applied sciences

MDPI

Article

Research on Rejoining Bone Stick Fragment Images: A Method Based on Multi-Scale Feature Fusion Siamese Network Guided by Edge Contour

Jingjing He [1], Huiqin Wang [1,*], Rui Liu [2], Li Mao [1], Ke Wang [1], Zhan Wang [3] and Ting Wang [1]

[1] School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China; jingjinghe@xauat.edu.cn (J.H.)
[2] Institute of Archaeology, Chinese Academy of Sciences, Beijing 100101, China
[3] Shaanxi Institute for the Preservation of Cultural Heritage, Xi'an 710075, China
* Correspondence: hqwang@xauat.edu.cn

Abstract: The rejoining of bone sticks holds significant importance in studying the historical and cultural aspects of the Han Dynasty. Currently, the rejoining work of bone inscriptions heavily relies on manual efforts by experts, demanding a considerable amount of time and energy. This paper introduces a multi-scale feature fusion Siamese network guided by edge contour (MFS-GC) model. Constructing a Siamese network framework, it first uses a residual network to extract features of bone sticks, which is followed by computing the L2 distance for similarity measurement. During the extraction of feature vectors using the residual network, the BN layer tends to lose contour detail information, resulting in less conspicuous feature extraction, especially along fractured edges. To address this issue, the Spatially Adaptive DEnormalization (SPADE) model is employed to guide the normalization of contour images of bone sticks. This ensures that the network can learn multi-scale boundary contour features at each layer. Finally, the extracted multi-scale fused features undergo similarity measurement for local matching of bone stick fragment images. Additionally, a Conjugable Bone Stick Dataset (CBSD) is constructed. In the experimental validation phase, the MFS-GC algorithm is compared with classical similarity calculation methods in terms of precision, recall, and miss detection rate. The experiments demonstrate that the MFS-GC algorithm achieves an average accuracy of 95.5% in the Top-15 on the CBSD. The findings of this research can contribute to solving the rejoining issues of bone sticks.

Keywords: bone stick rejoining; Siamese network; edge contour guidance; multi-scale feature fusion; similarity metrics

Citation: He, J.; Wang, H.; Liu, R.; Mao, L.; Wang, K.; Wang, Z.; Wang, T. Research on Rejoining Bone Stick Fragment Images: A Method Based on Multi-Scale Feature Fusion Siamese Network Guided by Edge Contour. *Appl. Sci.* **2024**, *14*, 717. https://doi.org/10.3390/app14020717

Academic Editors: Silvia Liberata Ullo and Li Zhang

Received: 16 November 2023
Revised: 5 January 2024
Accepted: 12 January 2024
Published: 15 January 2024

1. Introduction

A bone stick [1] refers to elongated bone pieces of approximately uniform size and fixed shape, which are processed from animal skeletons. The bone stick discovered [2] at Weiyang Palace in Han Dynasty's Chang'an City provides valuable historical data for studying Western Han history. It also preserves the written language used during that period, making it a crucial source material for studies in paleography, linguistics, and related fields. Excavated bone sticks often suffer damage and fractures due to prolonged burial underground, requiring rejoining for preservation and comprehensive analysis. Bone stick rejoining can be accomplished through two methods [3]: manual and computer-assisted with expert participation. In the manual method, experts use contextual semantics, joinable characters, literary materials, and knowledge of bone script morphology to speculate on potential rejoining outcomes. However, due to human resource limitations, manual rejoining is relatively inefficient. Experts in bone stick technology have consistently held high expectations for the use of computer technology in achieving the rejoining of bone sticks [4].

The field of image rejoining has two main categories of methods: local edge matching and deep learning. In the local edge-matching method, the techniques outlined in references [5–8] impose a high requirement for edge consistency in fragmented images, making them suitable for rejoining paper document edges. Among deep learning methods, References [9,10] describe methods that use neural networks to predict the relative positions of archaeological fragments, solving 3 × 3 jigsaw puzzles. However, these approaches are limited to square-shaped, equally sized two-dimensional fragments and are not suitable for irregular bone stick fragments. Le et al. [11] proposed the JigsawNet model, which uses a Convolutional Neural Network to detect the compatibility between image fragments. However, the network has limitations in detecting compatibility between defective cultural relic fragments because it is a polygonal approximation of the boundary contours. In their study, Ngo et al. [12] proposed a combination of residual networks and a pool of spatial pyramidal tandem networks to match images of exhumed wood fragments. However, its effectiveness is limited when the fragments suffer from edge erosion or fracture lines or when there are missing fragments. In such cases, the texture and text style of the fracture region provide more accurate high-level information. Zhang et al. [13,14] proposed a deep rejoining model for automatically rejoining oracle bone fragment images. The edge equidistant reconnection method is utilized to match and locate the edges of the two fragmented images. It then crops the target region image and evaluates its texture similarity using a convolutional neural network. However, this method has high network complexity and requires accurate edge contour information.

In this paper, a Multi-scale Feature Fusion Siamese network Guided by Edge Contour (MFS-GC) model is proposed. The limitation of previous methods using single feature information is addressed. Two improvements have been made to the Siamese network framework in this study: Firstly, when utilizing a residual network for feature extraction, the Batch Normalization (BN) layer tends to lose fine contour details, leading to an indistinct feature extraction effect on fractured edges. To address this issue, we incorporate the SPADE model to guide the normalization of bone stick edge contour images. This allows the network to learn multi-scale edge contour features at each stage. Secondly, as the network depth increases, information loss becomes a challenge. To mitigate this, we fused features at different scales and determined whether a match existed by comparing the L2 distances of the fused feature vectors. The proposed network focuses on capturing high-level semantic information in bone stick images while also considering texture and contour details. The advantage of this method lies in its improved preservation of detailed information, providing the network with richer data and enhancing the accuracy of rejoining bone sticks. To evaluate the effectiveness of the algorithm, we conduct a comprehensive experimental analysis comparing our algorithm with recent deep learning-based similarity matching algorithms. The evaluation is performed on the CBSD dataset, considering metrics such as accuracy, leakage rate, precision, etc. The results indicate a Top-15 accuracy rate of 95.5%, validating the superiority of our proposed algorithm.

2. Basic Methods

2.1. Siamese Network

In the context of comparing two images, Bromley [15] introduced the Siamese network (SN) in 1993 to authenticate whether the signature on a check corresponds to the signature stored by the bank. The Siamese network framework is illustrated in Figure 1. It involves obtaining two feature vectors, $h(x_{1\_1}, x_{1\_2})$ and $h(x_{2\_1}, x_{2\_2})$, by feeding two preprocessed images, X_1 and X_2, through a shared-weight feature extraction network. The L2 (Euclidean) distance between these two vectors is then calculated to derive the similarity value. During this process, the inputs are individually mapped by two neural networks into a new space, resulting in a representation of the inputs in that space. The similarity between the two inputs is assessed through the computation of the contrastive loss. In this study, applied to bone stick images, matched bone stick images are considered to exhibit the highest similarity, labeled as 1, representing positive samples. Conversely, non-matched

bone stick images are regarded as having the lowest similarity, labeled as 0, indicating negative samples. The training objective is to minimize the distance between matched bone specimens and maximize the distance between non-matched images.

