Oracle text positioning system based on improved YOLOv7-FC model

Ruoxu Tian *

Liaoning Normal University, Liaoning 116029, China

* Corresponding Author Email: syu_zsjy@sina.com

Abstract. In recent years, how to use computer to efficiently and accurately recognize oracle characters on oracle rubbings has become the focus of research. Due to the severe damage caused by the natural environment and the characteristics of complex background and loud noise, the oracle bone rubbings are difficult to separate the foreground and back of the oracle bone inscriptions. Therefore, it will be the future research direction to eliminate the negative impact of complex background and accurately locate the Oracle bone inscriptions. This paper puts forward the YOLOv7-FC neural network model and realizes the specific application in the oracle topology data set, and presents the system of single oracle character positioning results.The experimental results show that on the HWOBC-A dataset, the mAP value of the YOLOv7-FC neural network model proposed in this paper reaches 0.937. Compared with the original YOLOv7 model, it has improved by 0.22, effectively balancing the relationship between running speed and average accuracy, and providing feasibility for Oracle text localization machine learning method.

Keywords: Oracle bone inscriptions image, YOLOv7, Text positioning, Semantic segmentation, Object detection.

1. Introduction

Oracle bone inscriptions, also known as oracle bone inscriptions, were written on tortoise shells or animal bones for divination by the royal family in the late Shang Dynasty. Oracle bone script is the earliest mature writing system found in China so far, it is the source of Chinese characters and the root of China's excellent traditional culture, which deserves to be cherished and better inherited and developed.



Figure 1. Oracle rubbings image

At present, the research results of oracle bone inscriptions have been very rich, which can be divided into two research directions: oracle ontology research and oracle bone inscriptions application research. Among them, Oracle application research is mainly carried out from the perspective of computer technology application. The research focus is mainly on oracle bone fragment conjugate [2], Oracle text recognition [3], Oracle semantic retrieva [4], Oracle font generation [5] Other aspects.

The research task of this paper is to use rectangular frame to locate a single oracle character on the oracle topology image. The existing oracle text recognition system generally involves three main steps: text localization in the topology image, segmentation of text foreground and background, and oracle semantic recognition. According to previous researches on text detection methods in natural

scenes [6], The current text detection and location technology can be broadly divided into four directions: Method based on connected region [7], Method based on edge feature [8], Method based on texture feature [9]and A machine learning-based approach [10]. However, because of the particularity of oracle shape and the complexity of oracle topology background, it is difficult to accurately detect the specific position of oracle ontology text for text segmentation.

The research focus of this paper is to improve the YOLOv7-FC neural network model of activation function FReLU and attention-increasing mechanism CoT, and realize the specific application in oracle topology data set to present the positioning results of a single oracle character. The data set used in this paper selects 300 images from the Oracle topology test set in the HWOBC-A data set to train and test the model [11].

The rest of this paper is organized as follows: The second part is a detailed analysis and discussion of the previous research on oracle bone inscriptions; The third part introduces the main research methods and the improved network model. In the fourth part, the experiment is carried out by training the network model, and the hyperparameter is optimized through the verification process. The fifth part evaluates the data based on the results of the training and testing of the algorithm, and the sixth part summarizes the research work and looks forward to a better method of oracle text positioning in the future, so as to achieve a better inheritance and development of oracle text.

2. Related work

In recent years, researchers have made a lot of achievements in oracle image classification recognition and oracle image text localization.

In 2011, Li [12] proposed an oracle recognition algorithm based on graph isomorphism, which converts oracle into an undirected labeled graph and encodes oracle using the adjacency matrix of the undirected labeled graph. Compared with other approximation algorithms based on graph isomorphism, this algorithm has no misjudgment and short running time.

In 2014, Guo [13] proposed a new grid point feature extraction algorithm, which added point features into coarse grid features, converted absolute addresses into relative addresses, grided the point features, integrated the position relationship into the feature vector, and Auccracy of neatly written oracle rubbings reached 82.63%. On the basis of experiment, the accuracy of coarse mesh feature extraction algorithm is improved by 9%. In the same year, Shi [14] proposed an oracle bone rubbings text location method based on threshold segmentation and morphology. First, threshold values were selected to achieve rough text segmentation, and all text areas and noise were divided into white and background into black. Then morphological methods were used for fine localization, almost all non-text areas were excluded. The Precision was 68 percent and the Recall 61 percent.

In 2015, Gao [15] proposed a strategy based on the combination of context statistical analysis and Hopfield network for fuzzy oracle images, and finally calculated the fuzzy feature similarity according to OBI Optical Beat Interference to determine the shape of the target oracle. The final Auccracy for fuzzy oracle identification reached 92%.

