Ceramic Microscope Image Classification Based on Multi-Scale Fusion Bottleneck Structure and Chunking Attention Mechanism

Zhihuang Zhuang[1], Xing Xu*[2], Xuewen Xia[3], Yuanxiang Li[4], Yinglong Zhang[5]
School of Physics and Information Engineering, Minnan Normal University, Zhangzhou 363000, China[1,2,3,4,5]
Digital Strategy Development Research Institute of Hechi University, Hechi 546399, China[4]
School of Computer Science, Wuhan University, Wuhan 430072, China[4]

***Abstract*—In recent years, the status of ceramics in fields such as art, culture, and historical research has been continuously improving. However, the increase in malicious counterfeiting and forgery of ceramics has disrupted the normal order of the ceramic market and brought challenges to the identification of authenticity. Due to the intricate and interfered nature of the microscopic characteristics of ceramics, traditional identification methods have been suffering from issues of low accuracy and efficiency. To address these issues, there is a proposal for a multi-scale fusion bottleneck structure and a chunking attention module to improve the neural network model of Resnet50 and perform ceramic microscopic image classification and recognition. Firstly, the original bottleneck structure has been replaced with a multi-scale fusion bottleneck structure, which can establish a feature pyramid and establish associations between different feature layers, effectively focusing on features at different scales. Then, chunking attention modules are added to both the shallow and deep networks, respectively, to establish remote dependencies in low-level detail features and high-level semantic features, to reduce the impact of convolutional receptive field restrictions. The experimental results show that, in terms of classification accuracy and other indicators, this model surpasses the mainstream neural network models with a better classification accuracy of 3.98% compared to the benchmark model Resnet50, achieving 98.74%. Meanwhile, in comparison with non-convolutional network models, it has been found that convolutional models are more suitable for the recognition of ceramic microscopic features.**

***Keywords*—Deep learning; ceramic anti-counterfeiting; image classification; attention mechanism**

I. Introduction

Ceramics is a material and product discovered and produced by humans in their daily lives on Earth [1]. It is a hard product made from minerals such as clay through a series of physical and chemical reactions in a high-temperature environment. Due to its high practical value, the ceramic preparation process has been passed down through generations, and over time, this process has become increasingly refined. Throughout different historical periods, there are representative ceramic masterpieces characterized by a unique style, which to some extent, reflects the levels of productivity in different periods. Driven by this historical value, the trend of collecting ceramics has naturally flourished, while also endows ceramics with significant economic value. However, with the improvement of the technological level, the considerable economic benefits of ceramics have also given rise to the imitation industry. The rough and deliberately made products maliciously infiltrate every corner of the ceramics culture and trading market. This phenomenon not only infringes on the legitimate rights and interests of consumers. It also affects the dissemination and promotion of ceramic art and culture. Therefore, adopting a scientific and effective identification method is particularly important.

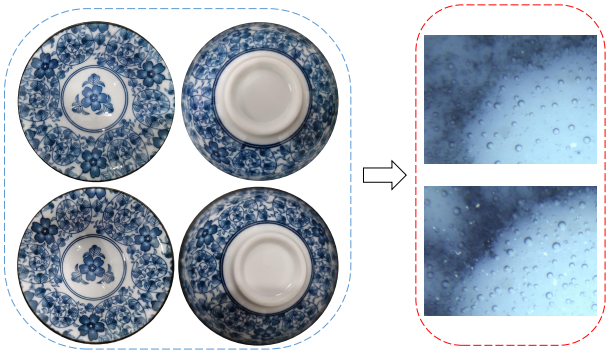


Fig. 1. Ceramic identification objects: macroscopic and microscopic images.

In recent years, with the rapid development of the field of computer vision, various universal visual tasks have been continuously refreshed with optimal indicators. At the early stage of the development of visual methods, visual feature extraction was mainly carried out by designing manual features [2]. In order to reduce the cost of feature engineering, the deep learning method represented by Convolutional neural network gradually has replaced the traditional manual feature method and achieved an excellent performance in basic visual tasks such as object detection and image segmentation [3], [4]. In the field of ceramic identification, this image-based identification method does not cause secondary damage to ceramics, and with the help of visual algorithms, it can achieve good differentiation of different ceramics. Therefore, scholars have also invested in the study of ceramic images [5]. At present, research on ceramic image identification has been mainly based on a macro perspective, by designing manual features or deep features for feature extraction, followed by feature classification. However, with the advancement of the ceramic manufacturing process, it is now possible to replicate the macroscopic appearance of ceramics completely (as shown in the left side of Fig. 1). It is less likely to guarantee the accuracy of the identification results by solely relying on

ceramic images for identification. During the physical and chemical process of ceramic firing, microscopic features such as crystallization and bubbles would emerge on the surface (as shown in the right side of Fig. 1). Even ceramics with similar macroscopic textures would exhibit certain differences when observed from a microscopic perspective. These randomly distributed microscopic features and texture variations are akin to the fingerprints of ceramics, endowing them with uniqueness. Therefore, the microscopic images of ceramics are more suited for the task of identification. On the other hand, currently, there is a lack of publicly available ceramic microscopic feature datasets in the market. In addition, personally-collected microscopic datasets of ceramics are limited in scale, and the complexity and non-uniformity of microscopic visual features pose challenges in feature recognition. Therefore, there is an urgent need for a specialized method in ceramic identification through microscopic visual analysis.

