoU during training. The horizontal axis represents the number of iterations, 99,500 iterations were performed for the three experiments, and the vertical axis is the loss value. It can be seen that the initial value of CIoU is the lowest, the convergence speed is the fastest, and the convergence value is the lowest. At the same time, we respectively compared the total loss value of the model when using these three regression loss functions. As shown in the figure, it can be seen that CIoU is still the best in convergence speed and convergence value.



(a) (b)

Fig. 6. (a) Comparison of three regression loss functions; (b) Comparison of the total loss with three loss functions respectively.

Post-processing Module. Under the same experimental environment, the Soft-NMS algorithm has a higher recall than the NMS algorithm, whether using linear or Gaussian weighting function, and the improvement is more obvious when using Gaussian weighting function: when the threshold is 0.5, the improvement is 0.4% and 1.1% on HWAD-e and HWAD-d, respectively.

5.3 Comparisons with Previous Methods

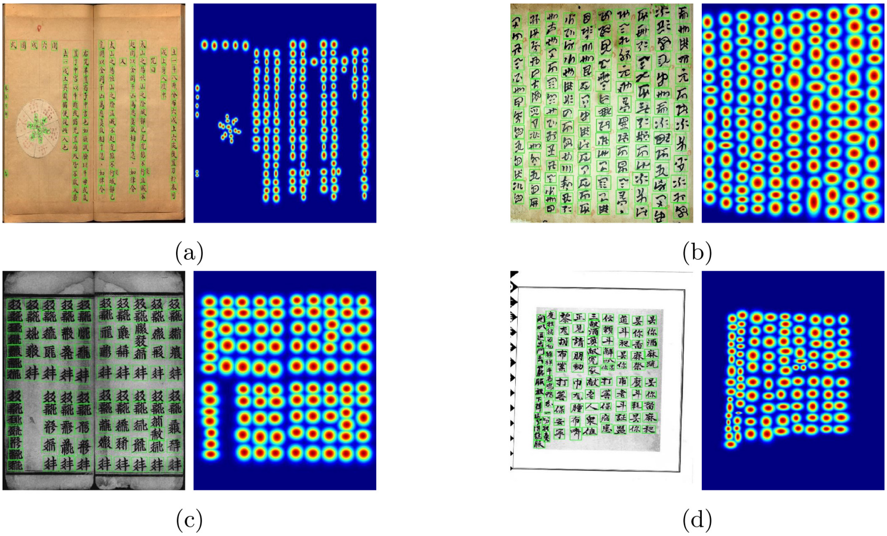
We compare the model we proposed with previous methods for scene text detection on the two datasets, and the experimental results are shown in Table 4 and Table 5. Specifically, our method achieves better performance than the other models on both HWAD-s and HWAD-c in terms of P, R and F1 (and leads more on HWAD-c). Among the methods for comparison, DBNet is the fastest to inference because the DB module is removed when predicting. Additionally, some test examples are visualized in Fig. 7.

Table 4. Comparison with prior arts on HWAD-s

Method	P	R	F1	FPS
CTPN [27]	0.593	0.587	0.590	7.1
SegLink [28]	0.603	0.594	0.598	8.4
RRD [29]	0.633	0.658	0.645	4.5
PixelLink [13]	0.588	0.573	0.590	13.6
EAST [11]	0.625	0.675	0.649	8.7
PSENet [14]	0.744	0.724	0.734	3.7
CRAFT [30]	0.858	0.842	0.850	7.4
DBNet [15]	0.880	0.883	0.881	**15.7**
DFCOS	**0.926**	**0.918**	**0.922**	12.2

Table 5. Comparison with prior arts on HWAD-c

Method	P	R	F1	FPS
CTPN	0.403	0.365	0.383	6.5
SegLink	0.463	0.322	0.380	6.3
RRD	0.433	0.338	0.380	3.2
PixelLink	0.388	0.350	0.368	9.6
EAST	0.429	0.401	0.415	5.7
PSENet	0.524	0.492	0.507	1.8
CRAFT	0.645	0.602	0.623	6.3
DBNet	0.722	0.694	0.707	**11.6**
DFCOS	**0.872**	**0.883**	**0.877**	8.4



(a)

(b)

(c)

(d)

Fig. 7. Detecting results: (a) Chinese; (b) Yi; (c) Tangut; (d) Tibetan.