In 2019, Chen [16] proposed an oracle recognition technology based on coding, converting oracle images into codes in the process of image pre-processing calculation, and on this basis proposed the implementation of oracle recognition system. Under the 3*3 template, the average image processing rate of the system is about 52 images/s. Each single word image has its own unique and significantly related encoded data. Auccracy is as high as 100% in the "CunZhongnan" Oracle bone dataset, with efficient recognition.

In order to solve the problem of time-consuming and labor-intensive entry of oracle characters, in 2020, LI [11] established a handwritten oracle character database named HWOBC-A, which contains 83, 245 character-level samples divided into 3,881-character categories. In addition, they also describe the performance of several methods based on deep convolutional neural networks, among which Melnyk Net [17] as a baseline model shows the best accuracy of 97.64%, which has contributed to the subsequent research on oracle recognition.

In 2021, Zhang [18] proposed a method of bone recognition based on cross-modal deep metric learning, which realized the cross-modal recognition of bone recognition through shared feature space modeling and nearest neighbor classification of the imitation bone text and the rubbing bone text. Finally, the Auccracy of CNN integrated with cross-modal information on oracle data set provided by Oracle Information Processing Laboratory of Anyang Normal University [19] reaches 88.4%, which has obvious advantages compared with traditional CNN classification framework and single mode recognition method.

In the same year, Men [20] used the EAST model [11] to train the character recognition model [22]and realized a system that could automatically locate and recognize the oracle bone characters one by one on the whole rubbing and automatically classify them according to the font. The Auccracy of the recognition system was approximately 90%. On this basis, the oracle font database with complete materials, convenient viewing and open editing of paragraph information is constructed, which makes the font database a convenient tool for image retrieval of Oracle.

The above analysis shows that the image processing of oracle rubbings based on machine learning has become a research focus in recent years. Although there are a lot of previous studies, they are limited to large model deployment and slow reasoning speed. Due to the impact of natural environment, the image of oracle bone rubbings is seriously damaged, the image background is complex, and the noise is large, so it is difficult to separate the foreground and rear scene of the oracle bone text. Therefore, the future research direction will be to eliminate the negative impact of the complex background and accurately locate the Oracle bone text.

3. Experimental method

3.1. YOLOv7-FC network architecture

Specific problems encountered in the process of Oracle topology text positioning in the original YOLOv7 model [23] have improved the overall network structure of YOLOv7 in two places. The improved network structure diagram, YOLOV7-FC, is shown in Figure 2.

Firstly, to solve the problem that YOLOv7 model is weak in spatial feature extraction, visual activation function FReLU [24] is used to replace SiLU. Compared with the traditional ReLU activation function, FReLU has stronger nonlinear expression ability and better learnability, so it can better adapt to the complex input data distribution. Using the FReLU activation function in YOLOv7 brings two advantages. First, FReLU can better detect nonlinear features in the input data, thus improving the expressiveness of the model. Second, FReLU can adaptively learn the parameters of the activation function during the training process, so that it can better adapt to different data distributions and improve the generalization ability of the model.

By using the FReLU activation function, YOLOv7 can better capture features in the image and therefore improve the Precision and Recall of target detection. Ablation experiment results showed that the mAP value of YOLOv7 was improved after the FReLU activation function was used.

Subsequently, CoT attention mechanism [3] was added to YOLOv7 model to replace 3x3 convolution in order to improve YOLOv7's performance in object detection tasks, aiming at the insufficient performance of YOLOv7 model in text positioning. Since CoT can effectively extract contextual information from images, applying it to YOLOv7 can help models better understand image content and improve the accuracy and recall rate of target detection. By using the CoT, 3x3 convolution can be replaced with some modules containing a self-attention mechanism and Transformer network, allowing the model to better capture contextual information in the image, thus improving the accuracy of the model.

