applied sciences

Article

Research on Bone Stick Text Recognition Method with Multi-Scale Feature Fusion

Mengxiu Du [1], Huiqin Wang [1,*], Rui Liu [2], Ke Wang [1] and Zhan Wang [1]

[1] School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China
[2] Institute of Archaeology, Chinese Academy of Sciences, Beijing 100101, China
* Correspondence: hqwang@xauat.edu.cn

Abstract: Bone sticks are composed of thin slices of animal bones created by ancient people, which mainly served the functions of fixing books, writing scripts, and divination. The bone stick script is an essential material for studying the history of Chinese Western Han script. Using a neural network for text recognition can quickly interpret ancient text, while extracting deeper semantic information, neural networks will also lose superficial image details. After multi-layer convolution and pooling of bone sticks, the continuous loss of superficial details affects classification accuracy. At the same time, the unbalanced distribution of bone stick quantity leads to a low recognition rate with small samples of bone sticks. Aiming to solve the above problems, a bone stick recognition method based on multi-scale features and focal loss function is proposed. Firstly, based on the residual network ResNet, the output features of the first layer and four Conv_x layers are pooled globally to reduce the feature dimension of each channel, and the channel splicing method is used to add different depths of base information to the original high-level features, which improves the detail feature extraction ability of the model. Secondly, in view of the unbalanced distribution of the bone stick data, the original cross-entropy loss function is replaced by the focus loss function, which increases the penalty for classification errors and improves the recognition rate of classes with few training samples. Experimental results show that the recognition accuracy of the proposed method on the bone stick data set is up to 90.5%.

Keywords: bone stick recognition; multi-scale; feature layer fusion; focus loss function

check for updates

Citation: Du, M.; Wang, H.; Liu, R.; Wang, K.; Wang, Z. Research on a Bone Stick Text Recognition Method with Multi-Scale Feature Fusion. *Appl. Sci.* **2022**, *12*, 12507. https://doi.org/10.3390/app122412507

Academic Editor: Lidia Jackowska-Strumillo

Received: 1 November 2022
Accepted: 2 December 2022
Published: 7 December 2022

1. Introduction

With the progress of modern science and technology, the identification methods of cultural relics have been gradually developing. More than 60,000 bone sticks have been unearthed from the third building site of the Weiyang Palace, of which more than 57,000 are inscribed with hundreds of thousands of characters. These are very precious archaeological materials for studying the history and culture of the Western Han Dynasty [1]. The Western Han Dynasty was a crucial period in the development and change of ancient Chinese calligraphy and a confirmed period for modern Chinese characters. The bone stick text is of great academic importance for studying the derivation of calligraphic strokes in the Western Han Dynasty and exploring Chinese characters' ancient and modern script changes [2]. The bone stick text not only provides a large amount of biological data for the study of Han official offices, weapons manufacturing management, imperial city architecture, and the history of the Han dynasty but also provides an essential basis for the study of Han archives and the history of ancient archives [3]. Correctly interpreting bone stick information is essential for exploring the history and culture of this period. However, bone stick text recognition relies only on experts to discriminate manually, which is time-consuming and laborious. Using image processing and artificial intelligence technology as a primary research method can assist scholars in quickly interpreting ancient texts or realizing image retrieval based on text content, improving work efficiency.

In recent years, convolutional neural networks have achieved inclusion in suitable applications in the field of image recognition. Ziyi Wu et al. [4] proposed Chinese character recognition based on an integrated attention layer convolutional neural network, which can comprehensively extract Chinese character feature information and distinguish shapes similar to Chinese characters. However, this method has high complexity and low training efficiency. Guoying Liu [5,6] detailed the research progress of oracle character recognition and oracle character detection from the perspectives of the application of traditional methods to the attempt to use deep learning techniques and described the technical details of the appropriate methods, the information of the used data set, and the actual performance. Mengting Liu [7] used strip convolution instead of ordinary convolution to reduce the number of neural network parameters and achieve lightweight recognition of oracle bone rubbing. Hao-Bin Wang [8] combined region-based complete convolutional networks and feature pyramid networks to design and build a basic oracle bone character recognition algorithm framework. Li Wenying et al. [9] proposed a deep learning-based inscription recognition method, which can accurately recognize text using a neural network model with two-stage feature mapping. However, the algorithm is limited by the amount of sample data. Ru et al. [10] introduced a deformation convolution module based on a traditional convolutional neural network to improve a network's modeling ability and training accuracy. However, the method can only recognize simple handwritten numerals, which has certain limitations. Yanlong Luo [11] et al. improved the ResNet network structure to achieve oracle handwriting recognition, but the recognition rate is not high for complex background characters. Chunshan Wang [12] et al. proposed multi-scale vegetable disease recognition to achieve a lightweight model, but the recognition accuracy is not high. Kuang-Bo et al. [13] proposed multi-scale bronze inscription topography detection using a repetitive feature pyramid network, dual-FPN (dual-feature pyramid network), to more effectively fuse the feature maps of each scale in the model.

In this paper, we propose a multi-scale fusion method for bone stick text recognition which enhances the recognition rate of bone stick text by effectively using the detailed information continuously lost during the convolution process. In the model-building process, the last layer of features of each residual block in the ResNet convolution process is firstly retained. These five scale features are used to reduce the high-dimensional information using the maximum pooling operation. Then, the different scale features are fused using the channel splicing method. The corresponding fully connected layer parameters are increased, and this paper designs a comparison test of multiple fusion methods between different layers. In order to balance the recognition accuracy between different categories, internal weighting is performed using the focal loss function to further improve the accuracy. By comparing each neural network experiment and the ablation experiment, it can be found that the method introduced in this paper has the best ability for bone stick text recognition.

2. Fundamentals

2.1. Multi-Scale Feature Fusion Profile

The neural network extracts the image features by quasi-overconvolution, and the perceptual field of the target can directly determine the merit of the acquired features. As the network deepens, the perceptual field of the deep features is larger than the shallow layer, the image information is compressed, the image detail information in the deep feature map is continuously weakened, and it has a more vital semantic representation in recognition. However, its resolution is low, and the shallow-detail information is weakly acquired. The shallow information perception field is relatively small, but the image resolution is high and contains detailed information but less semantic information and more noise.

This paper's multiscale feature fusion is inspired by the pyramid model (feature pyramid networks). FPNs are cost-effective and introduce the multiscale concept by reusing feature maps in feature extraction networks. FPNs mainly improves upsampling after network downsampling and obtains feature maps that are downsampled at the same scale through lateral links equivalent to the deep network. The information embedded in the feature map in the deep network is passed to the shallow layer for fusion by upsampling, and the feature map information is fully utilized to pass down the features from the higher layers to supplement the semantics of the lower layers, so that high-resolution and strongly semantic features can be obtained [14,15].