6 Conclusion and Future Work

In this paper, we propose an effective framework for detecting handwritten ancient texts called DFCOS. According to the characteristics of the text data of ancient books, we propose a new feature fusion network referring to FCOS. We have experimentally demonstrated that our method works well on ancient books. In the future, we hope to enlarge the dataset to improve the generalization performance and optimize the network structure to improve the inference and training speed.

References

1. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. In: ICCV (2019)
2. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
3. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature Pyramid Networks for Object Detection, arXiv preprint. arXiv: 1612.03144 (2017)
4. Handwritten Ancient-Books Dataset: HWAD. Unpublished Data
5. Dai, J., et al.: Deformable convolutional networks. In: ICCV (2017)
6. Bodla, N., Singh, B., Chellappa, R., Davis, L.: Improving object detection with one line of code. In: ICCV (2017)
7. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-IoU loss: faster and better learning for bounding box regression. In: AAAI (2020)
8. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS (2015)
9. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
10. Liao, M., Shi, B., Bai, X., Wang, X., Liu, W.: Textboxes: a fast text detector with a single deep neural network. In: AAAI (2017)
11. Zhou, X., et al.: East: an efficient and accurate scene text detector. In: CVPR (2017)
12. Zhang, C., et al.: Look more than once: an accurate detector for text of arbitrary shapes. In: CVPR (2019)
13. Deng, D., Liu, H., Li, X., Cai, D.: PixelLink: detecting scene text via instance segmentation. In: AAAI, pp. 6773–6780 (2018)
14. Wang, W., et al.: Shape robust text detection with progressive scale expansion network. In: CVPR (2019)
15. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization. In: AAAI (2020)
16. Su, X., Gao, G.: A knowledge-based recognition system for historical Mongolian documents. Int. J. Document Anal. Recogn. Neural Netw. **124**, 117–129 (2020)
17. Shi, X., Huang, Y., Liu, Y.: Text on oracle rubbing segmentation method based on connected domain. In: Proceedings of IEEE Advanced Information Management, Communicates Electronic and Automation Control Conference, pp. 414–418. IEEE Computer Society Press, Anyang (2016)
18. Hailin, Y., Lianwen, J., Weiguo, H., et al.: Dense and tight detection of Chinese characters in historical documents: datasets and a recognition guided detector. IEEE Access **6**, 30174–30183 (2018)
19. Han, Y.H., Wang, W.L., Wang, Y.Q.: Research on automatic block binarization method of stained Tibetan historical document image based on Lab color space. In: International Forum on Management, Education and Information Technology Application, pp. 327–338 (2018)
20. Huang, G., Liu, Z., Maaten, L., Weinberger, Q.: Densely connected convolutional networks. In: CVPR (2017)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: ICCV (2016)
22. Rezatofighi, H., Tsoi, N., Gwak, J.Y., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: a metric and a loss for bounding box regression. In: CVPR (2019)

23. Karatzas, D., et al.: ICDAR 2015 competition on robust reading. In: ICDAR 2015 (2015)
24. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: CutMix: regularization strategy to train strong classifiers with localizable features. In: ICCV (2019)
25. Paszke, A., et al.: Automatic differentiation in PyTorch (2017)
26. Chen, K., et al.: MMDetection: Open MMLab Detection Toolbox and Benchmark, arXiv preprint. arXiv: 1906.07155 (2019)
27. Tian, Z., Huang, W., He, T., He, P., Qiao, Y.: Detecting text in natural image with connectionist text proposal network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 56–72. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_4
28. Shi, B., Bai, X., Belongie, S.: Detecting oriented text in natural images by linking segments. In: Proceedings of CVPR, pp. 3482–3490 (2017)
29. Liao, M., Zhu, Z., Shi, B., Xia, G.-S., Bai, X.: Rotation-sensitive regression for oriented scene text detection. In: CVPR (2018)
30. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection. In: CVPR (2019)