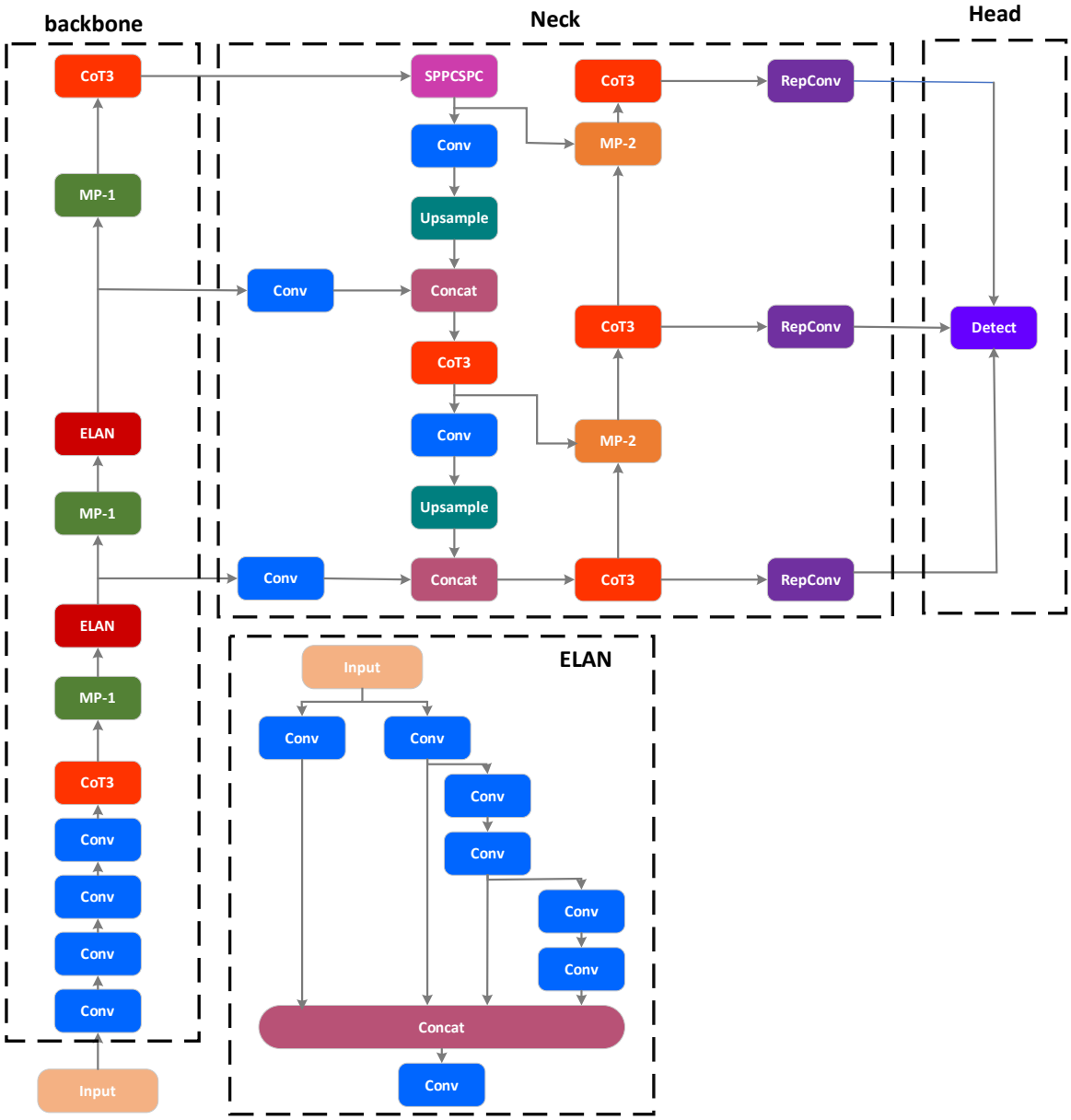


Figure 2. YOLOv7-FC network structure

3.2. FReLU activation function

FReLU [19] is a visual activation function that is simple in concept and can effectively complete image recognition tasks. By increasing the overhead of space conditions, ReLU and PReLU are extended to 2d activation. The expression of ReLU is shown in formula (1), the expression of PReLU is shown in formula (2), and the expression of FReLU is shown in formula (3). Where, T () is a two-dimensional space condition, and the pixel-level modeling capability is realized in a simple way, and the complex visual layout is captured by regular convolution.

$$f(x) = max(x,0) \tag{1}$$

$$f(x) = max(x,px) \tag{2}$$

$$f(x) = max(x,T(x)) \tag{3}$$

The original YOLOv7 model extended the activation function SiLU of YOLOv5 model, whose expression is shown in formula (4).

$$f(x) = x \times sigmond(x) \tag{4}$$

As shown in Fig. 3 and Fig. 4, by comparing the function images of SiLU and FReLU, it can be seen that the SiLU activation function image is derivable everywhere and continuously smooth. The application of activation function on the network model can improve the classification accuracy, but it also increases the calculation amount because of the introduction of exponential function. However, FReLU improves the spatial sensitivity of the activation function by adding a space condition, and only adds a negligible computing cost, and the implementation difficulty is not high.