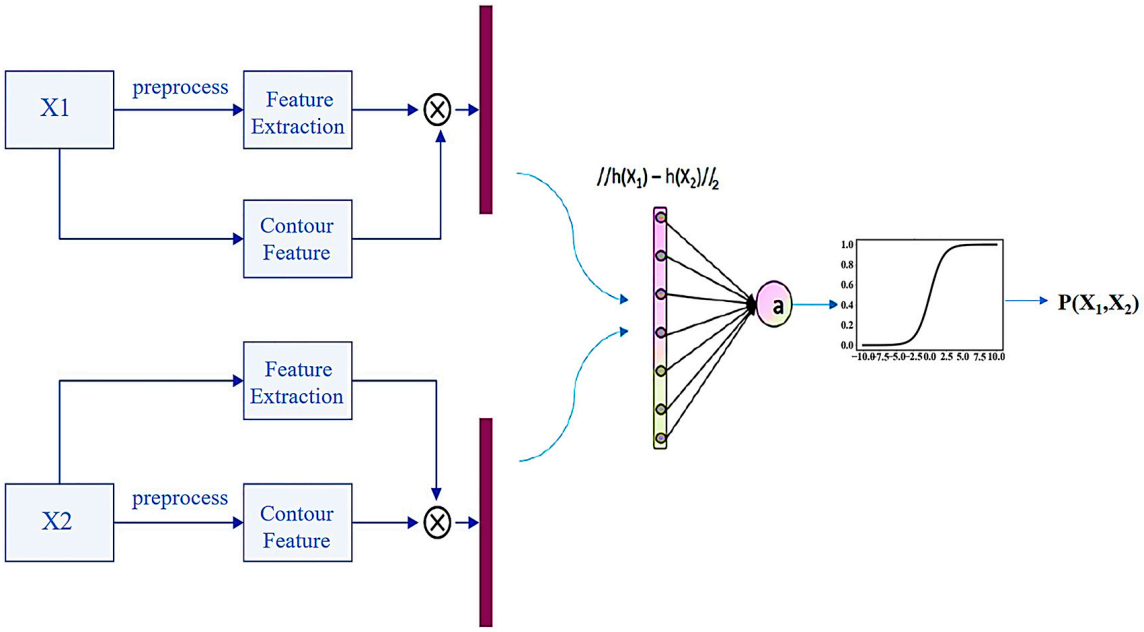

Figure 1. Siamese network. "a" refers to the similarity value between input images X1 and X2.

Meanwhile, this paper proposes the use of a Siamese network equipped with a Contrastive Loss function. This type of loss function effectively captures the degree of matching to the sample and is well-suited for training models in feature extraction. The definition is shown in Equation (1).

$$L = \frac{1}{2N}\sum_{n=1}^{N} yd^2 + (1-y)max(m-d,0)^2 \quad (1)$$

$$d(x_1, x_2) = \|x_1 - x_2\|_2 = \sqrt{\sum_{i=1}^{P}(x_{1i} - x_{2i})^2} \quad (2)$$

In this context, x_1 and x_2 represent two samples, while y denotes the label for the input sample pairs. A value of $y = 1$ indicates a match between the two samples, and $y = 0$ signifies a non-match. The threshold m is used to determine whether two input samples belong to the same category. We only consider distance values of dissimilar feature Euclidean distances within the range of 0 to m. N represents the number of samples. Equation (2) defines the Euclidean distance between the features x_1 and x_2 of two samples. The Siamese network aims to minimize the distance between sample pairs when they belong to the same category and increase the distance between sample pairs otherwise. Therefore, the contrastive loss function is employed to achieve the objectives of the Siamese network.

The utilization of Siamese networks primarily addresses two key challenges related to the rejoining of bone sticks [16]. Firstly, it tackles the issue of limited training data within the same category. The training dataset utilized in this study predominantly comprises images of bone sticks. Specifically, each upper segment of a bone stick corresponds to a lower image segment. Secondly, when new category data emerge, there is no necessity for retraining; the model can perform predictions directly.

2.2. Resnet Network Model

Theoretically, increasing the number of network layers allows the network to perform more complex feature pattern extraction, potentially leading to better performance with deeper models. However, practical experimentation has revealed that simply adding convolutional layers to the network not only fails to reduce the training error but also results in its escalation [17]. This phenomenon primarily arises due to the challenge of vanishing or exploding gradients in deep networks, where an increase in the number of layers exacerbates the training difficulty. Hence, the concept of skip connections was introduced in residual networks to address the issue of vanishing or exploding gradients caused by the stacking of numerous network layers. The formulation of the residual structure is depicted in Equation (3).

$$x_{l+1} = x_l + F(x_l, W_l) \tag{3}$$

Here, $F(x)$ represents the residual term. For any arbitrarily deep cell X_{l+1}, its features can be expressed as the features of a shallower cell X_l plus a residual function in the form of $F(x)$.

Figure 2 illustrates the structure of residual blocks. The application of residual blocks enables the deeper units to easily acquire results from shallower units. This method of endowing results from shallower networks to deeper networks is referred to as identity mapping. As depicted in Figure 2a, it can ameliorate the issue of feature loss during the learning process. Within convolutional neural networks, when the output X_l of the preceding layer does not align in dimensionality with the input feature map X_{l+1} of the subsequent layer, a 1×1 convolution is necessary to perform dimensionality augmentation or reduction. This scenario is exemplified by the residual block depicted in Figure 2b.

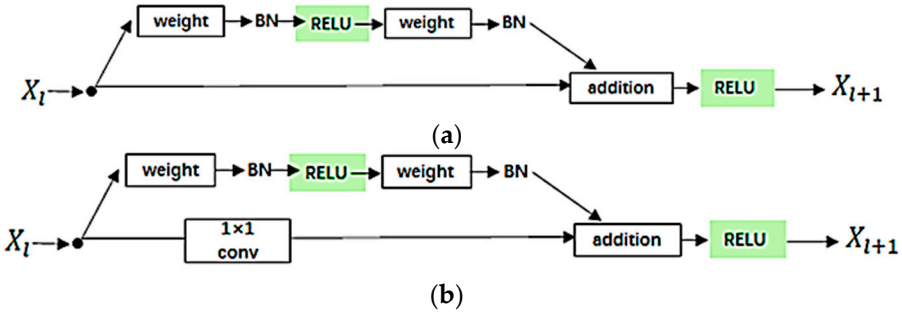

Figure 2. Residual block: (**a**) identity block; (**b**) conv block.

3. Multi-Scale Feature Fusion Siamese Network Guided by Edge Contour Model Design

3.1. Overall Model Design

Figure 3 illustrates the overall design of the multi-scale feature fusion guided by edge contour model (MFS-GC). The local regions of conjugable bone stick images exhibit a symmetrical structure in terms of color, texture, contours, and other information, thereby possessing similar core features. The approach presented in this paper utilizes a similarity measurement to assess the matching degree of bone sticks. A Siamese network framework is used for feature extraction, which is followed by a subsequent measurement of the extracted features' similarity.

In the image preprocessing stage, we normalize the dimensions of the input original bone stick images, x_1 and x_2, to 224×224 pixels, resulting in x_{1_1} and x_{2_1}. Subsequently, the images undergo grayscale conversion, binarization, and edge detection algorithms, yielding the corresponding edge contour images, x_{1_2} and x_{2_2}. As the size of the input contour images differs from the feature maps, x, at each stage, it becomes necessary to resize the contour images to match the size of the feature maps. This process generates multi-scale contour images.