2.2. Neural Network Model Selection

Deep learning has been successfully applied to image recognition in recent years, and a series of network models with excellent performance has emerged. The AlexNet [16] network is the first neural network framework, the model only has five layers of convolution and three layers of pooling, which has weak feature extraction ability for bone stick text recognition. In the Vgg16 [17] network, most convolutional kernels are of 3 × 3 size; when Vgg16, which is without a residual connection module, is compared to ResNet [18], the training time of bone stick text is too long and inefficient. DenseNet [19] stacks a large number of DenseBlock blocks, and each layer is connected to all previous layers in the channel dimension, which is a dense connection when compared to ResNet. The number of front and back dense connections is too large, the vast number of network parameters will lead to the overfitting of small-sample class training, and the accuracy of bone stick text recognition is not high. The ResNet model first proposed a residual jump connection structure that can effectively solve the problems of gradient disappearance, gradient explosion, and degradation. It solves the problem of the contradiction between neural network depth and recognition accuracy, can effectively extract more detailed features of images, and to a certain extent, improves the phenomenon of data loss caused by convolution. Moreover, the residual structure of ResNet only learns the difference between input and output, simplifies the target, and reduces the learning. In summary, the number of layers of the ResNet neural network, residual structure design, and a moderate number of network parameters is more suitable for bone stick text recognition, so the ResNet network is used as the improved base network in this paper.

2.3. ResNet Network Model Structure

The main innovation of ResNet is the idea of residual learning. ResNet's residual structure unit enables one layer's input to be connected to the next layer through bypass without loss. The basic residual structure is shown in Figure 1.

In the above structure, by taking a shortcut instead of learning the identity function, the function to be fitted is as in Equation (1).

$$H(x) = F(x) + x \tag{1}$$

In the above equation, $H(x)$ represents the function to be fitted, and $F(x) = 0$ is the residual term. The structure needs to be transformed into Equation (2) to achieve the same constant mapping purpose.

$$F(x) = H(x) - x \tag{2}$$

When $F(x) = 0$, that becomes a constant mapping in the learning process because the parameters of each layer of the network are initialized similar to 0, so $F(x) = 0$ is more accessible than learning $H(x) = x$, and the parameters converge faster.

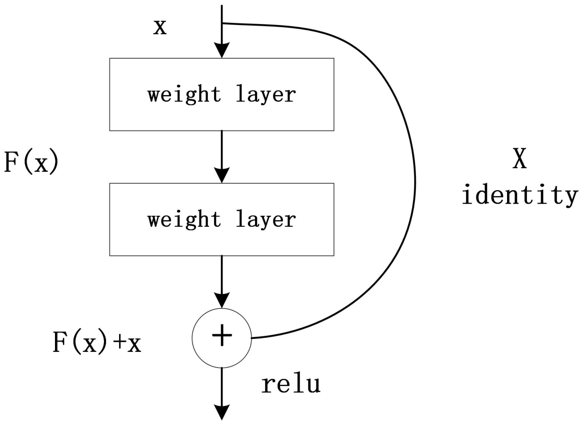

Figure 1. Schematic diagram of residual block.

2.4. Focal Loss Function

The proportion of each type of text sample in the bone stick text data set is unbalanced. The categories with large samples are close to 300 images, and some samples have only a few images. The uneven distribution of samples will lead to an over-recognition of classes with large samples and an insufficient recognition of classes with small samples. Lin et al. proposed focal loss [20] to solve the class imbalance problem by reducing the internal weight. This function focuses on using data sets with sparse and complex samples for training. Although the number of simple samples is large, it contributes little to the total loss. Focal loss is an improvement of the cross-entropy function. The definition of the original cross-entropy loss function [21] is shown in Equation (3).

$$L_{CrossEntropy} = -ylogy' - (1-y)log(1-y') \tag{3}$$

The modulation factor is multiplied by the original cross-entropy loss function, which is defined as Equation (4)

$$L_{FocalLoss} = -\alpha(1-y')^{\gamma}logy' - (1-\alpha)(1-y)y'log(1-y') \tag{4}$$

Where $L_{CrossEntropy}$ represents the loss value generated by the cross-entropy loss function, $L_{FocalLoss}$ represents the loss value generated by the focal loss function, and y is the actual category; y' is the predicted value of the category; the initial setting of $\alpha = NAN$ can be used to balance the uneven number of positive and negative samples themselves, i.e., category imbalance; γ is a positive adjustable parameter, and the setting of $\gamma = 2$ attenuates the error contribution of easy-to-identify samples so that the model can focus more effectively on the hard-to-classify samples during training, and for accurately classified samples y' tends toward 1 and the modulation factor $(1-y')^{\gamma}$ tends toward 0, so that the loss value is reduced; and for misclassified samples, $(1-y')^{\gamma}$ converges to 1, and the modulation factor has no effect on the loss function. Compared with the cross-entropy loss, the focal loss does not change for misclassified samples and becomes smaller for accurately classified samples. It is equivalent to increasing the weight of the inaccurately classified samples in the loss function. By reducing the weights of simple negative samples, the imbalance between positive and negative and complex and easy samples in the data set can be alleviated. Moreover, the direction of network optimization avoids being affected by a large number of simple negative samples. This solves the problem of sample imbalance and does not need to calculate complex weight mapping, so the model can better capture the signal features.

3. Multi-Scale Feature Fusion Model Design MS-ResNet

Bone sticks are old, the carriers are made of cow bones, different cow bones have different textures, the background color varies, the engraving strength varies, the bone stick stains and wears, decay damage is more serious, the bone stick writing style is variable, and there are cases of engraving in a "cursive" style where the traditional method can not be effectively recognized. Using deep learning methods in bone stick text recognition, the underlying feature map of the neural network contains some subtle features of the bone stick text, including texture, edge, angle, color, and other detailed features. In contrast, the higher-level feature map retains more important semantic information [10]. As the convolution continues to deepen the neural network, the bottom features have high resolution and contain detailed information but less semantic information and more noise. The top features have more important semantic information but less detailed information at the bottom of the resolution [22]. The feature maps extracted from the bone stick text after different depths of convolution layers are shown in Figure 2.