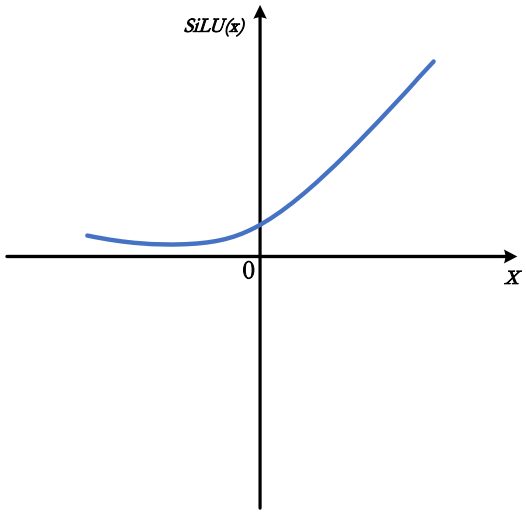


Figure 3. SiLU (x) function image

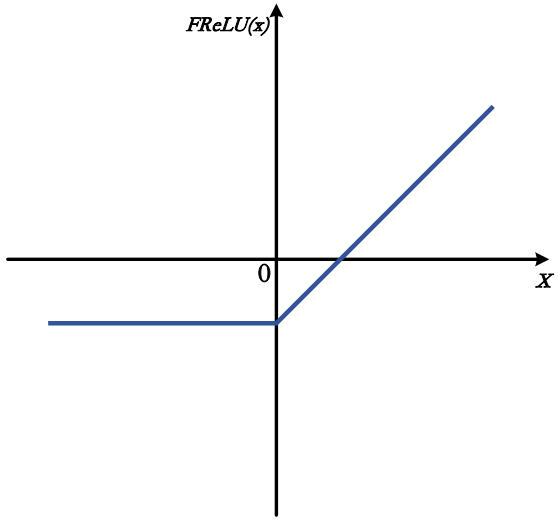


Figure 4. FReLU (x) function image

3.3. CoT attention mechanism

The CoT (Contextual Transformer) attention mechanism [20] is a novel Transformer-style module for visual recognition. Such a design makes full use of the context information between input keys to guide the learning of dynamic attention matrix, thus enhancing the ability of visual representation. Technically, the CoT block first encodes convolution of the input keys by context, resulting in a static context representation of the input. The model further concatenates the encoding key with the input query through two successive convolutions. The learned attention matrix is multiplied by the input values to achieve a dynamic context representation of the input. The fusion of static and dynamic context representations is finally taken as output.

The CoT Block replaces the 3x3 convolution in the ELAN structure, and the Self-Attention structure in the original Transformer (shown above) only calculates the attention matrix based on the interaction of query and key, thus ignoring the connections between adjacent keys. First, 3x3 convolution is used on the key to model the static context information, and then the query and the key after the context information is modeled are concat, and two consecutive 1x1 convolution is used to generate the dynamic context. The static and dynamic context information is finally merged into the output, Fig. 5 is the traditional Transformer and Fig 6 is the CoT module.