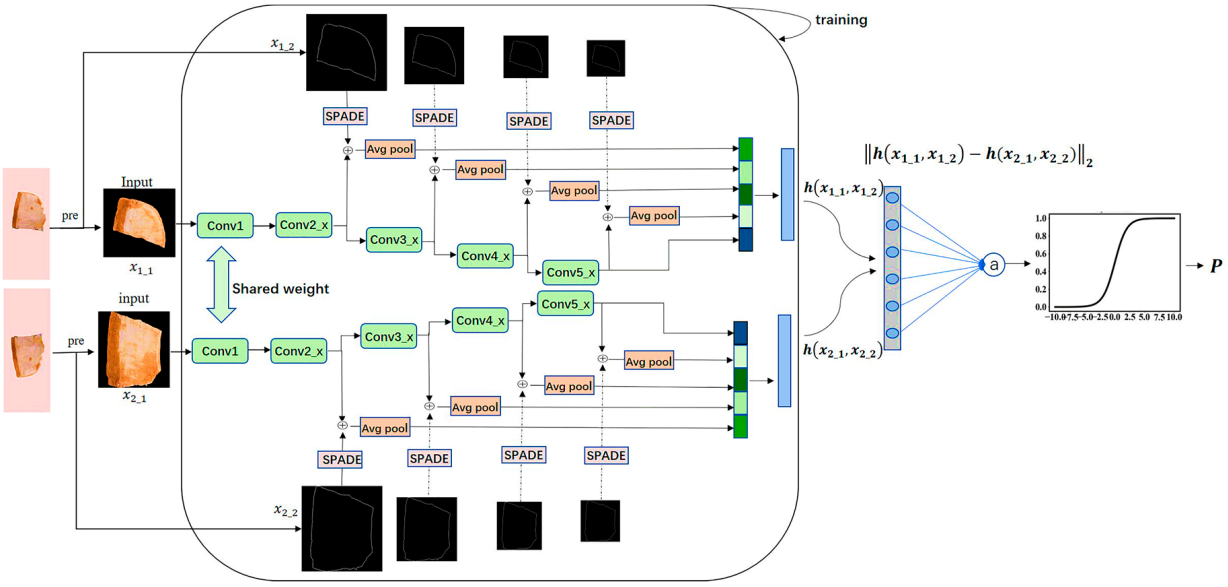

Figure 3. MFS-GC model framework. In the figure, from top to bottom, the colors represent green, light green, dark green, light blue, and blue, indicating the fused features R1, R2, R3, and R4 obtained through the SPADE model in the conv2_x to conv5_x stages of the backbone network, and the deep features extracted from the conv5_x stage; "a" refers to the similarity value between input images X1 and X2.

During the training phase, preprocessed bone stick images are input into a feature fusion residual network with two shared parameters for training. The contrastive loss function for the two bone stick images is computed, and network parameters are iteratively updated through backpropagation. The objective is to minimize the L2 distance between feature vectors of positively pairable samples and maximize the L2 distance between feature vectors of non-pairable bone stick samples, thereby achieving discrimination between the two.

During the inference phase, the bone stick image to be appended is input, along with each bone stick image from the dataset, into the trained network to extract feature vectors, and L2 distances are calculated. Similarity scores for each pair are sorted, and the top-T ranked bone stick fragment images are selected to form the candidate set for conjugable bone sticks.

3.2. SPADE Model

Unlike panoramic image stitching, adjacent bone stick fragments only possess matching geometric shapes and textures along the fractured boundaries. Addressing the issue of losing lower-level details in the backbone feature extraction network caused by the use of BN normalization layers, this paper introduces the SPADE module to incorporate feature information from contour maps. The feature information from contour maps is utilized to guide the normalization of feature maps at various hierarchical levels.

The Spatially Adaptive DEnormalization (SPADE) model [18] is a model designed to prevent BN from truncating semantic information of input images, and its basic framework is illustrated in Figure 4.

The implementation process of the SPADE model [19] can be summarized as modifying the computation of γ and β based on Batch Normalization. Specifically, convolution is utilized to learn the γ and β of feature maps, which are then used as the normalization coefficients and biases, respectively, acting on the previous-level feature maps after the normalization layer following convolution. As illustrated in Figure 2a, assuming the input to the SPADE module comprises the previous-level feature map represented as x_{in}, the

input contour map as l_s, and the output feature map as x_{out}, the formulation can be expressed as shown below:

$$l\_s' = RELU(conv(Resize(l\_s))) \tag{4}$$

$$x_{out} = bn(x_{in}) \cdot conv_{\gamma}(l\_s') + conv_{\beta}(l\_s') \tag{5}$$

Due to the discrepancy in size between the input contour map l_s and the feature map x, l_s is resized to match the size of x. Subsequently, it undergoes a convolutional layer followed by ReLU activation, resulting in an intermediate layer denoted as l_s'. Following this, two distinct convolutional layers are applied to obtain the corresponding γ and β parameters. These derived parameters are then, respectively, used as normalization coefficients and biases, influencing $bn(x_{in})$.

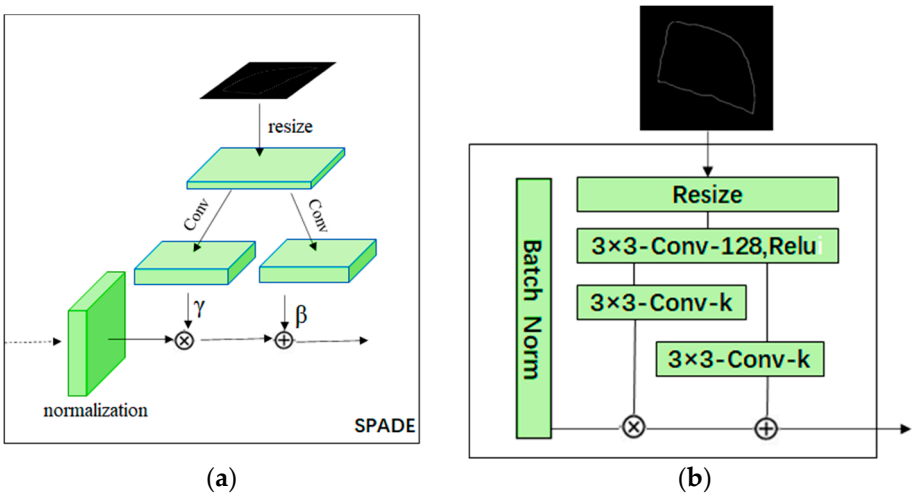

Figure 4. Basic framework of the SPADE model: (**a**) schematic diagram of the SPADE model; (**b**) internal structure of the SPADE model. The direction of the arrows represents the direction of data transfer.

In the SPADE model, γ and β are three-dimensional matrices that have width and height dimensions in addition to the channel dimensions. Therefore, in Equation (6), the subscripts for γ and β include symbols for c, y, and x. This embodies the concept of "spatial adaptation," indicating differences and adaptability in spatial dimensions. This is distinct from the BN layer [20,21], where the parameter vectors γ and β only have the subscript c (representing channels). Clearly, the computation method of BN, which does not distinguish spatial dimensions, is prone to losing information from the input image. The calculation of the mean μ and standard deviation σ in Equation (6) is illustrated in Equations (7) and (8).

$$\gamma^i_{c,y,x}(m)\frac{h^i_{n,c,y,x} - \mu^i_c}{\sigma^i_c} + \beta^i_{c,y,x}(m) \tag{6}$$

Among them:

$$\mu^i_c = \frac{1}{NH^iW^i}\sum_{n,y,x} h^i_{n,c,y,x} \tag{7}$$

$$\sigma^i_c = \sqrt{\frac{1}{NH^iW^i}\sum_{n,y,x}\left(h^i_{n,c,y,x}\right)^2 - \left(\mu^i_c\right)^2} \tag{8}$$

The advantage of the SPADE model is that it better preserves semantic information against public normalization layers. The semantic map focuses differently on different regions of the input image; in this paper, for the edge contour binary image, the contour region is 1 and the other regions are 0. Therefore, the SPADE module is added to allow the network to focus on the contour region features. This is accomplished in such a way that

the positional information (x, y) of the mask can be taken into account. So, SPADE can be seen as a general definition of other condition normalization.

3.3. Edge Contour-Guided Feature Fusion Residual Network

The MFS-GC model designed in this paper introduces the SPADE module into the backbone feature extraction network at the feature extraction stage, and then the output feature maps of the four stages are used as inputs to the SPADE model, which uses the edge contour binary image information of the bone stick image to guide the feature maps of the Resnet50 neural network to be normalized, and it intervenes by continually adding edge contour binary images so that each stage of the Resnet50 network is able to learn the edge contour features. Finally, the fused features Ri output from the four stages are obtained.

The MFS-GC model proposed in this study includes the SPADE module in the core feature extraction network during the feature extraction phase. The output feature maps generated from the four stages are then used as inputs to the SPADE module. This module uses information from the edge contour of the bone stick image to guide the normalization of the feature maps within the Resnet50 neural network. Each stage of the Resnet50 network is able to learn the edge contour features This results in the fused features, denoted as Ri, being outputted from the four stages.