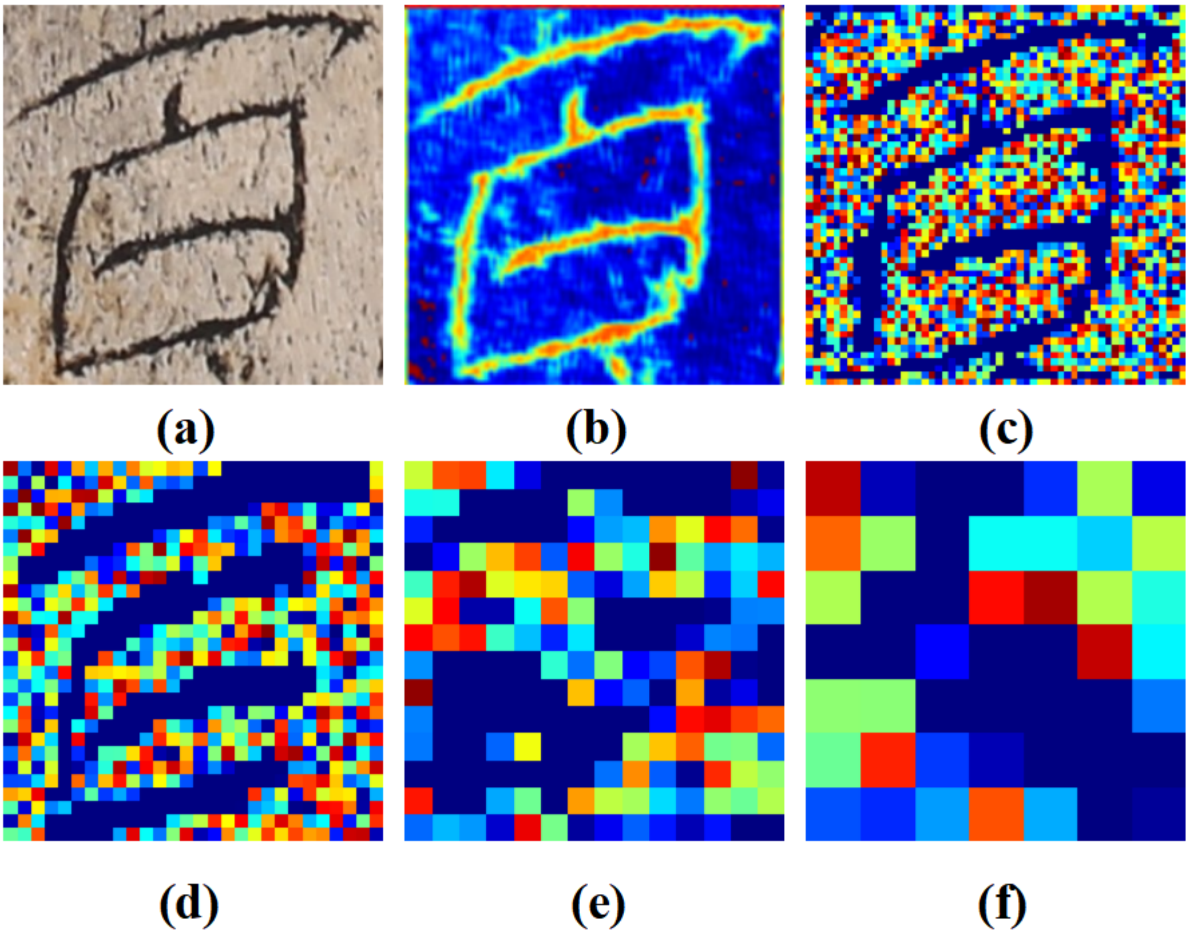

Figure 2. Characteristic display of each layer of bone stick: (**a**) original bone stick, (**b**) after 1 layer of convolution, (**c**) after 7 layers of convolution, (**d**) after 19 layers of convolution, (**e**) after 27 layers of convolution, (**f**) after 33 layers of convolution.

In bone stick text recognition, as the network deepens, the perceptual field gradually increases, the image information is compressed, the image detail information in the high-level feature map is continuously weakened, and text classification using only high-level features is too singular, so the detail perception is poor. Using a multi-scale model can effectively combine the advantages of both features and take into account the bone stick text details and semantic features, which is the key to improving the model feature extraction and thus enhancing the text recognition accuracy. In order to effectively extract different scale features, this paper designs a MS-ResNet model with multi-scale fusion. It introduces a multi-scale fusion strategy based on the ResNet34 neural network to fuse different levels of image features and enhance the text's shallow detail features. In order to avoid the bottom detail features extracted from the features being covered by the higher-level features, resulting in the loss of detail features, channel stitching is used to fuse the feature maps of different scales, which not only retains the deep features but also adds the bottom features of different depths to obtain a multi-scale feature representation with richer information. Channel splicing means that two feature vectors are added in the channel dimension, the new vector retains both sets of vector features, and the channel splicing process is schematically shown in Figure 3.

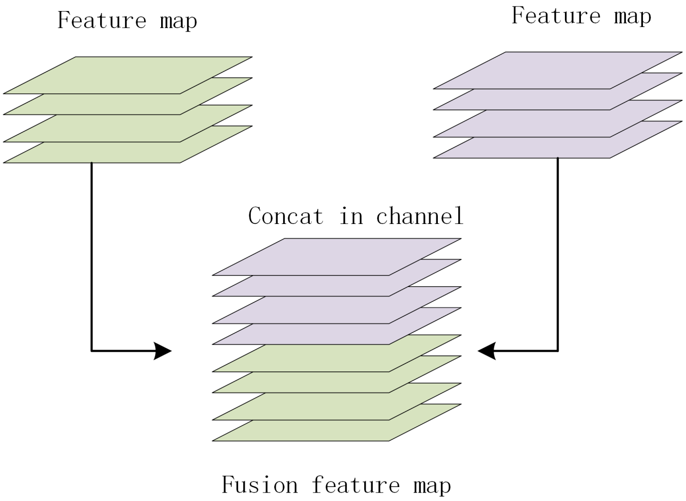

Figure 3. Illustration of the channel stitching process.

The multi-scale feature fusion method for bone stick text designed in this paper, the first layer 7×7 convolutional Conv1 features of ResNet, and the last layer output features Conv2_x, Conv3_x, Conv4_x, and Conv5_x in each residual block are selected. The above five feature maps are pooled on average and stitched in the channel dimension, and the feature fusion module can be expressed as Equation (5)

$$F = C(P(F_1) \bullet P(F_2) \bullet P(F_3) \bullet P(F_4) \bullet P(F_5)) \tag{5}$$

where F_1 is the Conv1 feature; F_2, F_3, F_4, and F_5 denote the last residual output feature of the Conv2_x, Conv3_x, Conv4_x, and Conv5_x layers, respectively; P denotes the global average pooling operation; and $C(\bullet)$ denotes the splicing operation along the channel dimension.

ResNet first layer convolution of Conv1 features and the last residual output features in each Conv2_x, Conv3_x, Conv4_x, and Conv5_x layer yield Scale1, Scale2, Scale3, Scale4, and Scale5 with dimensions of (112,112,64), (56,56, 64), (28,28,128), (14,14,256), (7,7,512), respectively. The above five feature maps are pooled by averaging to obtain features with sizes of (1,1,64), (1,1,64), (1,1,128), (1,1,256), and (1,1,512), respectively. The five-layer feature vectors are stitched in the channel dimension to obtain (1,1,1024)-dimensional fused feature vectors. The fusion process is shown in Figure 4.

After multi-scale fusion, the Conv5_x features are retained, and the Conv1, Conv2_x, Conv3_x, and Conv4_x layer features are incorporated using the channel splicing method. The number of the original 512 classification features is increased to 1024, which increases the underlying detail of the bone stick text information. Finally, it uses the obtained fused features for fully connected layer classification.