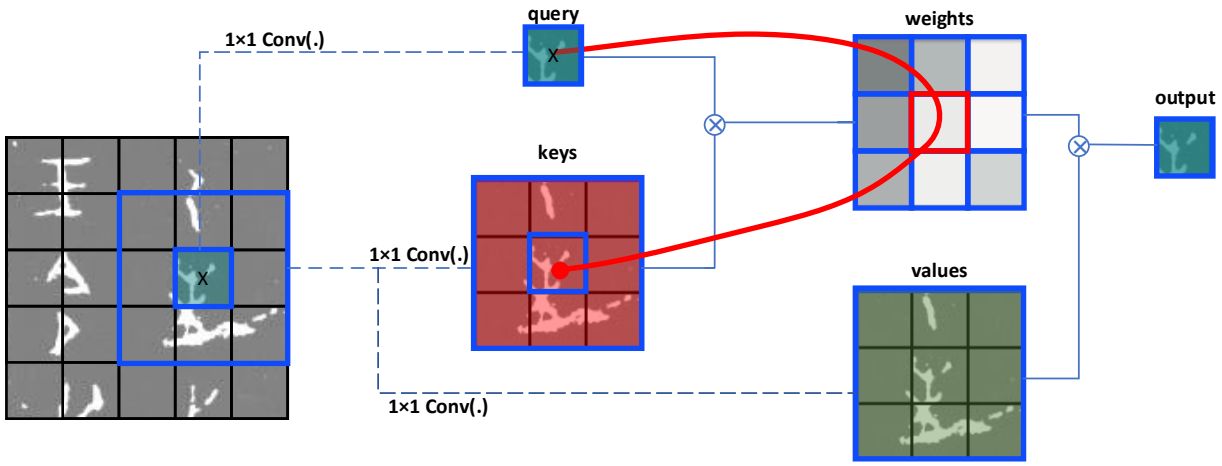


Figure 5. Traditional attention mechanisms

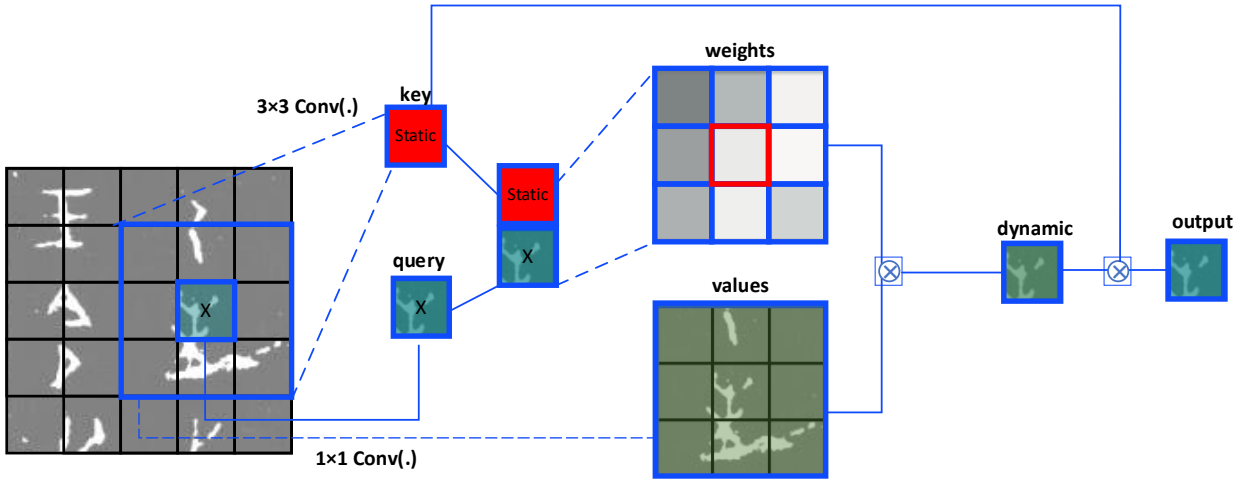


Figure 6. CoT attention mechanism

4. Experiment

4.1. Data set introduction

In this paper, the test set in HWOBC-A oracle handwriting dataset [11] is selected for training and testing. Among them, the test set in HWOBC-A contains 9516 scanned images of oracle rubbings, which is rich in targets and more difficult to classify. In the experiment, 300 rubbings images were selected from the collection and divided into three categories, as shown in Fig 7 below. The first type is the rubbings with clear background and less noise, but incomplete strokes of the text image; The second type is the rubbing image with heavy background fuzzy noise; The third type is the rubbings with clear background and less noise, and the number of each type of image is controlled at 100. To train our model, we used 300 images for training and 60 images for testing. The ratio of the training data set to the test data set is five to one, and the ratio is the same for the above three categories. in this experiment, tailor, crop, overturn and paste in Mosaic data enhancement were used to enhance the data set.

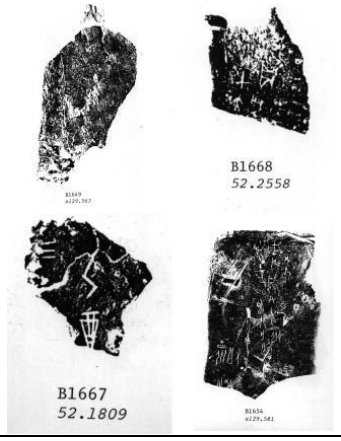
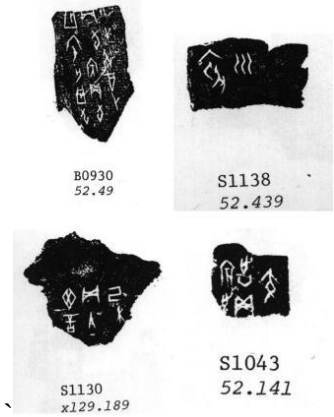
The first type Clear background and text with less noise	The second type Background blur noise is heavy	The third type Clear background and low noise
		

Figure 7. Test set classification

4.2. Data preprocessing

In the data enhancement part, Mosaic data enhancement module [4] was used to enhance the existing samples. Mosaic is a data enhancement technique used to increase the diversity and number of training datasets. The basic idea of Mosaic is to combine multiple different images into one large image, and randomly crop out several small images for training.

Specifically, Mosaic stitched together four different images into one large image, as shown in Fig 8. Each small image covered a quarter of the larger image, and then randomly cropped multiple small images for training. During cropping, different parts of the larger image may be combined into a new image, increasing the diversity of the training data.