Therefore, this paper proposes a multi-scale fusion bottleneck structure and chunking attention module to solve the above problems and constructs a deep residual multi-scale network for feature classification of micro images of Jingdezhen and Dehua ceramics. The main contributions can be summarized as follows:

1) Shifting the research object in the field of ceramic identification from ceramic composition and macroscopic images to the study of microscopic images of ceramics. By collecting 12 pairs of microscopic images of ceramics with similar macroscopic features and conducting classification experiments, the effectiveness of this study was verified, and to some extent, the risks brought by ceramic imitation were solved.
2) Proposing a multi-scale fusion bottleneck structure and a chunking attention module for capturing features of different scales in images and reducing the computational cost of establishing feature remote dependencies. They can be easily embedded into deep neural networks.
3) Making a model improvement was made on the classic deep residual network and incorporating the two modules mentioned above. A deep residual network based on multi-scale fusion and attention mechanism was proposed, and in the collected ceramic micro datasets, it surpassed the current mainstream classification models and achieved the optimal results of benchmark testing.

The structure of this paper is arranged as follows: Section II will discuss related work. Section III mainly illustrates the relevant modules and algorithm processes in the model. Section IV mainly focuses on the microscopic data and experimental situation of ceramics and analyzes them. Section V discusses and summarizes the research content of this paper.

II. Related Work

In recent years, there have been many studies using images as a medium in the field of porcelain product recognition. Mu et al. [6] constructed manual visual features based on the contour, texture, and other information of macroscopic images of ancient ceramics and achieved a recognition rate of over 95% in ceramic recognition. However, this manual feature-based recognition method is only applicable to ancient ceramics with relatively fixed shapes and cannot adapt to the increasingly diverse types of modern ceramics. The development of deep learning has somewhat solved the limitations brought by manual features. Jiapeng et al. [7] used neural networks to classify ceramic images of different macroscopic shapes and achieved an accuracy of 92.62%. Yi et al. [8] constructed a set of ceramic classification standards for visual elements such as shape, color, and pattern of ceramics and achieved 72% pattern classification accuracy through target detection by using neural networks, ultimately formed a ceramic classification system. Chetouani et al. [9], [10] automatically classified the ceramic fragment images by constructing a Convolutional neural network and achieved the best accuracy. These studies have improved the archaeological efficiency and verified the superiority of neural networks in the field of ceramic classification. The above research on macroscopic images of ceramics still has certain limitations in scenarios with similar macroscopic features.

Therefore, another type of ceramic recognition research designed the microscopic images of ceramics. Wang et al. [11] proposed a fractal reconstruction method for high-temperature ceramic surface images and established a fractal Convolutional neural network model for image recognition, which achieved a classification accuracy of 93.74%. Min et al. [12] recognized the microscopic characteristics of ceramics through a Convolutional neural network and then carried out feature detection. Although these studies have shown some significant effects in their respective application fields, they have not yet taken into account the similarity in macroscopic appearance in ceramic identification. In addition, ceramic microscopic images can also be used for studying the properties and identifying the composition of ceramics. Hogan et al. [13] discovered the relationship between compression testing and microstructure changes by conducting uniaxial and biaxial compression experiments on ceramics and conducting stress analysis while observing changes in ceramic microscopic images. Aprile et al. [14], [15] identified the composition of ceramics through microscopic image acquisition methods such as OM and conducted modal analysis. This method of detection can avoid complex component extraction processes. In terms of detection, Guang et al. [16] improved the YOLO v5 model by combining the attention mechanism and depth separable convolution to detect defects in ceramic tile surface images. Huiliang et al. [17] used a graph structure clustering algorithm and Convolutional neural network to detect defects on ceramic tile surfaces. These studies have shown that Convolutional neural networks can also be used in ceramic detection tasks.

On the other hand, convolutional neural networks are not exclusive to pure image modal data. Yong et al. [18] used a full Convolutional neural network to classify the components of Jian kiln black glaze porcelain from the Song Dynasty in Fujian Province, thereby assisting in the classification of ceramics. This research also brought the possibility of multimodal analysis of ceramics through a Convolutional neural network and also had the prospect of using this technology in the field of ceramic identification. In terms of ceramic anti-counterfeiting, in addition to studying the characteristics of ceramics themselves, some scholars have also added anti-counterfeiting components to achieve ceramic anti-counterfeiting. Jae et al. [19] invented an anti-counterfeiting

material through spray pyrolysis and applied it to the field of ceramic anti-counterfeiting. However, the identification cost and threshold of the identification end cannot be avoided by anti-counterfeiting in this way. Therefore, the pure image method has its unique advantages in the identification of ceramic authenticity. Nevertheless, current image recognition algorithms face certain bottlenecks in the identification of ceramic microscopic images. Methods based on manual features exhibit low accuracy and poor generalization in identifying ceramic microstructures, failing to meet the demands of complex and diverse ceramic microscopic image recognition. In order to address this issue, this paper enhances deep residual networks by combining multi-scale fusion and attention mechanisms, aiming to achieve high-accuracy identification of ceramic microscopic images.