The structure of MS-ResNet is shown in Figure 5. The MS-ResNet model mainly consists of 34 layers of the network. The first layer uses 7 × 7 convolution with a step size of 2 and a channel number of 64, followed by a maximum pooling operation to reduce the image size by half to obtain the feature map S1. Then, we enter the Conv2_x, Conv3_x, Conv4_x, and Conv5_x layers with channel numbers 64, 128, 256, 512, and 512, respectively. Each Conv_x has 3, 4, 6, and 3 residual blocks, respectively, and each residual block is stacked with two 3 × 3 convolutions to increase the network depth. The last residual convolution of each Conv_x yields S2, S3, S4, and S5, respectively, and then these five feature maps are pooled equally to obtain vectors with 1 × 1 × (number of channels) features. These five one-dimensional feature vectors are fused and stitched together using the channel stitching method to obtain new features, increasing the number of features to 1024 from the original 512 features, and finally, the fully connected layer is used.

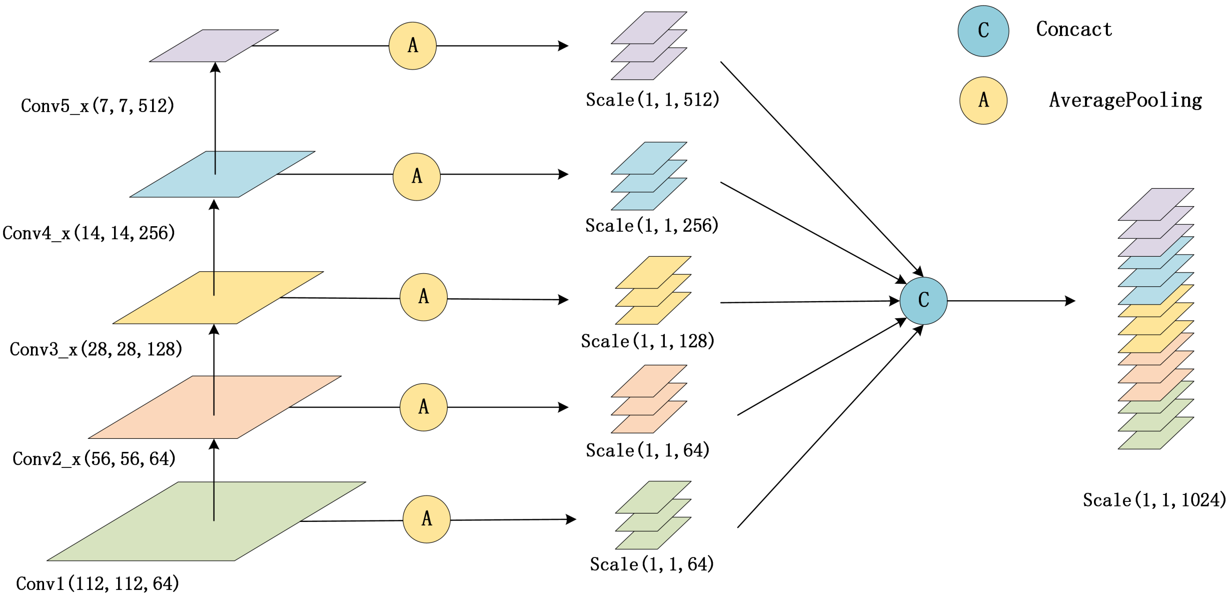

Figure 4. Schematic of the feature fusion process.

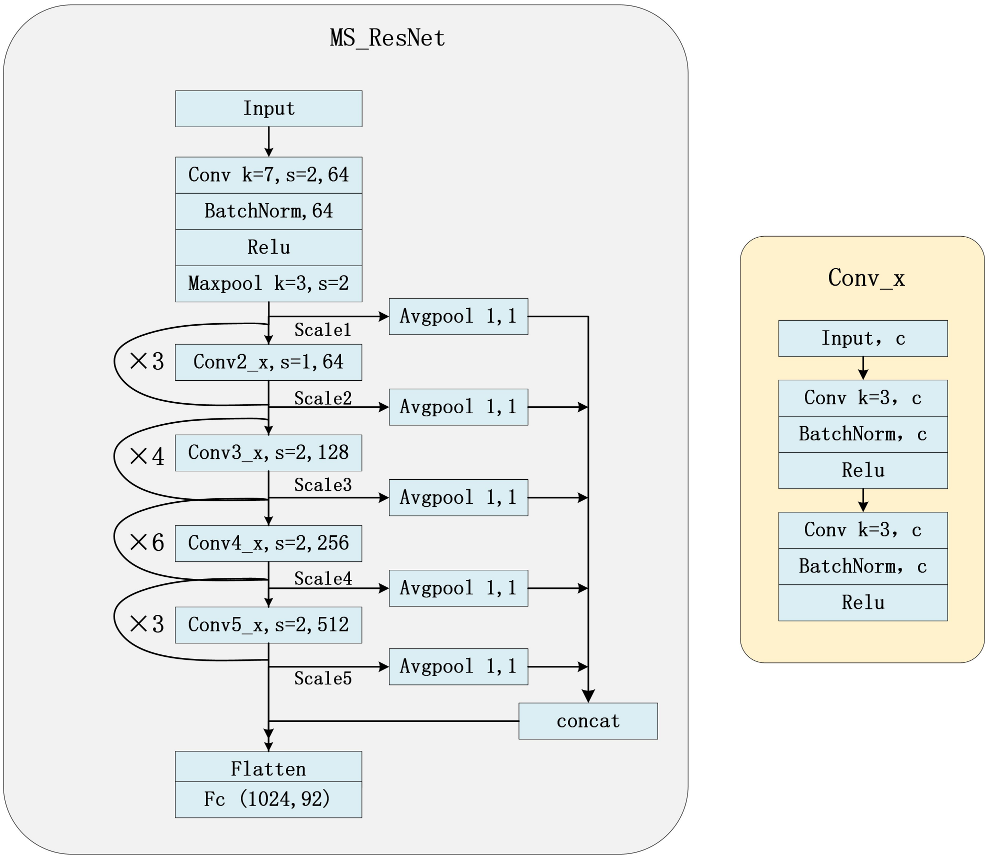

Figure 5. MS-ResNet structure.

4. Experimental Analysis and Discussion

4.1. Introduction of Experimental Data

The source of experimental data in this paper is a data set of pictures of bone stick text provided by the Institute of Archaeology of the Chinese Academy of Sciences, and a total of 2000 bone stick texts in two volumes were selected as experimental data. The pictures are whole bone sticks, and it is necessary to use LabelImg software to frame the bone stick text and input the corresponding word against the bone stick interpretation provided by the expert to realize bone stick text data labeling, and the data labeling process is shown in Figure 6.

After labeling the text data of the bone stick, an XML file was generated, the XML file was parsed using python, and the labeled text was cropped and sorted into different folders. From this, a bone stick text data set was established in which some text samples were insufficient. This experiment selected bone stick text with a sample number of more than 10 as the experimental object, which created a data set that had 92 types of bone stick text and a total of 7312 images. In the face of the complex bone stick text situation, denoising the pictures will destroy the original morphological structure of the bone stick text. Therefore, this paper uses the original pictures for text training recognition without uniform denoising and binarizing of the bone stick text pictures. The bone stick text is cropped from the bone stick image, and its image size varies. In order to improve the recognition accuracy of the algorithm, a common image size normalization method, bilinear interpolation, is used to normalize the image size to 224 × 224 pixels. In order to expand the training sample, avoid overfitting the neural network, and obtain better recognition results, data enhancement methods were used to expand the bone stick text images before model training. In this paper, we used image flip, rotation, contrast enhancement, and brightness enhancement methods to expand the number of images to five times the original number of images and establish a unified bone stick text database with 36,560 images. Some of the bone stick text images are shown in Figure 7.