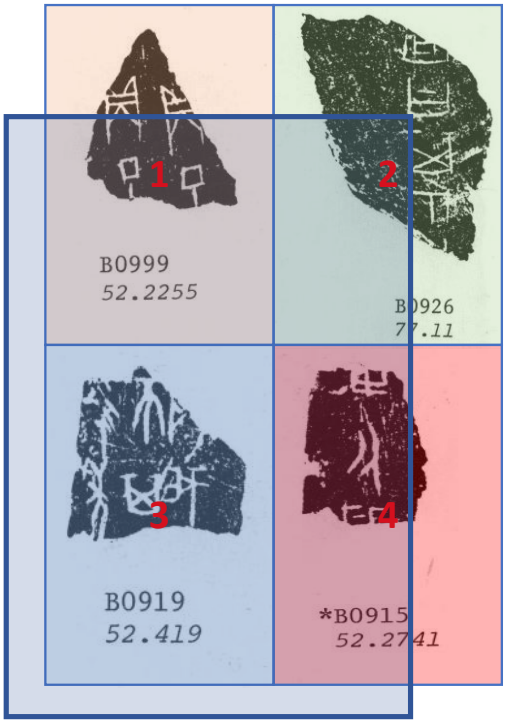


Figure 8. Enhanced with Mosaic data

Mosaic data enhancement technology can effectively increase the amount and diversity of training data, so as to improve the generalization ability and robustness of the model. In addition, Mosaic can also reduce the occurrence of overfitting phenomenon and improve the performance of the model. Mosaic data enhancement technology has been verified in many computer vision applications, including object detection, image classification, semantic segmentation and so on.

4.3. Experimental environment

The hardware accessories and experimental environment of this paper are shown in the following table. The GPU configuration used in this paper is NVIDIA RTX A4000, and the whole experiment is developed in Python language, using PyTorch deep learning framework.

Table 1. Software and hardware environments

Graphics card	NVIDIA RTX A4000
Internal memory	16GB
python	3.8.10
CUDA	11.2
torch/torchvision	1.12.1/0.13.1

4.4. Ablation experiment

Ablation experiments were designed for the YOLOv7 model proposed in this paper to verify the effectiveness and accuracy of each new module. Average accuracy is mainly adopted as the main experimental evaluation standard in the experiment. On the training set based on manual annotation, 300 epochs are used, batch size is 8, and the loss function is shown in Fig 9. To train the original YOLOv7 model, the YOLOv7-Frelu model that changes the activation function on the baseline model, the YOLOv7-COT model that adds the CoT attention mechanism to the baseline model, and the YOLOv7-FC model that finally integrates the first two parts, as shown in the following table.

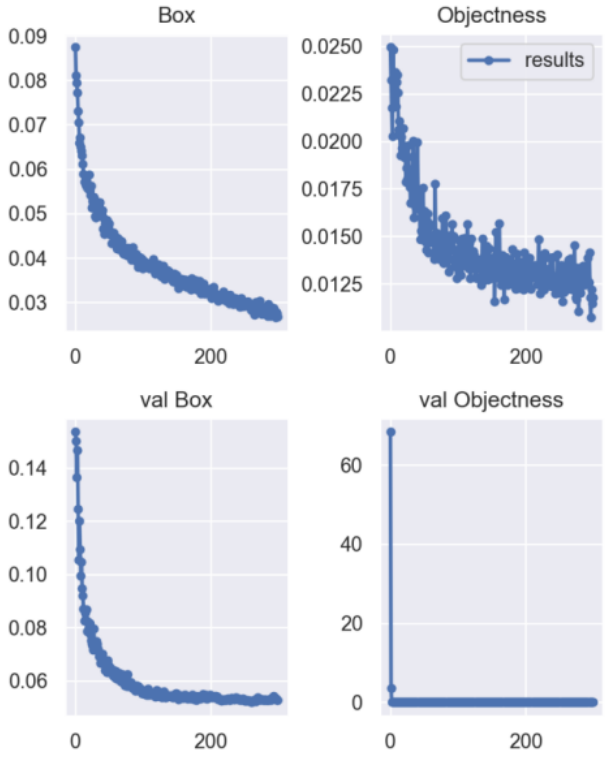


Figure 9. Loss function

Table 2. ablation data

Improvement section			Models	mAP(%)	F1 Score
SiLU	FReLU	CoT			
	√		YOLOv7	0.915	0.952
√			YOLOv7-FReLU	0.935	0.908
		√	YOLOv7-CoT	0.915	0.890
√	√	√	YOLOv7-FReLU-CoT	0.937	0.923
Improvement section			Models	mAP(%)	F1 Score

The test batch results are shown in Figure 10. These are the results of our training and testing phase on the final improved YOLOv7-FC model. We can see that the first type of oracle text incomplete topology image and the second type of topology image with loud noise and fuzzy background in the training set can successfully identify the box and select the text area in the YOLOv7-FC model, achieving the text positioning effect.