III. Model Design

In the field of image classification, the Convolutional neural network has always been a simple and effective model. The depth residual Convolutional neural network proposed by this paper, which is based on the combination of multi-scale and attention mechanisms, is an efficient and effective Convolutional neural network that has obvious performance advantages in the field of ceramic microscopic image data and can effectively characterize complex ceramic microscopic characteristics. In this chapter, we will introduce the principles and techniques related to the proposed multi-scale fusion bottleneck structure and chunking attention mechanism. Additionally, we will present the main details of the model used for this ceramic microscopic image classification task.

A. Multi-scale Fusion Module

In the design process of a Convolutional neural network, to improve the feature extraction ability of the model, it is often necessary to expand the scale of the model in a variety of ways, the most representative of which is widening and deepening [20], [21].Since the proposal of Resnet [22], this field has, for the first time, expanded the depth of the model to a scale greater than three digits, while also avoiding the risks of model degradation and overfitting. The residual connection and bottleneck structure proposed in this paper have also had a profound impact on subsequent research [23]. In addition, some researchers believe that the performance of the model is confined by the local dependency of convolution operations. Therefore, to make the model globally dependent, models represented by attention mechanisms have emerged [24].

On the other hand, it has been observed that there are a large number of bubble features and micro-texture information in ceramic micro images, and there are also certain differences in the background information of these key features. Therefore, this background information is also worth utilizing. Based on this motivation, a feature pyramid approach has been introduced to fuse ceramic micro background information at different scales. This multi-scale feature extraction method can effectively improve the recognition ability of the model [25].

However, the current mainstream multi-scale feature extraction methods have just simply stacked features, ignoring the correlation and importance between different scale feature maps. In order to integrate multi-scale features of the model and explore the correlation between different scale features for weighted fusion, this module establishes cross-correlations for different scale features through the attention mechanism and improves on the traditional bottleneck structure to form a new multi-scale fusion bottleneck structure.

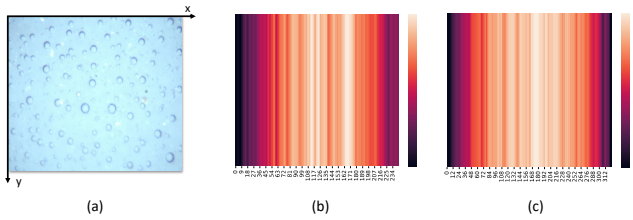


Fig. 2. Example diagram of ceramic microscopic characteristic coordinate axis pooling.

1) Cross-scale coordinate attention mechanism: The cross-scale coordinate attention mechanism can be seen as a feature weighting operation for features from different scales. From Fig. 2 (a), it can be observed that ceramic micro features exhibit different distributions along the X and Y coordinate axes, and their grayscale pooling features thermal maps are shown in Fig. 2 (b) and Fig. 2 (c). Therefore, modeling the information extracted from the X-axis and Y-axis directions in ceramic microscopic images can enhance the model's attention to important features.

This attention mechanism accepts any two feature tensor inputs of different scales, let it be set as $X_x = [x_1^x,\ x_2^x,\ ...,\ x_C^x]$, $X_y = [x_1^y,\ x_2^y,\ ...,\ x_C^y]$, where $X_x,\ X_y \in \mathbb{R}^{C\times H\times W}$. Firstly, encode the different channels of input features along the X and Y axes to form two one-dimensional feature sequences. The calculation process uses two global feature pooling operations, represented as follows:

$$\begin{cases} y_c^x\left(h\right) = \frac{1}{W}\sum_{i=1}^{W} x_c^x\left(h,\ i\right) \\ y_c^y\left(w\right) = \frac{1}{H}\sum_{i=1}^{H} x_c^y\left(w,\ i\right) \end{cases} \tag{1}$$

where, $Y_x = [y_1^x,\ y_2^x,\ ...,\ y_C^x]$ and $Y_y = [x_1^y,\ x_2^y,\ ...,\ x_C^y]$ represent the two coding sequences $Y_x \in \mathbb{R}^{C\times H\times 1}$ and $Y_y \in \mathbb{R}^{C\times W\times 1}$. This step captures the global position information of the coordinate axis direction from different scale feature maps, enhancing information sharing in the direction.

The second step is to concatenate the above features to form a new feature whole. In order to make the features from two scales interact effectively, it is usually considered to map the tensor to another linear space. Therefore, using a 1×1 convolutional kernel can perform linear transformations on the channel of the feature map, followed by feature activation and other operations, represented as follows:

$$\mathcal{X} = \mathcal{R}\left(\mathcal{B}\left(Conv_{c\to c/r}^{1\times 1}\left(\left[Y_x,\ Y_y\right]\right)\right)\right) \tag{2}$$

where, $\mathcal{X} \in \mathbb{R}^{\frac{C}{r}\times(H+W)\times 1}$. $Conv$ refers to convolution operation, and r refers to the compression ratio of the feature channel. Generally, this number is an integral power of 2. In this paper, $r = 32$, $\mathcal{B}$ refers to Batch Normalization, and $\mathcal{R}$

refers to the Activation function of a kind of deformation of RELU, which can limit the data range to 0 to 1 to better adapt to image characteristics. After that, the calculated feature is taken as the initial attention score of the attention mechanism, weight the feature, and then segment the size feature corresponding to the original X-axis and Y-axis. For the features corresponding to the X-axis and Y-axis, we also pass two sets of 1×1 convolution kernel is inversely mapped into the linear space of the original input, and the activation function is used to normalize the corresponding axis attention score, which is expressed as follows:

$$\begin{cases} \mathcal{D}_x = \mathcal{F}\left(Conv_{c/(hr)\to c}^{1\times 1}\left(\mathcal{V}^H\right)\right) \\ \mathcal{D}_y = \mathcal{F}\left(Conv_{c/(wr)\to c}^{1\times 1}\left(\mathcal{V}^W\right)\right) \end{cases} \tag{3}$$

where, $\mathcal{V}^H \in \mathbb{R}^{\frac{C}{r}\times H\times 1}$, $\mathcal{V}^W \in \mathbb{R}^{\frac{C}{r}\times W\times 1}$ represent the feature inputs corresponding to X-axis and Y-axis, and $\mathcal{D}_x \in \mathbb{R}^{C\times H\times 1}$, $\mathcal{D}_y \in \mathbb{R}^{C\times W\times 1}$ represent the feature outputs corresponding to X-axis and Y-axis. The activation function $\mathcal{F}$ is a sigmoid function, which exhibits an S-shaped growth curve in biology. By applying this function during normalization, it performs a nonlinear transformation of features, enabling the model to recognize more complex features.

Finally, the original tensor has been weighted with the attention fraction of the coordinate axis in the X-axis and Y-axis directions as follows:

$$\mathcal{Y} = X_x \odot \mathcal{D}_x \odot \mathcal{D}_y \tag{4}$$

where, $\odot$ represents the point multiplication operation of the tensor, and $\mathcal{Y} \in \mathbb{R}^{C\times H\times W}$ represents the weighting result of the tensor along the channel for its own X-axis feature and the Y-axis feature of other scale tensors.

The above are the details of the cross-scale coordinate attention mechanism. It should be noted that this module focuses on the cross-influence between scales. Therefore, for feature input, it is necessary to ensure that the size of the feature tensor of the two scales is consistent. The specific process is shown in Algorithm 1.

Algorithm 1 Cross-scale coordinate attention algorithm.

Input: Different scale feature X_x, X_y, squeeze ratio r.
Output: Cross-scale coordinate attention-weighted feature $\mathcal{Y}$.

Compute Y_x, Y_y according to Eq. (1)
Compute $\mathcal{X}$ according to Eq. (2)
Compute S according to Eq. (2) without $\mathcal{R}$
$\mathcal{V} = \mathcal{X} \odot S$
$\mathcal{V}^H, \mathcal{V}^W = Split\left(\mathcal{V}\right)$
Compute $\mathcal{D}_x, \mathcal{D}_y$ according to Eq. (3)
Compute $\mathcal{Y}$ according to Eq. (4)
return $\mathcal{Y}$

2) Multi-scale fusion bottleneck structure: The traditional residual bottleneck structure is composed of two 1×1 and a 3×3 convolution kernel stack. On this basis, the multi-scale fusion bottleneck structure replaces the 3×3 convolution kernel with multiple 3×3 convolution kernels and introduces the cross-scale coordinate attention mechanism to mine the correlation between different scales.

Specifically, after the feature passes through the first 1×1 convolution kernel, it is divided into s parts according to the number of channels. Where the first $s-1$ sub-features each have a 3×3 convolution kernel corresponding to them one-to-one, and the features entering the current convolution operation are those formed as a result of the mutual accumulation of the output of the previous convolution operation and the current sub-feature. After the scale feature pyramid operation, distinctive features such as bubbles will undergo further enhancement through a chain of convolutional operations. This operation can be represented as follows:

$$y_i = \begin{cases} Conv_{c/r\to c/r}^{3\times 3}\left(x_i\right), & i = 1 \\ Conv_{c/r\to c/r}^{3\times 3}\left(y_{i-1} + x_i\right), & 1 < i < s \\ x_i, & i = s \end{cases} \tag{5}$$

where, $x_i \in \mathbb{R}^{\frac{C}{s}\times H\times W}$ and $y_i \in \mathbb{R}^{\frac{C}{s}\times H\times W}$ denote the input and output features, respectively. The variable s represents the number of scale divisions, which is a factor of channel number C.

However, it is important to note that the presence of noise points in the feature map can contaminate the subject features during the convolutional operation chain. This method overlooks the differences and correlations between adjacent scales, and the straightforward superposition of features can magnify this error. Therefore, a cross-scale coordinate attention mechanism has been introduced between adjacent scale features to enable the model to accurately identify important

Algorithm 2 Multi-scale fusion bottleneck structure.

Input: Input feature X, split number s.
Output: Output feature Y.

$[x_1, x_2, ..., x_s] = Split\left(Conv2D_{in\to hidden}^{1\times 1}\left(X\right)\right)$
out initial value is $\oslash$
for each i $\in [1,\ s)$ **do**
  if $i \neq 1$ **then**
    $cur\_feat = Conv_{c/r\to c/r}^{3\times 3}\left(x_i + pre\_feat\right)$
    $cur\_feat = Relu\left(\mathcal{B}\left(cur\_feat\right)\right)$
    $z_{i-1,\ i} = \Phi\left(pre\_feat,\ cur\_feat\right)$
    $z_{i,\ i-1} = \Phi\left(cur\_feat,\ pre\_feat\right)$
    $out = Concat\left(out,\ z_{i-1,\ i},\ z_{i,\ i-1},\ cur\_feat\right)$
  else
    $cur\_feat = Conv_{c/r\to c/r}^{3\times 3}\left(x_1\right)$
    $cur\_feat = Relu\left(\mathcal{B}\left(cur\_feat\right)\right)$
    $out = Concat\left(out,\ cur\_feat\right)$
  end if
  $pre\_feat = cur\_feat$
end for
$z_{s-1,\ s} = \Phi\left(pre\_feat,\ x_s\right)$
$z_{s,\ s-1} = \Phi\left(x_s,\ pre\_feat\right)$
$Y = X + Concat\left(out,\ z_{s-1,\ s},\ z_{s,\ s-1},\ x_s\right)$
return Y