Inspired by the Feature Pyramid Networks (FPNs) [22–24], this study introduces a multi-scale approach by reusing feature maps during the feature extraction phase. This method effectively integrates deep-level features with shallow-level features. The channel concatenation method, as illustrated in Figure 5, involves adding two feature vectors along the channel dimension. Deep neural networks often neglect low-level information while extracting high-level semantic features. However, in the domain of bone stick rejoining, both the low-level and high-level information of two-dimensional damaged bone stick images are equally significant. Hence, outputs from four convolutional layers of different scales collectively contribute to the representation of the feature vectors. This approach overcomes the limitations of conventional convolutional neural networks in extracting texture and contour feature information, thereby enhancing the model's capability to extract detailed features.

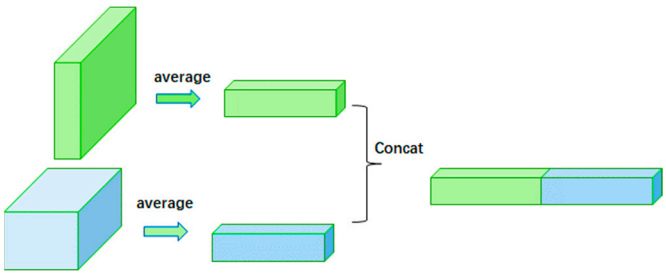

Figure 5. Multi-scale feature channel stitching.

This paper proposes the construction of two contour-guided feature fusion residual networks with shared weights. The feature extraction network of a single branch is illustrated in Figure 6 with Resnet50 serving as the backbone of the network. The network consists of 49 convolutional layers and 1 fully connected layer. The overall structure of Resnet50 is divided into five stages: Stage 0, Stage 1, Stage 2, Stage 3, and Stage 4. Among them, the Stage 0 structure is simple and is considered as the preprocessing stage for input images. The subsequent four stages all include residual network units, where each residual unit is composed of three convolutional layers. The unit first employs a 1×1 convolution to achieve dimension reduction; then, it utilizes a 3×3 convolution to extract features channel-wise and finally employs a 1×1 convolution to achieve dimension expansion.

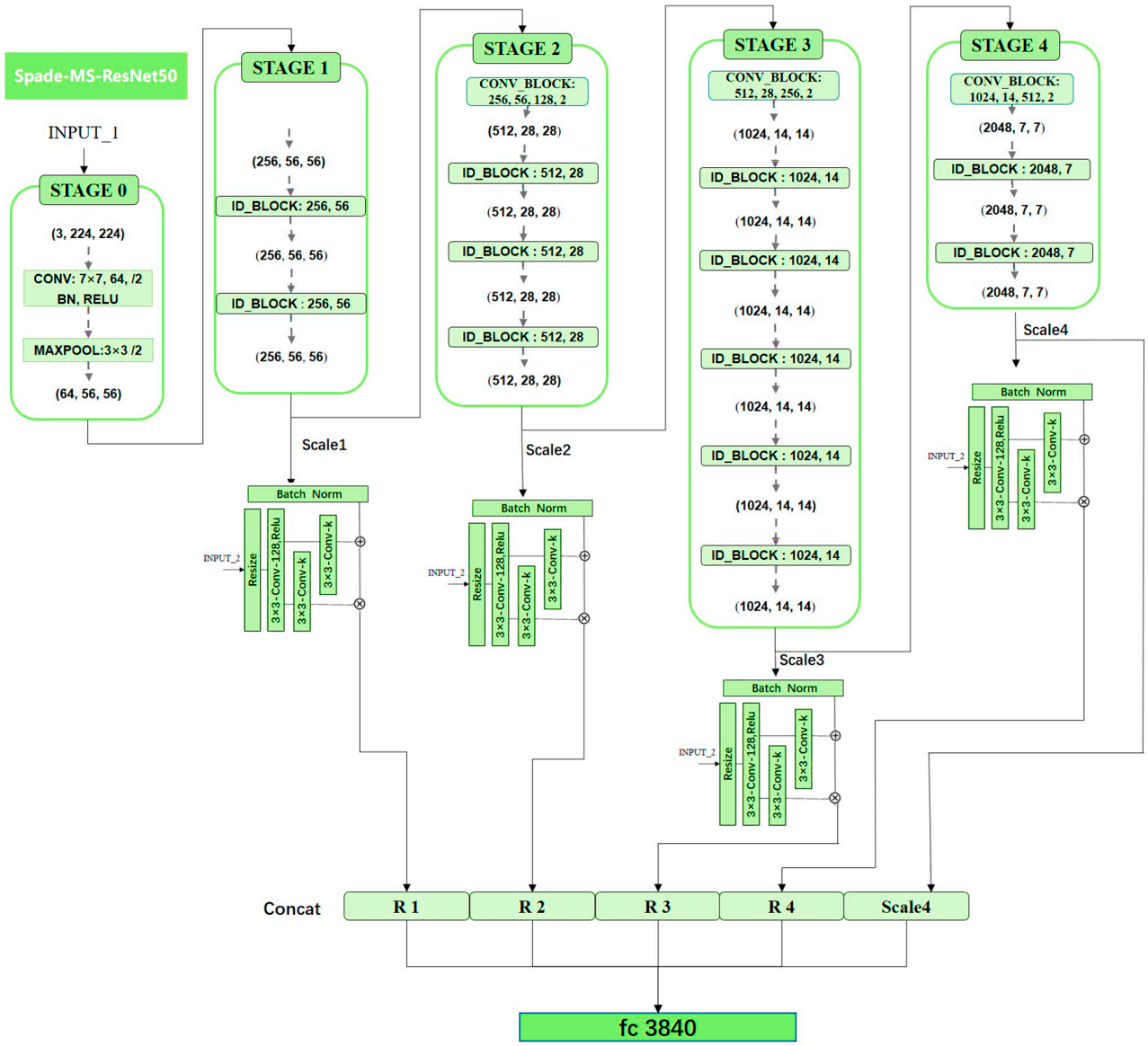

Figure 6. Edge contour-guided feature fusion Resnet network.

The entire network begins with Stage 0, where the input image has a shape of (224, 224, 3). It undergoes convolutional layers, Batch Normalization (BN) layers, Rectified Linear Unit (RELU) activation functions, and a max-pooling layer, resulting in an output shape of (56, 56, 64). In Stage 1, the output from the previous stage, (56, 56, 64), serves as input to obtain the output scale 1 with dimensions (56, 56, 256). Subsequently, the SPADE model takes scale 1 as input 1 and the contour map of the bone marker as input 2. It learns detailed contour features of the bone marker. The contour map is first resized and passed through a convolutional layer, which is followed by two 3 × 3 convolutions. The outputs of these two convolutional layers are then subjected to element-wise multiplication and addition with scale 1, yielding the fused output result R1 with a feature size of (56, 56, 256). In Stage 2, the backbone network consists of four bottleneck residual blocks, resulting in an output feature map scale 2 of dimensions (28, 28, 512). The feature map of the bone marker is then combined with scale 2 to produce the fused feature R2 with dimensions (28, 28, 512). Stage 3 comprises six bottleneck residual blocks, generating the feature map scale 3, which, after SPADE model processing, yields the fused feature R3 with dimensions (14, 14, 1024). In Stage 4, with three bottleneck residual blocks, outputs scale 4 with dimensions (7, 7, 2048) and the fused feature R4 has the same dimensions (7, 7, 2048).

On the basis of edge contour guided Resnet50, a multi-scale fusion strategy is introduced to enhance the shallow detail features of bone marker images. While preserving the deep feature R4, additional fusion features of different scales are incorporated with dimensions being (56, 56, 256), (28, 28, 512), (14, 14, 1024), and (7, 7, 2048), respectively. The aforementioned four feature maps undergo adaptive average pooling layers to obtain features with dimensions (1, 1, 256), (1, 1, 512), (1, 1, 1024), and (1, 1, 2048). Subsequently, these five layers of feature vectors are concatenated along the channel dimension, yielding a fused feature with dimensions (1, 1, 3840). This fused feature is then processed through a fully connected layer (fc) for output with the incorporation of a Dropout layer in the fully connected layer to prevent overfitting.