In the experimental training phase, the training set, validation set, and test set were divided into 8:1:1. In order to expand the training sample, avoid overfitting the neural network, and obtain better recognition results, data augmentation methods were used to expand the data of the bone stick text images before model training. In this paper, we mainly considered brightness, contrast, flip, and rotation, a method that has no glyph changes and can retain the original image information. The training data set was expanded by making a series of changes to the training images to produce similar but different training samples. With the ResNet model, the test set's accuracy was improved by 3% after experimental data enhancement. After data enhancement, the final number of images was expanded to five times the original number of images, totaling 36,560 bone stick text images in the training set. The graphs before and after data enhancement are shown in Figure 8.

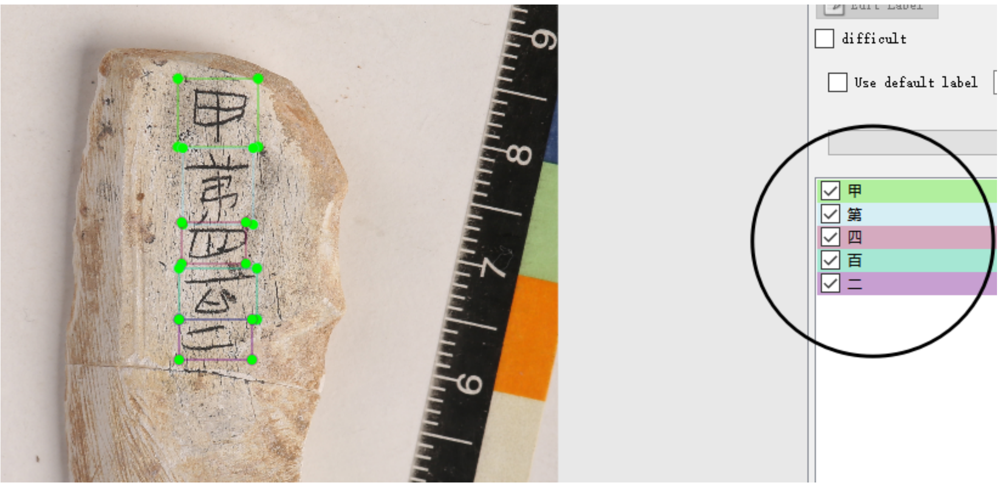

Figure 6. Data annotation display. (jia di si bai er).

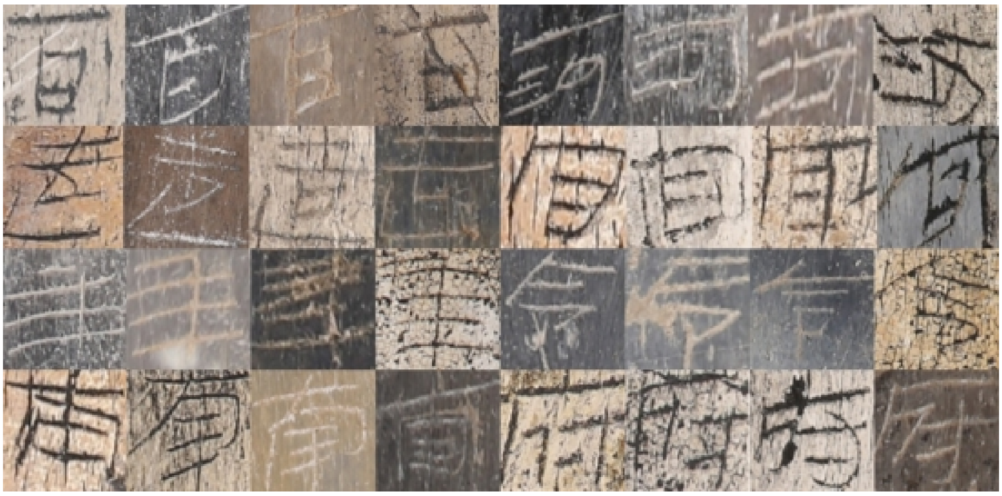

Figure 7. Partial images of the data set.

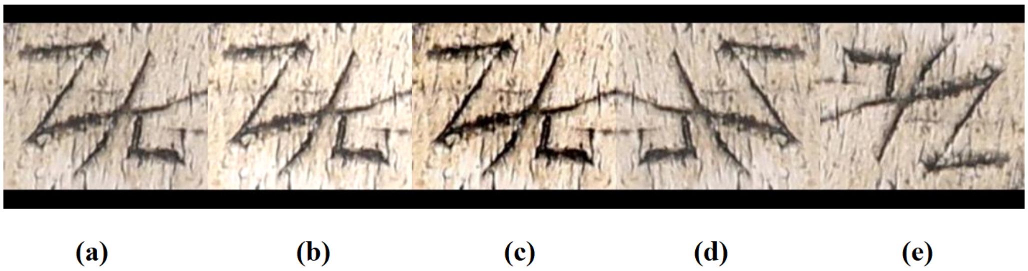

Figure 8. (**a**) original image (**b**) brighter (**c**) contrast enhance (**d**) flip (**e**) rotating.

4.2. Experiment Environment

This experiment used Windows, GeForce RTX 3090 GPU, a Pytorch framework environment, model training with a cross-entropy loss function and Adam optimizer, a ReLu hidden layer activation function, a batch size of 32, an initial learning rate of 0.001, a cosine learning rate decay method, and the number of model training iterations was 100.

During the training of the neural network text recognition model, excessive recognition will degrade the neural network recognition performance. In order to explore the best bone stick text recognition model, the training weights of each iteration of the model were kept during the training process, the test set was tested, and the best experimental results in 100 iterations were selected for comparison. In order to quantitatively evaluate the performance of each model, the experiments used recall, precision, f1, accuracy (Acc), and fixed test set loss values as evaluation indexes to comprehensively evaluate the recognition and classification results of different algorithms.

4.3. Comparison of Different Fusions of Different Feature Layers with Each Other

In order to verify the effectiveness of the fusion effect using five-layer features, this paper designed 10 groups of fusion methods and compared the experiments using the bottom-layer features alone for recognition. Various feature layers fused with each other, S1, S2, S3, S4, and S5, respectively, denote the first layer of the MS-ResNet features and the five layers of the last output layer of each residual from shallowest to deepest. The bottom layer features S5 and the middle layers are fused with each other. For example, the first-, third-, and fifth-layer features are fused as S1+S3+S5. The following experiments were conducted to fuse different feature layers on bone stick text recognition, and their various effects are shown in Table 1.