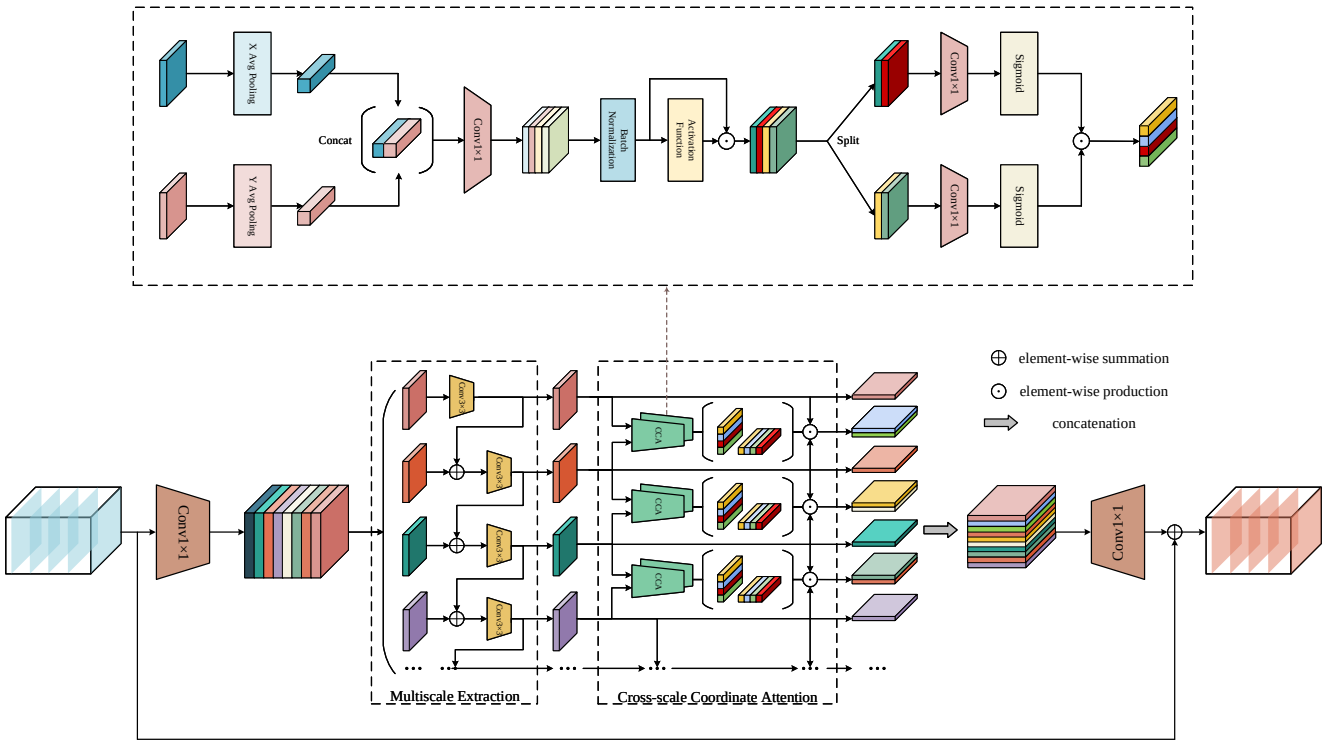


Fig. 3. Flowchart of the multiscale fusion bottleneck structure when s = 4.

features. When calculating attention features between different scales, it is important to consider both the influence of oneself on adjacent scales and the influence of adjacent scales on oneself. As a result, the total output quantity of the operation is $s + 2 \times (s - 1)$. Here, s represents the amount of channel segmentation for the original feature. This operation can be represented as follows:

$$z_i = Concat\left(\Phi\left(y_{i-1},\ y_i\right),\ \Phi\left(y_i,\ y_{i-1}\right)\right),\ 1 < i \leq s \quad (6)$$

where, $z_i \in \mathbb{R}^{2\times\frac{C}{s}\times H\times W}$ represents the cross-scale coordinate attention feature between adjacent scales, $Concat$ represents the connection operation between channels and Φ represents the cross-scale coordinate attention operation.

Fig. 3 illustrates the multi-scale fusion bottleneck structure when $s = 4$. After the aforementioned computations have resulted in a feature map with a channel number of $3 \times s - 2$, a second 1×1 convolution is utilized to transform the channel number to the standard quantity. The specific process is as shown in Algorithm 2.

B. Chunking Attention Module

Since Self Attention [26], [27] has been proposed, it has played a role in both computer vision and Natural language processing. On this basis, many classic architectures have also emerged [3], [4]. In contrast, the Convolutional neural network lacks the ability to establish a long-distance global dependency, so intuitively, it is very likely to establish such global dependency for the Convolutional neural network by introducing the Self Attention mechanism. However, the computational cost of Self Attention is unexpectedly high, so there has been a series of efforts to improve the problem and propose corresponding solutions for different fields [28], [29]. In this section, a method has been proposed based on Self Attention to establish attention-weighted features between local and global blocks through block partitioning. This approach effectively reduces computational complexity while still enabling the establishment of long-distance dependencies. This section describes the techniques and principles of spatial chunking attention and channel chunking attention.