4. Experimental Analysis

In order to verify the effectiveness and reliability of this paper's algorithm in the field of bone stick rejoining, this paper carries out comparative experiments of different algorithms on the constructed bone stick dataset.

4.1. Dataset Production

The data used in this paper comes from the bone stick images excavated from Weiyang Palace in Chang'an City of the Han Dynasty, provided by the Institute of Archaeology, Chinese Academy of Social Sciences. A total of 935 pairs (1870 images) of bone stick images recognized by the experts as rejoining and 800 images of bone stick images that are not verified by the experts are selected as the experimental data. In the experimental training phase, the dataset CBSD is divided into three parts: the training set, validation set, and test set. The data are divided as shown in Table 1 below. The test set is named CBSD_T. Simultaneously, to test the reliability of the model, 800 original bone stick images not verified by experts are added to the test set, which are named CBSD_I.

Table 1. Dataset segmentation.

CBSD	Training Set	Validation Set	Test Set	
			CBSD_T	CBSD_I
1870	1496	174	200	200 + 800

Due to the non-uniform size of bone stick images in the dataset, it is necessary to standardize the images to the same dimensions to improve the matching accuracy of the algorithm. The image size is normalized to 224 × 224 pixels using the bilinear interpolation algorithm. Simultaneously, to increase the training samples, enhance the model's generalization ability, and prevent neural network overfitting for better matching results, this paper employs image flipping and rotation data augmentation methods on bone stick images before model training. The number of images is expanded to twice the original quantity.

Each pair of rejoined bone stick images is divided into upper and lower slices. The naming convention is illustrated in Figure 7. For instance, for a bone stick image numbered 27356, it is divided into upper slice 27356_01 and lower slice 27356_02. The fractured local area of the rejoined bone stick image exhibits a symmetrical structure in terms of color, texture, contour, and other information, thereby possessing similar core features. Therefore, images of the local areas of the bone stick are selected as the dataset, and they are named 27356_01_01 and 27356_02_01; subsequently, the local area images undergo grayscale transformation, binarization, and boundary tracking algorithms to obtain their corresponding edge contour images, named 27356_01_02 and 27356_02_02.

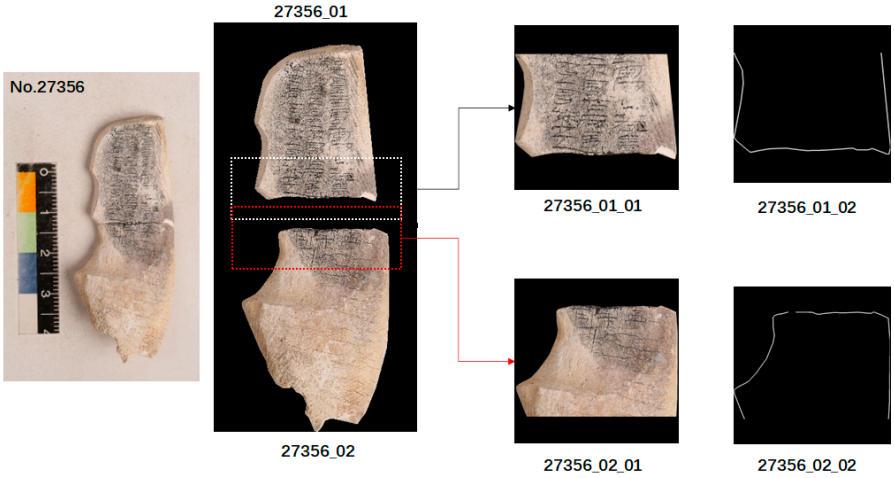

Figure 7. Dataset production In the figure, 27356_01_01 and 27356_02_01 denote the acquired local region images of bone sticks, while 27356_01_02 and 27356_02_02 represent the corresponding edge contour images.

4.2. Experimental Configuration and Evaluation Metrics

The experiment was conducted using the Windows 10 operating system with hardware comprising a GeForce RTX 3090 GPU and an AMD Ryzen 9 5900X 12-Core Processor. Programming was performed using the Python language and the PyTorch deep learning framework. The model training employed the Adam optimization algorithm, setting the initial learning rate (lr) to 0.001. The model's loss function adopted the contrastive loss function with the RELU function serving as the activation function for the hidden layers. The batch size was set to 20, and the model underwent a total of 30 epochs during training. To ensure the optimal bone stick rejoining model, the training retained the training weights of the model at each iteration, subjected the test set to evaluation, and selected the experiment's best-performing iteration for comparative analysis.

In order to maximize the feature extraction capability of the backbone feature extraction network, under the selection of Resnet50 as the base network, we use the idea of transfer learning to use the initial weights pre-trained in the Image Net dataset and then further train them in the bone stick dataset for fine-tuning.

Bone stick rejoining is defined as follows: given a broken bone stick image as input, the algorithm returns the top-T bone stick images with the highest match to this image as a candidate rejoining set. This set should include bone stick images that can genuinely be rejoined with the input image among the T images. The rejoining is considered satisfactory if this condition is met. In the constructed CBSD dataset, the true rejoined objects for each image are known, allowing the algorithm to automatically calculate its Top-T accuracy without the involvement of experts.

However, in real-world scenarios of bone fragment rejoining, since the true rejoining objects of a given bone stick image are not known in advance, bone stick experts are required to use their professional knowledge to filter and confirm the candidate rejoining results. If the value of T is large, there will be too many candidate results, and reviewing the matching results of each bone stick fragment image will consume a significant amount of time for the bone stick experts. Additionally, due to the large number of bone stick images, it becomes challenging for bone stick scholars to complete the filtering and confirmation of candidate rejoining results for all bone stick fragment images. In order to effectively alleviate the workload of bone stick experts in real-world bone fragment adhesion scenarios, and based on the recommendations of bone stick experts, the value of T is set to 15.

Therefore, the evaluation metrics are defined as follows:

Top-*T* rejoining accuracy (*ACC*) is the ratio of the number of images that can correctly find the concatenation result to the number of all images that actually have rejoining objects

in the case of returning the Top-T candidate images with the highest rejoining match for each image. Then, the rejoining accuracy ACC can be expressed as Equation (9).

$$ACC = \frac{TP}{M} \tag{9}$$

where TP denotes the number of correctly predicted images, and M represents the total number of images.

$$\text{Missed detection rate} = 1 - \text{Accuracyrate};$$

Precision refers to the number of correct positive identifications divided by the sum of the rankings of the true positive identifications among the Top-T candidate results. If two algorithms find the same number of correct positive identifications, their Top-T accuracy and missed detection rate are also the same. In such cases, if a method's correct positive identifications are consistently ranked higher among the Top-T results, its precision value is higher. If a method's correct positive identifications are consistently ranked first (Top-1) among the Top-T results, its precision metric is 100%. Precision can be expressed as Equation (10).

$$PRE = \frac{TP}{\sum_{i=1}^{N} Rank_i} \ (i = 1, 2, \ldots N) \tag{10}$$

$Rank_i$ refers to the position of the true positive identification of the correct rejoining image among T candidate results, and N represents the number of correctly predicted images.

In conclusion, high Top-T accuracy and high precision index values are the desired effects of the bone stick rejoining algorithm.

4.3. Experimental Results and Analysis

4.3.1. Comparative Experiments with Classical Feature Extraction Networks

In the backbone feature extraction stage, Table 2 illustrates the performance evaluation of classical network models, including Vgg16, Resnet34, Resnet18, and DenseNet121, on the CBSD training image dataset. The MFS-GC model achieves an accuracy of 95.5% in the Top-15, surpassing the performance of other models. Furthermore, the training loss of our proposed model is 0.024, which is lower compared to other models. Among the six models considered, our model demonstrates state-of-the-art performance.