As can be seen from the table, S5 represents the accuracy of recognition originally using only single-layer features, which is only 88.6%, and the classification results of fusing different underlying features are all higher than 88.6%. Multi-scale feature fusion improves the accuracy of bone stick text recognition to different degrees for different layer fusion methods compared with single high-level feature classification, indicating that each intermediate feature layer contributes to bone stick text recognition. Therefore, the effectiveness of the fusion of different underlying features to enhance the fine-grained features of text and improve the accuracy of bone stick text recognition is verified, and the superiority of the multi-scale fusion approach of this paper's model is validated.

Table 1. Classification results of different features and their integration and comparison.

Model	Recall/%	Precision/%	F1/%	Acc/%
S5	79.3	88.2	83.5	88.6
S1+S5	80.9	88.7	84.6	89.1
S2+S5	80.2	86.2	83.1	89.2
S3+S5	81.6	88.1	84.7	89.6
S4+S5	81.9	88.0	84.9	88.6
S2+S3+S4	80.0	87.4	83.5	88.2
S1+S3+S5	81.5	89.9	85.6	89.4
S1+S2+S5	81.8	87.7	84.7	89.5
S2+S3+S5	83.0	88.3	85.6	89.9
S2+S3+S4+S5	82.9	88.0	84.8	89.8
MS_ResNet	83.4	88.9	86.0	90.3

In Figure 9 below, there are 23 samples for "Man" and 23 samples for "Shuo", which are similar in number and shape. The original ResNet is weak in extracting details, and "Man" is identified as "Shuo". There are 41 samples for "fruit" and 70 samples for "public opinion", which have a similar number of samples and a similar morphology. The original network model incorrectly identifies "Guo" as "Yu". The MS-ResNet model with multi-scale feature fusion allows the features to better capture text details and shallow contours and reduces the recognition of misspellings with similar morphological features.

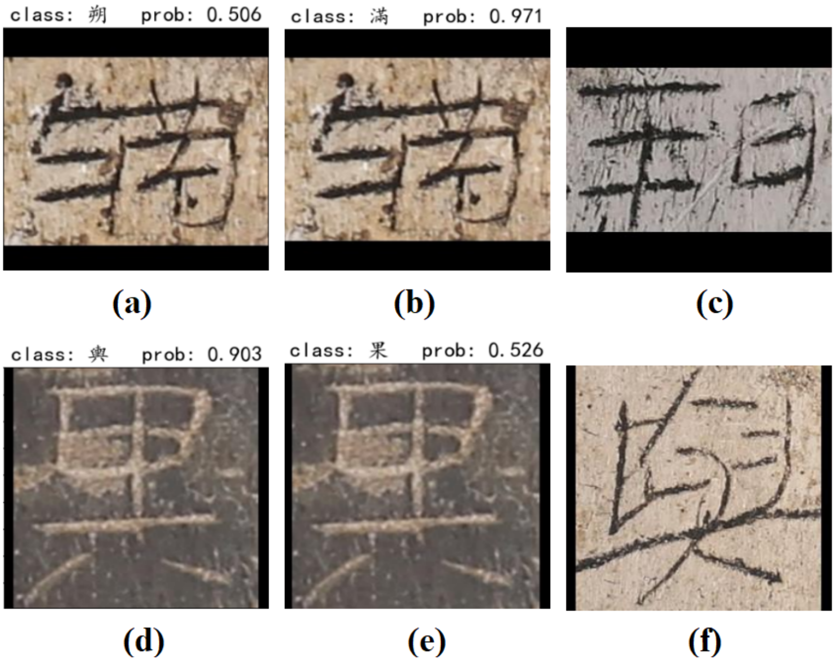

Figure 9. Validation of the MS-ResNet model (**a**) ResNet incorrectly identifies "Man" as "Shuo"; (**b**) MS-ResNet correctly identifies "Man"; (**c**) the image of "Shuo" is shown, and its stroke morphological features are similar to those of "Man"; (**d**) ResNet incorrectly identifies "Guo" as "Yu"; (**e**) MS-ResNet correctly identifies "Guo"; (**f**) the image of "Yu" is shown, and its stroke morphological features are similar to those of "Guo".

4.4. Comparison Test with Classical Network Model Method

Each classical network has a wide range of applications in Chinese character recognition. In this paper, we compared the classical AlexNet, DenseNet121, GooLeNet, Vgg16, ResNet34 network models and the MS-ResNet model of a multi-scale feature fusion network. Among the classical networks, ResNet was the best in recognition performance, with a slightly longer running time than AlexNet but excellent accuracy performance. The improved MS-ResNet neural network model in this paper based on ResNet can effectively improve the recognition rate of bone stick text with an accuracy rate of 90.3%, which is higher than AlexNet, DenseNet121, GooLeNet, Vgg16, and ResNet34, with accuracy rates of 8.6%, 5.0%, 4.2%, 2.7%, 1.7%, respectively. At the same time, the algorithm in this paper also has a higher recall, accuracy, and F1 value, and the algorithm produced the lowest loss value.

This experiment shows that classical convolutional neural network algorithms have a lower recognition rate for bone stick text. The multi-scale feature fusion model enhances the underlying detail features of bone stick text, and the overall recognition effect on the data set of this paper is better than other neural network methods. The final recognition rate results of different methods on the data set are shown in Table 2 below.

Table 2. Recognition rate of bone stick text by different algorithms.

Model	Recall/%	Precision/%	F1/%	Acc/%	Loss
AlexNet	72.2	78.0	75.0	81.7	1.093
DenseNet	75.7	81.0	78.2	85.3	0.98
GooLeNet	76.1	83.0	79.4	86.1	1.025
Vgg16	79.5	82.8	81.1	87.6	1.257
ResNet	79.3	88.2	83.5	88.6	0.454

Figures 10 and 11 below show the test set's accuracy curve and the training set's loss value of the bone stick text dataset in the first 80 iterations of each algorithm. The figure shows that the MS-ResNet network model converges the fastest, reaches the highest accuracy rate at more than 30 rounds of training, and outperforms the other networks overall, achieving the highest accuracy rate of 90.3%. The data show that excessive network training while yielding smaller loss values, can lead to performance degradation due to the overtraining of the network. ResNet has the second-highest accuracy and convergence speed, AlexNet has the slowest convergence speed and the lowest recognition rate of 81%, and DenseNet has the fastest convergence speed. However, the recognition accuracy is not high, remaining at 84%, andthe recognition rate of Vgg16 is not much different from that of ResNet, but the model convergence speed is slower, and the training time is increased by three times.

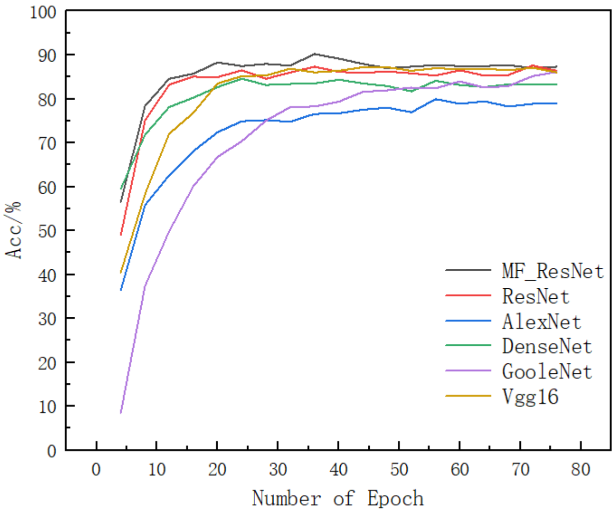

Figure 10. Change in accuracy value of bone stick recognition by different algorithms.