1) Overview of self attention: The Self Attention mechanism considers a feature tensor, respectively, as Query, Key, and Value, and obtains its important features through the operation between them. The calculation method is as follows:

$$Attention\left(Q,\ K,\ V\right) = Softmax\left(\frac{Q \otimes K^T}{\sqrt{d_k}}\right) \otimes V \quad (7)$$

where, $\otimes$ represents the Matrix multiplication of the tensor, and $Q,\ K,\ V \in \mathbb{R}^{n\times d_k}$ represents the input characteristic tensor. T represents the matrix transpose operation. In the field of vision, the value of n is generally the size of the image $h \times w$, and d_k represents the number of feature channels. This formula includes two Matrix multiplications. First, through the operation between Query and Key, and $Softmax$, the attention score of each pixel in the global is calculated. The Softmax formula is as follows:

$$Softmax\left(X_{ij}\right) = e^{X_{ij}} / \sum_{z=1}^{n} e^{X_{iz}} \quad (8)$$

Then, the global pixels are weighted by the second Matrix multiplication. It is not difficult to find that the computational complexity of this operation is $\Omega\left(2n^2d_k\right)$. However, this computational complexity is unacceptable before the feature map undergoes multi-layer downsampling. It is noticed that pixels interact with the entire feature map during the calculation of the global attention score, which is the fundamental reason for the increase in computational complexity. Therefore, Query, Key, and Value have been redesigned to reduce computational complexity.

2) Spatial chunking attention: The Spatial Chunking Attention module starts by dividing the feature map into uniform spatial patches. Inspired by the divide-and-conquer algorithm, it independently computes attention-weighted features for each sub-patch. Finally, these features are merged to form the Spatial Chunking Attention features. Fig. 4 illustrates the calculation process of the Spatial Chunking Attention module, which consists of the following specific steps:

Set the input feature tensor as $X \in \mathbb{R}^{C\times H\times W}$. Similarly, before calculating attention scores, form preliminary Query, Key, and Value tensors through a set of learnable convolutional kernels, which are represented as follows:

$$\begin{cases} Query = Conv^{1\times1}_{C\to C/r}(X) \\ Key = Conv^{1\times1}_{C\to C/r}(X) \\ Value = Conv^{1\times1}_{C\to C}(X) \end{cases} \tag{9}$$

where, $Query,\ Key \in \mathbb{R}^{\frac{C}{r}\times H\times W}$, $Value \in \mathbb{R}^{C\times H\times W}$. Next, we divide the Query and Value inputs uniformly to create two sets of patch sequences ($K_h = \sqrt{H},\ K_w = \sqrt{W}$), namely $Q \in \mathbb{R}^{K_h\times K_w\times\sqrt{HW}\times\frac{C}{r}}$ and $V \in \mathbb{R}^{K_h\times K_w\times\sqrt{HW}\times C}$. For the definition of Key, if the Self Attention setting is followed, the sub-patch will only have its own local dependency. Therefore, perform global feature mean pooling and sample a set of globally abstract features as K, which is represented as follows:

$$\begin{cases} K^{<i,\ j>}_{c/r} = \frac{1}{\sqrt{H\times W}}\sum_{s=1}^{\sqrt{H}}\sum_{t=1}^{\sqrt{W}} Key_{c/r}(u,\ v) \\ u = (i-1)\times\sqrt{H}+s, \quad 1\le i\le K_h \\ v = (j-1)\times\sqrt{W}+t, \quad 1\le j\le K_w \end{cases} \tag{10}$$

where, $K_{c/r} \in \mathbb{R}^{\sqrt{HW}\times\frac{C}{r}}$. By solving attention scores with globally abstract features, global dependencies can be effectively established. At this point, the Spatial Chunking Attention feature can be obtained through Eq. (8), which is calculated as follows:

$$Y^{<i,\ j>}_c = Softmax\left(\frac{Q^{<i,\ j>}_{c/r}\otimes K^T_{c/r}}{\sqrt{c/r}}\right)\otimes V^{<i,\ j>}_c \tag{11}$$

where, $Y^{<i,\ j>}_c \in \mathbb{R}^{\sqrt{HW}\times C}$ represents the attention weighted features of each patch. Afterward, merge the patch features by location to restore spatial attention features $Y \in \mathbb{R}^{C\times H\times W}$.

Finally, a learnable feature has been proposed for modeling momentum representation in Spatial Chunking Attention, which is represented as follows:

$$\begin{cases} momentum = 0.5\odot gamma/(1+|gamma|)+0.5 \\ Z = momentum\odot Y+(1-momentum)\odot X \end{cases} \tag{12}$$

where, $momentum \in \mathbb{R}^{1\times1}$ represents the learnable momentum value, with a range of $[0,1]$. $Z \in \mathbb{R}^{C\times H\times W}$ represents the Spatial Chunking Attention weighted feature. At this point, the solution for Spatial Chunking Attention is obtained. It is worth noting that when partitioning the feature map, it is necessary to ensure that the dimensions H and W are perfect square numbers. Therefore, has been the prerequisite is not met, the feature map needs to be padded along the borders. This paper employs mirror padding, where the mirrored content of the original feature map is filled symmetrically with respect to the border.