Table 2. Comparison of the performance of different backbone networks.

Model	Accuracy	Precision	Missed Detection	Loss
Vgg16	0.750	0.192	0.25	1.456
Resnet34	0.890	0.301	0.11	0.369
Resnet18	0.83	0.225	0.17	0.384
DenseNet121	0.735	0.175	0.264	0.91
MFS-GC	0.955	0.356	0.045	0.24

The experiments indicate that the MFS-GC model introduced in this paper enhances the underlying fine-grained features of bone stick images, yielding overall results superior to alternative methods on the dataset employed in this study. Figures 8 and 9 illustrate the training loss and accuracy curves for each algorithm. From the graphs, it is evident that all models tend to converge with the network model in this paper achieving the highest accuracy of 95.5% after 30 training epochs, which is accompanied by a training loss of 0.24. This performance is, in general, superior to that of other networks. DenseNet121 exhibits the slowest convergence speed with an accuracy as low as 73.5%, while the Vgg16 model converges rapidly but achieves a relatively lower accuracy of only 75%. Resnet18 and Resnet34 demonstrate accuracies of 83% and 89%, respectively, indicating that for the bone stick image dataset in this study, the performance of residual networks surpasses that of other networks. Additionally, a comparison among different residual networks reveals

that the performance of the MFS-GC model, utilizing the Resnet50 backbone network, outperforms Resnet18 and Resnet34.

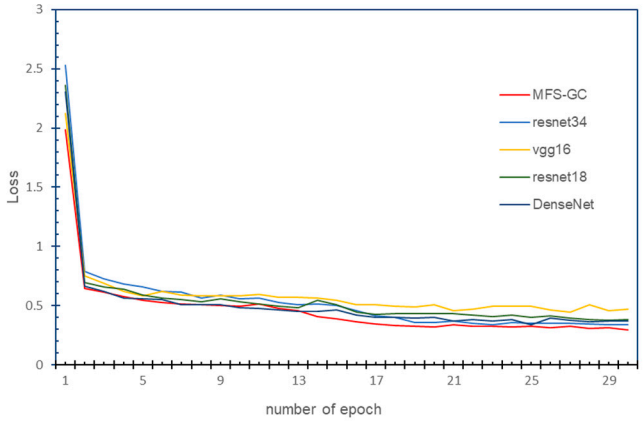

Figure 8. Training set loss values.

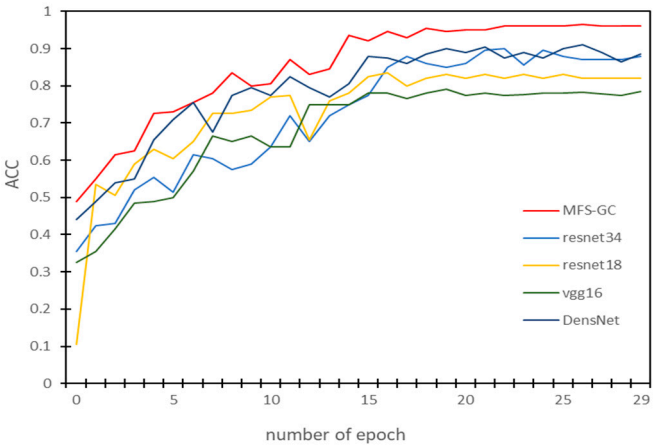

Figure 9. Accuracy curves.

4.3.2. Complexity Analysis

To objectively measure the complexity of the proposed model, the parameters (Params) and floating-point operations (FLOPs) were computed in this study [25–27]. Additionally, a comparative analysis was conducted with other CNN network models for reference. The GFLOPs were calculated based on a 224 × 224 input scale, and the experimental results are presented in Table 3.

Table 3. Comparison of different model.

Model	Number of Parameters	Floating-Point Operations (FLOPs)
VGG16	138 M	32.2
Densenet121	8.1 M	5.9
MFS-GC	27.6 M	4.2

Generally, a model's computational complexity is directly proportional to its parameter count and FLOPs. Larger models typically entail more parameters and computational requirements. Choosing an appropriate model depends on the specific application requirements and the availability of computational resources.

The data from Table 3 reveal that the proposed methodology in this study has a parameter count of 27.6 million and a computation workload, measured in FLOPs (Floating Point Operations), of 4.2 G. In comparison, the VGG16 network possesses a parameter count

of 138 million and a FLOP value of 32.2 G. Hence, it is asserted that the model presented in this paper demonstrates superior performance. Subsequently, a comparative analysis between the MFS-GC model and the DenseNet121 network reveals that while DenseNet121 has a relatively lower parameter count, it exhibits higher FLOPs. Consequently, after carefully weighing computational resources against model performance, the MFS-GC model was selected due to its provision of enhanced feature representation. Notably, this model achieves a significant reduction in parameter count and computational workload while concurrently improving performance.

4.3.3. Different Test Subsets

Considering that the authentic application scenario of bone stick rejoining involves unearthed damaged fragments of bone stick, the rejoining relationships are thus unknown. To assess the model's high feasibility in addressing the task of bone stick rejoining, we conducted experiments and constructed the dataset CBSD_I, comprising 800 authentic interfered bone inscription fragments and 200 pairs of bone stick fragments with known concatenation relationships. The evaluation metric used is rejoining accuracy. Figure 10 shows the rejoining capability of the MFS-GC model on the COBD_T and COBD_I test datasets. The vertical axis represents accuracy, ranging from 0 to 1, while the horizontal axis indicates the judgment requirement Top-T for rejoining accuracy. The figure illustrates rejoining accuracy from Top-1 to Top-20. It can be observed from the graph that the accuracy of bone stick fragments with an added authentic interference has experienced a certain decrease compared to CBSD_T. However, overall, the model proposed in this paper exhibits good rejoining capability on both test sets.

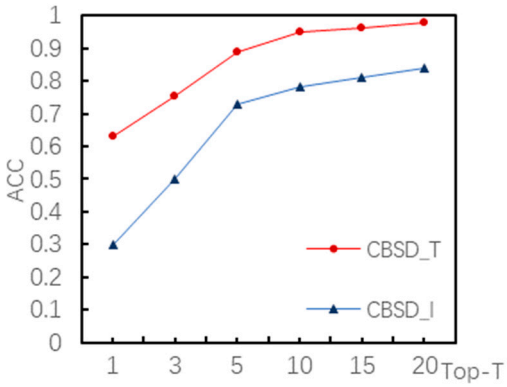

Figure 10. Accuracy under different test sets.

4.3.4. Comparison with Other Literature Methods

To assess the reliability of the model proposed in this paper, comparative experiments were conducted with the models presented in references [12,28,29]. As indicated in Table 4, reference [28] employs a fundamental Siamese network framework, utilizing Resnet34 as the backbone feature extraction network to address the issue of predicting radiance in sky images. However, it lacks the incorporation of multi-scale contour guidance, resulting in a simplistic network model and consequently lower accuracy. Reference [29] utilizes an enhanced Siamese depth feature fusion method, leading to improved rejoining effects on bone stick images. Nevertheless, this method merely overlays convolutional features of different scales without specifically focusing on edge information, leaving room for enhancement in the context of bone stick images. In contrast, reference [12] combines residual networks with spatial pyramid pooling, employing a two-dimensional Siamese neural network for matching excavated wooden pieces and obtaining a matching probability. While achieving a high accuracy rate of 89.8%, it predominantly emphasizes texture features, leading to decreased performance when dealing with complex bone stick images. Through comparison, the MFS-GC model proposed in this paper achieves an accuracy of 95.5%. This model demonstrates superior applicability in the domain of bone stick rejoining.

Table 4. Comparison of different literature algorithms.