Figure 10. Test result

4.5. Evaluation criteria

In order to conduct a comprehensive and scientific evaluation of the model of the four ablation experiments, the two evaluation indexes mainly used in this paper are mAP (average accuracy) and F1 Score.

The IoU (intersection ratio) threshold for judging positive and negative samples set in this paper is 0.5. When the intersection ratio between the actual annotation box and the algorithm detection box is greater than the set threshold, the sample is correctly predicted and is in the detection box, marked as positive sample negative sample (TP). When the intersection ratio between the actual annotation frame and the algorithm detection frame is greater than the set threshold, the sample will be marked as false positive (FP). When the sample is the actual labeled box region but not the algorithm detection box region, the region of the sample that is correct is predicted to be false. It will be marked as a false negative (FN).

The ratio of the number of true positives (TP) samples to the total number of detected target class t is accuracy, indicating the proportion of all results that are correctly retrieved, see formula (5).

$$P_t = \frac{TP_t}{TP_t + FP_t} \tag{5}$$

The ratio of the number of true positive (TP) samples to the number of correctly predicted samples and the false negative (FN) that correctly predicted errors into the target class t is the recall rate, indicating the probability of the model finding the correct object, see formula (6).

$$R_t = \frac{TP_t}{TP_t + FN_t} \tag{6}$$

With precision P as the vertical axis and recall rate R as the horizontal axis, a two-dimensional P-R curve can be drawn. The area of the besieged city under the P-R curve is the average accuracy, see Formula (7).

$$AP_t = \int_0^1 P_t(R_t)dR_t \tag{7}$$

The AP values of each category were calculated separately, and the sum of all AP values was divided by the number of categories to get the average precision mAP. However, since the experiment in this paper only involved the separation of oracle text area and topology background, that is, there was only one category, mAP = AP.

F-Measure was used to evaluate the model detection ability. The average harmonic values of accuracy P and recall rate R were represented by factors, as shown in Formula (8).

$$F1_t = \frac{1}{|T|}\sum_{t\in T}\frac{2P_tR_t}{P_t+R_t} \tag{8}$$

5. Summery

In the experiment of this paper, the average accuracy is used to evaluate the effect of the improved module, as shown in the following table. Firstly, the baseline model of YOLOv7 and the trained YOLOv7 were used as the pre-training model for the annotated data set. The Precision was 0.967, the Recall was 0.94, and the mAP was 0.915. Then the 3*3 convolution in ELAN can be replaced by the CoT3 attention mechanism to better obtain the context information of the topology image. Compared with the original YOLOv7 model, the Precision is 0.855, the Recall is 0.97, and the mAP is 0.915, as shown in Fig 2. Recall improves the running speed of the algorithm by 0.3 points, but the accuracy is lower than that of the baseline model, and the average accuracy does not change. Then, the study on accuracy improvement was carried out on the original YOLOv7 model. The activation function in the original YOLOv7Conv layer was changed from SiLU activation function to FReLU activation function, which enhanced the sensitivity of the activation space and significantly improved the image vision. The Precision was 0.831. Recall is 0.96 and mAP is 0.935, as shown in Fig 3. The average accuracy is improved by 0.2 points, but the training speed drops to two to three times that of the original model. In order to solve the shortcomings of the above two improved directions, the YOLOv7-FC model was finally formed by the fusion of the two directions. The Precision was 0.882, the Recall was 0.97, and the mAP was 0.937, as shown in Fig4. The average precision was increased by 0.22 on the basis of the baseline model, and the running speed was increased. In general, the improved YOLOv7 network model not only improves the text localization of single characters in Oracle dataset, but also effectively balances the relationship between running speed and average accuracy, providing feasibility for the machine learning method of Oracle text localization.

Table 3. Experimental data

Models	Numbers of Images	Precision	Recall
YOLOv7	300	0.967	0.94
YOLOv7-CoT	300	0.855	0.97
YOLOv7-FReLU	300	0.831	0.96

6. Conclusion

In order to solve the difficulty of oracle text localization due to fuzzy background and loud noise of oracle topology image, an improved detection algorithm based on YOLOv7 was proposed. First, the visual activation function FReLU was introduced to replace the original SiLU to better extract spatial visual information and improve the Precision of Oracle text positioning. Then, CoT attention mechanism was introduced to better obtain picture context information, pull up Recall, and improve the running speed of the algorithm. Compared with the training and test data of the original YOLOv7, the improved YOLOV7-FC model has the highest mAP value and fast operation efficiency, but it still cannot achieve accurate text positioning for the noisy oracle image topology. The future research

direction is to further study the text localization of fuzzy oracle topology image by comparing with more classical object detection neural network models.