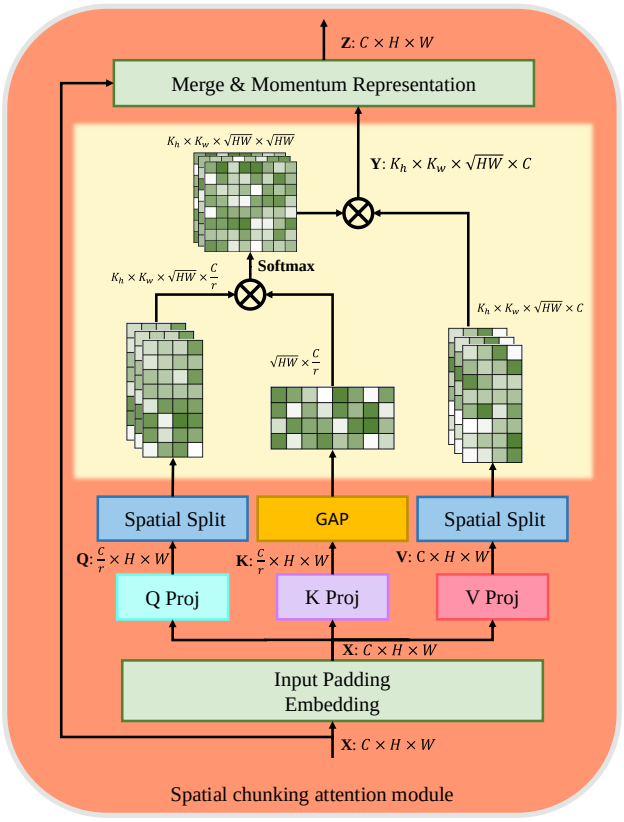


Fig. 4. Flowchart of the spatial chunking attention module.

3) Channel chunking attention: In the process of image downsampling, each highly abstract channel graph can be regarded as a special class response, and the response relationship between these channels together constitutes the semantic information of the target. In order to identify complex and disordered bubbles and other features in ceramic micro images,

the analysis of the response relationship between channel graphs can help the model understand the semantic information of ceramic micro images, and then improve the accuracy of recognition. Similarly, using Self Attention for channel global interaction can easily cause the unacceptable computational complexity of the model. Also, the idea of the divide and conquer algorithm con be used to divide the channel graph evenly, calculate the attention characteristics of each sub-patch independently, and finally merge to form the Channel Chunking Attention. The calculation process is as shown in Fig. 5. The specific calculation steps are as follows:

Firstly, transform the number of channels in the feature map using a set of 1×1 convolutional kernels, which is represented as follows:

$$\begin{cases} Query = Conv_{C \to SC}^{1\times 1}(X) \\ Key = Conv_{C \to SC}^{1\times 1}(X) \\ Value = Conv_{C \to SC}^{1\times 1}(X) \end{cases} \tag{13}$$

where, $X \in \mathbb{R}^{C\times H\times W}$ and $Query,\ Key,\ Value \in \mathbb{R}^{SC\times H\times W}$. SC is the perfect square number greater than the original number of channels. Similarly, after reconstructing the number of channels, perform the average segmentation to form two sets of $K_{ch}\times K_{cw}$ patch sequences ($K_{ch} = K_{cw} = \sqrt[4]{SC}$), namely $Q,\ V \in \mathbb{R}^{K_{ch}\times K_{cw}\times \sqrt{SC}\times HW}$, for Query and Value. For Key, use global feature mean pooling to obtain a set of feature K, whose operation is represented as follows:

$$\begin{cases} K_{sc}^{<i,\ j>} = \frac{1}{\sqrt{SC}} \sum_{s=1}^{K_{ch}} \sum_{t=1}^{K_{cw}} Key_{sc}(u,\ v) \\ u = (i-1)\times K_{ch} + s, \quad 1 \le i \le K_{ch} \\ v = (j-1)\times K_{cw} + t, \quad 1 \le j \le K_{cw} \end{cases} \tag{14}$$

where, $K_{sc} \in \mathbb{R}^{\sqrt{SC}\times HW}$. Afterward, we can establish a global dependency on the channel through Eq. (8), which is represented as follows:

$$Y_{sc}^{<i,\ j>} = Softmax\left(\frac{Q_{sc}^{<i,\ j>} \otimes K_{sc}^{T}}{\sqrt{H\times W}}\right) \otimes V_{sc}^{<i,\ j>} \tag{15}$$

where, $Y_{sc}^{<i,\ j>} \in \mathbb{R}^{\sqrt{SC}\times HW}$ represents the attention-weighted feature of each channel patch. In addition, to restore the feature dimension, it is necessary not only to merge each channel patch but also apply a set of 1×1 convolutional kernels to restore the number of channels. The representation is as follows:

$$P = Conv_{SC \to C}^{1\times 1}(Y) \tag{16}$$

where, $Y \in \mathbb{R}^{SC\times H\times W}$ represents the result of merging patch features by channel, and $P \in \mathbb{R}^{C\times H\times W}$ represents the result of restoring the number of feature channels. Similarly,