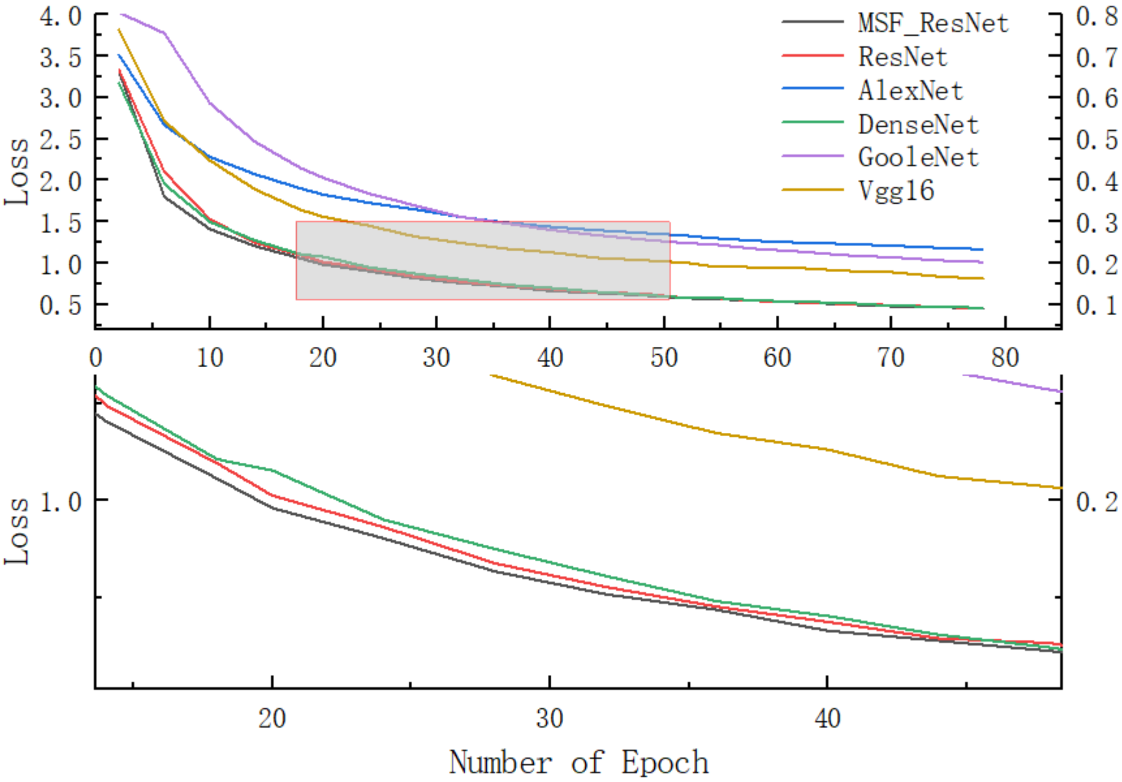

Figure 11. Change in loss value of bone stick recognition by different algorithms.

4.5. Experimental Validation of the Effectiveness of the Focal Loss Function

In order to show that the average performance of the methods in this paper is meaningful, the experiments are deliberately divided into different test subsets, with the number of text images in each category in the test set being three. The F1 value is used as the evaluation criterion for each method. It can be seen that the MS-ResNet network model has a higher F1 value and performs optimally in different test subsets. The results are shown in Table 3.

Table 3. Test results of different subsets.

Model	F1/%	Model	F1/%	Model	F1/%
AlexNet	76.8	S1+S5	82.0	S1+S3+S5	80.0
DenseNet	74.7	S2+S5	80.6	S1+S2+S5	81.2
GooLeNet	78.8	S3+S5	80.3	S2+S3+S5	84.5
Vgg16	81.6	S4+S5	83.8	S2+S3+S4+ S5	84.0
ResNet	82.4	S2+S3+S4	81.8	MS-ResNet	88.3

4.6. Experimental Validation of the Effectiveness of the Focal Loss Function

In order to verify the effectiveness of the improved multi-scale MS-ResNet model and the incorporation of the focal loss function for ablation experiments, the MSF-ResNet (multi-scale and focal loss ResNet) model combining the two achieved the highest recognition accuracy of 90.5%. Compared with the classical ResNet34 residual network model, the recall, precision, F1 score, and accuracy are improved by 4.9%, 1.1%, 3.2%, and 1.9%, respectively. Compared with the ResNet network model using only the focal loss function, the above indexes are improved by 3.4%, 1.0%, 2.4%, and 1.6%, respectively. Compared with the multi-scale network model without the focal loss function, the above indexes are improved by 0.8%, 0.4%, 0.7%, and 0.2%, respectively. The experimental results are shown in Table 4.

The focus loss function can effectively balance the recognition rate of small sample categories, such as in Figure 12, where the"Zhi" data sample size of 18 can be expanded to 90. However, for "Man" there are 166 training samples, and after the expansion there are 830. After adding the focal loss function, the category weight of small samples is increased, and the small samples that were incorrectly identified before can be correctly identified, strengthening the accuracy of small sample text recognition.

Table 4. Text recognition of bone sticks of different modules.

Model	Recall/%	Precision/%	F1/%	Acc/%	Loss
ResNet	79.3	88.2	83.5	88.6	0.454
ResNet+Focal	80.8	88.3	84.3	88.9	0.438
MS-ResNet	83.4	88.9	86.0	90.3	0.409
MSF-ResNet	84.2	89.3	86.7	90.5	0.385

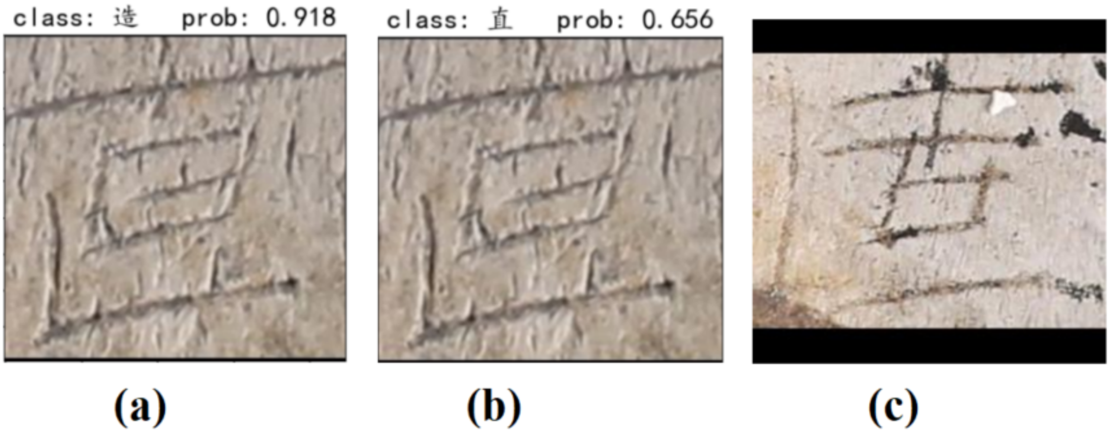

Figure 12. Experimental results using the FocalLoss loss function are shown: (**a**) without focal loss function, MS-ResNet incorrectly identifies "Zhi" as "Zao"; (**b**) with focal loss function, MS-ResNet correctly identifies "Zhi"; (**c**) the image of "Zao" is displayed.