	Accuracy (%)	Precision	Time (s)
Literature [28]	76.4	0.16	44.86
Literature [29]	83.6	0.224	23.7
Literature [12]	89.8	0.289	42.62
MFS-GC	95.5	0.356	34.6

Table 5 presents the comparative performance of the MFSGC algorithm designed in this study and algorithms from references [12,28,29] on the CBSD dataset, which are evaluated based on the Top-T accuracy metric. From the table, it can be observed that the MFS-GC algorithm achieves a significant improvement of nearly 11% in the Top-1 accuracy metric compared to the algorithm in reference [12]. In terms of Top-5, 10, and 15 accuracy metrics, the MFS-GC algorithm demonstrates noticeable performance enhancements. Overall, the MFS-GC model devised in this study achieves the best composite performance, particularly with a Top-15 recall rate reaching 95.5%, highlighting the superiority of this algorithm.

Table 5. Comparison of accuracy at different Top-T (%).

Top-T	Literature [28]	Literature [29]	Literature [12]	MFS-GC
Top-1	31.5	38.0	53.5	64.0
Top-5	48.5	55.5	69.0	75.5
Top-10	69.0	72.5	80.5	89.0
Top-15	76.4	83.6	89.8	95.5

4.3.5. Ablation Experiments

1. Comparison of accuracy under different layer feature fusion structures

To illustrate the role of multi-scale feature fusion, experiments were conducted on networks with different feature fusion structures. Table 6 presents the accuracy of the networks without feature fusion scales R1, R2, R3, and R4, respectively. During testing, the connections of the corresponding scales in the concat layer in Figure 6 were selectively disconnected.

Table 6. Comparison under different feature fusion structures.

Feature Fusion Architecture	Accuracy	Missed Detection	Time
Without R1	91.63	8.37	32.91
Without R2	87.46	12.54	27.31
Without R3	81.33	18.67	22.68
Without R4	74.02	25.98	19.34
Only scale4	69.96	30.04	10.66
MFS-GC	95.50	4.50	35.63

From Table 6, it can be observed that compared to models with full-scale feature fusion, the absence of any feature fusion structure leads to a certain degree of reduction in rejoining accuracy. This indicates that each intermediate feature layer contributes to the conjunction accuracy. Therefore, the superiority of the multi-scale fusion approach proposed in this paper has been validated.

2. Validation of SPADE to guide edge contours

To validate the effectiveness of SPADE guidance, we compared networks utilizing the SPADE model for contour guidance with the original backbone feature extraction network without the SPADE model. As shown in Table 7, compared to the network without the SPADE model, the proposed MFS-GC model in this study preserves the contour features of the bone stick fragment images, resulting in an improvement of 6.0% in rejoining accuracy.

Table 7. Comparative experiments with and without SPADE guidance.

SPADE	Resnet50 (Backbone)	Accuracy
×	√	Top-15: 0.895 Top-10: 0.745 Top-5: 0.615 Top-1: 0.38
√	√	Top-15: 0.955 Top-10: 0.89 Top-5: 0.755 Top-1: 0.64

4.3.6. Distance Metric Performance

In practical analysis, understanding the magnitude of differences between individuals is crucial for evaluating their similarities and categories. Similarity measurement is a method for assessing how data samples are interrelated or close to each other. Typically, this involves calculating the distance between the features of entities. Similarity measures are usually represented as numerical values, with higher values indicating greater similarity between data samples. These values are often transformed to a scale between 0 and 1, where 0 signifies low similarity (data objects are dissimilar), and 1 indicates high similarity (data objects are very similar). In other words, as the correlation between data objects strengthens, the distance decreases, and the similarity score increases. In this context, Euclidean distance is considered optimal and valuable for distance computation.

To clarify this issue, we take the example of four pairs of conjugate bone stick images from the test set, where the upper image A_01 corresponds to the lower image A_02, and their similarity scores are presented in Table 8. The closer the distance measurement, the higher the similarity score for bone stick images. In Table 8, the average similarity score among the four pairs of bone stick images is above 0.8, indicating a high correlation between the matching results of the bone stick images and the distance measurements.

Table 8. Distance and similarity scores of pairs of bone stick images.

Bone Stick Pair	Similarity Score	Distance
No. 02750_01 No. 02750_02	0.91	0.6697
No. 02816_01 No. 02816_02	0.81	0.9265
No. 03136_01 No. 03136_02	0.86	0.8246
No. 119_01 No. 119_02	0.88	0.80331

5. Bone Stick Rejoining Results

After obtaining a candidate set of bone stick images for rejoining, two bone stick images were identified by experts for rejoining through coordinate transformations. Based on the outcomes in the IMS-GC model, Figure 11 illustrates partial successful examples of rejoining results. As the corresponding damaged bone stick exhibits severe fractures and increasingly indistinct fracture features, the difficulty of conjunction progressively intensifies. Figure 12 presents a failed case of bone stick rejoining, which is attributed to wear and discoloration in the fractured regions of the bone stick fragments. Wear creates gaps between fragments, complicating the alignment of adjacent pieces, while discoloration generates false edges, broadening the scope of potential matches. Consequently, the Top-T rejoining results fail to encompass bone stick images that could genuinely rejoin, resulting in rejoining failure.

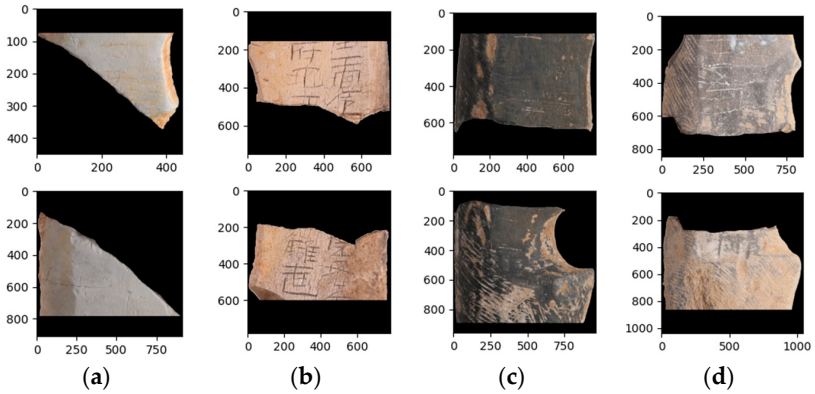

Figure 11. Display of correct rejoining results. (**a**) No. 02750, Similarity: 0.9105; (**b**) No. 119, Similarity: 0.8764; (**c**) No. 02816. Similarity: 0.8147; (**d**) No. 03136. Similarity: 0.8638.

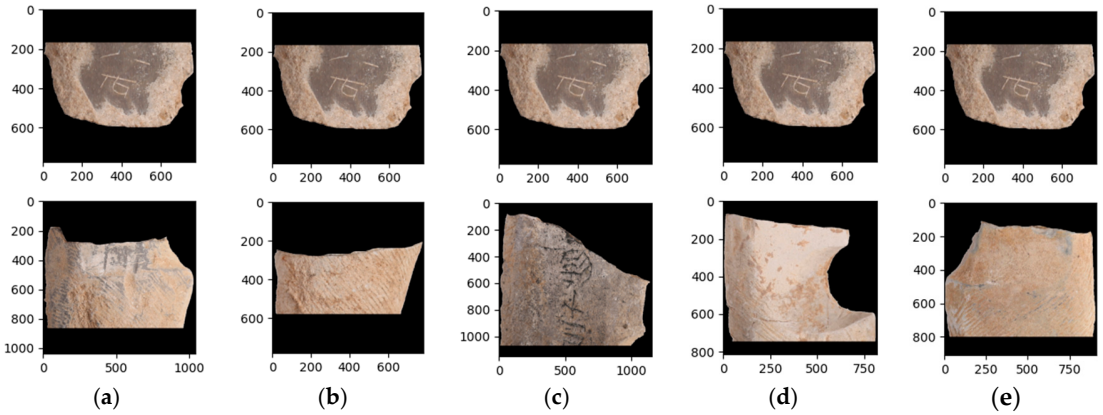

Figure 12. Display of the result of error rejoining as No. 25734. (**a**) Top-1, Similarity: 0.7962; (**b**) Top-3, Similarity: 0.7760; (**c**) Top-5, Similarity: 0.7670; (**d**) Top-10, Similarity: 0.7644; (**e**) Top-15, Similarity: 0.7539.