References

[1] Wang Aimin, Zhong Luo, Ge Yanqiang, et al. Research on Key Technologies for Intelligent Conjugation of Oracle Bone Fragments [J]. Journal of Wuhan University of Technology, 2010 (20): 194 - 199.

[2] Liu Fang, Li Huabiao, Ma Jin, et al. Research on Automatic Detection and Recognition of Oracle Rubbings Based on Mask R-CNN [J] Data Analysis and Knowledge Discovery, 2021 (12): 88 - 97.

[3] Han Jiaohong. Research on ontology-based semantic retrieval model for oracle bone literature [J] Library Science Research, 2013 (07): 51 - 57.

[4] Li Qingsheng, Wu Qinxia, Yang Yuxing, et al. Research on oracle bone script font dynamic description library and its font generation technology [J] Journal of Peking University (Natural Science Edition), 2013 (01): 61 - 67.

[5] Li Dongqin, Wang Fang, Zhou Wanhuai, et al. Research on Text Detection Methods in Natural Scene Images [J] Journal of Chongqing University of Science and Technology (Natural Science Edition), 2016, 18 (06): 108 - 111.

[6] Yi C, Tian Y L. Text string detection from natural scenes by structure-based partition and grouping [J]. IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, 2011 (9): 2594 - 2605.

[7] Zhao, X, Lin, K.-H, Fu, Y, Hu, Y, Liu, Y, Huang, T. S. Text from Corners: A Novel Approach to Detect Text and Caption in Videos[J]. Image Processing, IEEE Transactions on, 2011, (3): 790 - 799.

[8] Shivakumara P, Phan T Q, Tan C L. New Fourier-Statistical Features in RGB Space for Video Text Detection [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2010 (11): 1520 - 1532.

[9] Jung C,Liu Q, Kim J.A stroke filter and its application to text localization [J]. Pattern Recognition Letters, 2009 (2): 114 - 122.

[10] Li B, Dai Q, Gao F, et al. HWOBC-A Handwriting Oracle Bone Character Recognition Database [J]. Journal of Physics Conference Series, 2020, 1651: 012050.

[11] Li Qingsheng, Yang Yuxing, Wang Aimin A Graph Isomorphism Method for Oracle Bone Inscription Recognition [J] Computer Engineering and Applications, 2011, 47 (8): 3.

[12] Guo Lei. Research on feature extraction algorithm for neatly written oracle bone rubbings [J]. Computer Applications and Software. 2014 (6): 4.

[13] Shi Xiaosong, Huang Yongjie, Liu Yongge. A text localization method based on threshold segmentation and morphology for oracle bone rubbings [J]. Journal of Beijing University of Information Technology (Natural Science Edition), 2014 (06): 7 - 10.

[14] Feng, G., Jing, X., & Yongge, L. (2015). Recognition of Fuzzy Characters on Oracle-Bone Inscriptions. IEEE International Conference on Computer & Information Technology Ubiquitous Computing & Communications Dependable. IEEE.

[15] Chen Tingzhu, Wu Shaoteng, Wu Jiang, et al.Research on Encoding Based Oracle Recognition Technology [J]. Chinese Character Research, 2019 (1): 12.

[16] Melnyk P, You Z, Li K. A high-performance CNN method for offline handwritten Chinese character recognition and visualization [J]. Soft Computing, 2019, 24 (8).

[17] Zhang Yikang, Zhang Heng, Liu Yongge, et al. Oracle bone character recognition based on cross modal depth metric learning [J]. Journal of Automation, 2021 (4): 10.

[18] Gaofeng, Liu Yongge, Han Shengwei. From "Digital Oracle Bone" to "Intelligent Oracle Bone" - A New Paradigm of "Yin Qiwen Yuan" Assisting Oracle Bone Research [J]. Oracle Bone Inscription and Yin Shang History, 2021 (00): 600 - 610.

[19] Men Yi, Zhang Chongsheng. Artificial intelligence-based oracle bone recognition technology and font database construction [J]. Chinese Character Research, 2021 (1): 8.

[20] Zhou X, Yao C, Wen H, et al. EAST: An Efficient and Accurate Scene Text Detector [J]. IEEE, 2017.

[21] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. Computer Science, 2014.

[22] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors [J]. arXiv e-prints, 2022.

[23] Ma N, Zhang X, Sun J. Funnel Activation for Visual Recognition [J]. 2020.

[24] Li Y, Yao T, Pan Y, et al. Contextual Transformer Networks for Visual Recognition [J]. 2021.

[25] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection [J]. 2020.