The MSF-ResNet model with multi-scale feature fusion using focal loss function enhances the feature extraction ability while strengthening the ability to balance small samples, enabling better bone stick text model recognition applications.

4.7. Comparison Experiments with Other Literature Methods

In order to evaluate the classification performance of the improved MS-ResNet network for bone stick text, references [8,12,13] are used as the control group to compare the classification accuracy with the method of this paper. Their experimental results are shown in Table 5.

Table 5. Comparison with other literature methods for recognition accuracy.

Method	Recall%	Precision%	F1%	Acc%	Running Time Per Epoch
[13]	53.8	62.8	57.9	70.13%	48 s
[12]	80.7	87.3	83.8	88.2%	158 s
[8]	73.9	84.2	78.1	84.6%	53 s
Ours	83.4	88.9	86.0	90.5%	65 s

From the table, it can be seen that the improved multi-scale ResNet in the literature [13] has a low recognition accuracy on the bone stick data set and is not designed for the characteristics of the data set. The introduction of multi-scale also reduces the recognition accuracy of the bone stick text, and this method is only applicable to the classification of fewer categories. The ResNet with an improved network structure in the literature [12] does not improve the recognition effect of the bone stick text, there are too many convolutional layers, and the recognition time is slow.The long convolutional bars in the improved oracle recognition method in the literature [8] have no positive effect on bone stick text recognition and the network model is too simple. After the above comparison, the proposed multi-scale fusion bone stick text recognition method in this paper achieves good results in terms of accuracy and time effectiveness. The proposed method has more potential for applications in bone stick text recognition.

5. Conclusions

In this paper, a bone stick recognition model using a multi-scale feature fusion focal loss function is constructed based on the actual application of the bone stick text. To address the problems of insufficient extraction of underlying detail features, unbalanced distribution of data categories, and low recognition accuracy of existing methods in the recognition of Western Han bone sticks, this paper uses the ResNet neural network model as the basis, reduces the feature size dimension by global average pooling of different scale features, and uses the channel splicing method to fuse the features of each layer to enhance the extraction ability of underlying detail features of bone sticks. This paper further improves the recognition accuracy of the category with a small number of samples by using the focal loss function. The method in this paper can solve the problem of an unbalanced number of samples while improving the extraction ability of the network for bone stick features and then improving the recognition accuracy of the model. The experimental results show that the MSF-ResNet model proposed in this paper has a higher recognition accuracy of 90.5% on the self-built data set compared with other deep learning recognition methods and models. In future work, we will further expand the bone stick text data set and explore the effects of other multi-scale fusion methods on bone stick text recognition and image structure features.

Author Contributions: Conceptualization, M.D. and H.W.; methodology, M.D. and K.W.; software, M.D.; validation, M.D.; formal analysis, Z.W.; investigation, Z.W.; resources, Rui Liu; writing—review and editing, M.D. and K.W.; visualization, M.D.; supervision, H.W. and K.W.; funding acquisition, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Social Science Fund Special Project of China under grant number 20VJXT001.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data used to support the findings of this study are part of a private image data set.

Acknowledgments: Thanks to Zhe Yang for polishing this article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MS-ResNet	Multi-Scale ResNet
MSF-ResNet	Multi-Scale and Focal Loss ResNet

References

1. Qi, H. *Probe into the Archives of Bone Signet in Han Dynasty*; Lantai World: Shenyang, China, 2014; pp. 58–59.
2. Zhang, G. Annotated Bone stick of Chang 'an City in Han Dynasty. Master's Thesis, Capital Normal University, Beijing, China, 2012.
3. Lu, L. *A Preliminary Study on Bone Signage Excavated from Weiyang Palace in Chang 'an City of Han Dynasty*; Northwestern University: Xi'an, China; 2013; pp. 2–3. .

4. Wu, Z.; Liu, L.; Zhang, Z. Chinese Character recognition based on integrated Attention layer convolutional neural network. *Comput. Technol. Dev.* **2018**, *28*, 4.
5. Liu G. Deep learning based Oracle bone character detection and recognition. *Yindu J.* **2020**, *41*, 6.
6. Liu Y; Liu G. Oracle bone character recognition based on SVM. *J. Anyang Norm. Univ.* **2017**, *3*, 54–56.
7. Liu, M.; Liu, G.; Liu, Y.; Jiao, Q. Oracle bone inscriptions recognition based on deep convolutional neural network. *J. Image Graph.* **2020**, *8*, 114–119. [CrossRef]
8. Wang H. Oracle Bone Script Detection and Recognition Based on Deep Learning. Master's Thesis, South China University of Technology, Guangzhou, China, 2018.
9. Li, W.; Cao, B.; Cao, C.; Huang, Y. A Deep learning-based Method for Bronze Inscriptions recognition. *Acta Autom. Sin.* **2018**, *44*, 8.
10. Ru, X.; Hua, G.; Li, L.; Li, L. Handwritten digit recognition Based on Deformable convolutional Neural network. *Microelectron. Comput.* **2019**, *36*, 5.
11. Luo, Y.; Bi, X.; Wu, L.; Li, X. Dongba hieroglyphic recognition based on improved residual learning. *J. Intell. Syst.* **2022**, *17*, 79–87.
12. Wang, C.; Zhou, J.; Wu, H.; Teng, G.; Zhao, C.; Li, J. Improved Multi-scale ResNet for vegetable leaf disease recognition. *Trans. Csae* **2020**, *36*, 9.
13. Kuang, B.; Chen, Y.; Su, B. Detecting for Bronze Inscriptions. In Proceedings of the 2020 4th International Conference on Electronic Information Technology and Computer Engineering, Xiamen, China, 6–8 November 2020; pp. 555–559.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
15. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
16. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
17. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 July 2016; pp. 770–778.
19. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
20. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
21. Chockler, H.; Farchi, E.; Godlin, B.; Novikov, S. Cross-entropy based testing. In Proceedings of the Formal Methods in Computer Aided Design (FMCAD'07), Austin TX, USA, 11–14 November 2007; pp. 101–108.
22. Pan, T.-S.; Huang, H.-C.; Lee, J.-C.; Chen, C.-H. Multi-scale ResNet for real-time underwater object detection. *Signal Image Video Process.* **2021**, *15*, 941–949. [CrossRef]