6. Conclusions

This paper proposes a multi-scale feature fusion Siamese network guided by edge contour (MFS-GC) model to address the problem of rejoining damaged bone stick fragments. To tackle the issue of the Batch Normalization (BN) layer losing low-level detailed features, the SPADE model intervenes in the residual network by incorporating bone stick edge contour images. This ensures that the network captures multi-scale edge contour features at each layer. The features extracted are then combined across various scales, and similarity measurement is performed using L2 distance to match bone stick fragment images locally. The experimental results demonstrate that the model achieves a rejoining accuracy of 95.5% on the test dataset. The algorithm described in this paper converts the task of reassembling damaged bone stick fragments into a similarity matching problem. This provides a useful reference for the reassembly of damaged bone sticks that were unearthed from the Wei-yang Palace in Chang'an City during the Han Dynasty, and it assists experts. However, if the features of bone stick fragments in the fractured region are not distinct, the algorithm may fail to identify matching fragments, leading to potential errors in judgment. Future work involves enhancing the robustness of the IMS-GC model.

Author Contributions: Conceptualization, J.H. and H.W.; methodology, J.H. and L.M.; software, J.H.; validation, J.H., L.M. and K.W.; formal analysis, Z.W.; investigation, T.W.; resources, R.L.; data curation, J.H. and R.L.; writing—original draft preparation, J.H.; writing—review and editing, J.H. and L.M.; visualization, K.W.; supervision, K.W., L.M. and K.W.; project administration, 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: This study did not require ethical approval for not involving humans or animals.

Informed Consent Statement: Not applicable.

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

Acknowledgments: This research was supported by experts from the Shaanxi Institute for the Preservation of Cultural Heritage, and we express our thanks for their assistance.

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

Abbreviations

The following abbreviations are used in this manuscript:
CBSD—Conjugable Bone Stick Dataset; MFS-GC—multi-scale feature fusion Siamese network guided by edge contour model; L2 distance—Euclidean distance.

References

1. Qi, H. Probe into the Archives of Bone Signet in Han Dynasty. *Lantai World Shenyang China* **2014**, *26*, 58–59.
2. Gao, M. The Bone Sticks of the Weiyang Palace in Chang'an City of Han Dynasty (9 Rules). *J. Bohai Univ. (Philos. Soc. Sci. Ed.)* **2022**, *44*, 86–89.
3. Gao, J. Restudy of the Name and usage of the bone tallies unearthed from the Han period Chang'an city-site. *Huaxia Archaeol.* **2011**, *3*, 109–113.
4. Zhang, C.; Zong, R.; Cao, S.; Men, Y.; Mo, B. AI-powered oracle bone inscriptions recognition and fragments rejoining. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021; pp. 5309–5311.
5. Shi, B.; Li, M. Automatic stitching and restoration algorithm for paper fragments based on angle and edge features. *J. Comput. Appl.* **2019**, *39*, 571–576.
6. Fang, R.; Huang, F.; Xin, H. Local matching for 2-D fragments reassembling. *Mod. Electron. Tech.* **2015**, *38*, 54–56.
7. Zhao, X.; Du, L. An Automatic and Robust Image Mosaic Algorithm. *J. Image Graph.* **2004**, *9*, 417–422.
8. Zhang, K.; Li, X. A graph-based optimization algorithm for fragmented image reassembly. *Graph. Models* **2014**, *76*, 484–495. [CrossRef]
9. Paumard, M.M.; Picard, D.; Tabia, H. Deepzzle: Solving Visual Jigsaw Puzzles With Deep Learning and Shortest Path Optimization. *IEEE Trans. Image Process.* **2020**, *29*, 3569–3581. [CrossRef] [PubMed]
10. Noroozi, M.; Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 69–84.
11. Le, C.; Li, X. JigsawNet: Shredded Image Reassembly Using Convolutional Neural Network and Loop-Based Composition. *IEEE Trans. Image Process.* **2019**, *28*, 4000–4015. [CrossRef] [PubMed]
12. Ngo, T.T.; Nguyen, C.T.; Nakagawa, M. A Siamese Network-based Approach For Matching Various Sizes Of Excavated Wooden Fragments. In Proceedings of the 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), Dortmund, Germany, 8–10 September 2020; pp. 307–312.
13. Zhang, Z.; Guo, A.; Li, B. Internal Similarity Network for Rejoining Oracle Bone Fragment Images. *Symmetry* **2022**, *14*, 1464. [CrossRef]
14. Zhang, Z.; Wang, Y.-T.; Li, B.; Guo, A.; Liu, C.-L. Deep Rejoining Model for Oracle Bone Fragment Image. In Proceedings of the Asian Conference on Pattern Recognition, Jeju Island, Republic of Korea, 9–12 November 2021; pp. 3–15.
15. Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; Shah, R. Signature verification using a "siamese" time delay neural network. *Adv. Neural Inf. Process. Syst.* **1993**, *6*, 737–744. [CrossRef]
16. Zagoruyko, S.; Komodakis, N. Learning to compare image patches via convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4353–4361.
17. 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 June 2016; pp. 770–778.
18. Park, T.; Liu, M.-Y.; Wang, T.-C.; Zhu, J.-Y. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2337–2346.
19. Tan, Z.; Chen, D.; Chu, Q.; Chai, M.; Liao, J.; He, M.; Yuan, L.; Hua, G.; Yu, N. Efficient Semantic Image Synthesis via Class-Adaptive Normalization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 4852–4866. [CrossRef] [PubMed]

20. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
21. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.
22. Lu, Z.; Bian, Y.; Yang, T.; Ge, Q.; Wang, Y. A New Siamese Heterogeneous Convolutional Neural Networks Based on Attention Mechanism and Feature Pyramid. *IEEE Trans. Cybern.* **2023**, *53*, 37021890. [CrossRef] [PubMed]
23. Wang, Z.; Zhu, J.; Fu, S.; Mao, S.; Ye, Y. RFPNet: Reorganizing feature pyramid networks for medical image segmentation. *Comput. Biol. Med.* **2023**, *163*, 107108. [CrossRef] [PubMed]
24. Yang, M.; Jiao, L.; Liu, F.; Hou, B.; Yang, S.; Jian, M. DPFL-Nets: Deep Pyramid Feature Learning Networks for Multiscale Change Detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 6402–6416. [CrossRef]
25. Rajevenceltha, J.; Gaidhane, V.H.; Anjana, V. A novel approach for Drowsiness Detection using Local Binary Patterns and Histogram of Gradients. In Proceedings of the 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, United Arab Emirates, 19–21 November 2019; pp. 1–6.
26. Yelampalli, P.K.R.; Nayak, J.; Gaidhane, V.H. A novel binary feature descriptor to discriminate normal and abnormal chest CT images using dissimilarity measures. *Pattern Anal. Appl.* **2019**, *22*, 1517–1526. [CrossRef]
27. Al Sameera, B.N.; Gaidhane, V.H.; Rajevenceltha, J. Image Focus Measure Based on Polynomial Coefficients and Reduced Gerschgorin Circle Approach. *IETE Tech. Rev.* **2023**. [CrossRef]
28. Chong, G.; Tian-Yuan, Q. Radiance Illumination Prediction of Sky Images Based on Siamese Networks. *Inf. Technol. Informatiz.* **2023**, *1*, 150–153.
29. Gang, D.; Xiang-Ning, W.; Yu-Jiao, D.; Yu, T.; Feng, Z.; Heng, F. PCB Defect Classification Model Based on Siamese Depth Feature Fusion Residual Network. *Comput. Syst. Appl.* **2023**, *32*, 211–219.