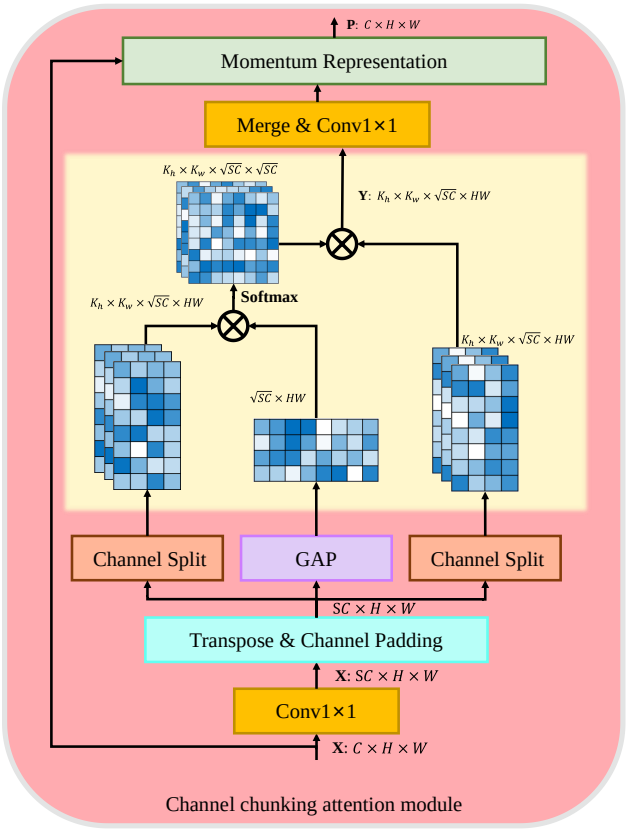


Fig. 5. Flowchart of the channel chunking attention module.

according to Eq. (12), a set of learnable momentum representations have also been designed for Channel Chunking Attention. At this point, the solution for Channel Chunking Attention is completed.

4) Complexity analytics: The Spatial Chunking Attention mechanism divides the spatial plane into $K_h \times K_w$ sub-patches on average. To simplify the analysis, it is assumed that the tensor dimensions H and W are perfect square numbers, so the size of each sub-patch is $\sqrt{H} \times \sqrt{W} \times C/r$. Therefore, the Time complexity of solving attention score is $\Omega\left(K_h K_w \left(\sqrt{HW}\right)^2 C/r\right)$. The second part of the operation is Matrix multiplication between the spatial attention score and the $K_h \times K_w$ sub-patches of V, where $V \in \mathbb{R}^{K_h\times K_w\times \sqrt{HW}\times C}$ and $Score \in \mathbb{R}^{K_h\times K_w\times \sqrt{HW}\times \sqrt{HW}}$. Therefore, the Time complexity of this part is $\Omega\left(K_h K_w \left(\sqrt{HW}\right)^2 C\right)$. Based on the above, the overall time complexity of this module is denoted as $\Omega\left(\left(\sqrt{HW}\right)^3 (1/r+1)\, C\right)$.

Similarly, the Channel Chunking Attention mechanism divides channels into $K_{ch} \times K_{cw}$ sub-patches on average. The number of channels is assumed as a second-order perfect square number, that is, SC and $\left(\lfloor \sqrt[4]{SC} \rfloor\right)^4$ are equal. For the first part of the calculation of attention score, it is the Matrix multiplication between $Q \in \mathbb{R}^{K_{ch}\times K_{cw}\times \sqrt{SC}\times HW}$ and $K^T \in \mathbb{R}^{HW\times \sqrt{SC}}$, and the Time complexity is

$\Omega\left(K_{ch}K_{cw}\left(\sqrt{SC}\right)^2 HW\right)$. The operation of the second part is Matrix multiplication between the channel attention score and the $K_{ch} \times K_{cw}$ sub-patches of V. Where $V \in \mathbb{R}^{K_{ch}\times K_{cw}\times\sqrt{SC}\times HW}$ and $Score \in \mathbb{R}^{K_{ch}\times K_{cw}\times\sqrt{SC}\times\sqrt{SC}}$. It is not difficult to find that the Time complexity of this part is $\Omega\left(K_{ch}K_{cw}\left(\sqrt{SC}\right)^2 HW\right)$. Therefore, the time complexity of the Channel Chunking Attention module is $\Omega\left(2\left(\sqrt{SC}\right)^3 HW\right)$.

C. Model Backbone Architecture

The overall structure of this model is as shown in Fig. 6 (a), where the backbone network will be designed according to the 4-stage principle [22]. Due to the different image representation capabilities of each stage block, shallow stage blocks retain more ceramic microscopic details, while deep stage blocks have a higher level of abstraction ability for ceramic microscopic images, which can extract higher-level semantic information. Based on the above characteristics, the chunking attention module has been added to the first and fourth stage blocks, respectively, to enable the model to model global key features, thereby combining low-level detail features with high-level semantic features. In this way, the model can fully utilize global contextual information and generate a more accurate characterization of ceramic microscopic images.

The number of 4-stage bottleneck structures in the entire backbone network is 3, 4, 6, and 3, respectively. In terms of feature channel changes, the model first performs feature convolution on the input image through a 7×7 large convolution kernel, and its output is a 64-channel feature tensor. Next, in the first bottleneck structure in stage-1, the number of channels is expanded to four times the original number, and the number of internal channels remains unchanged. Therefore, the feature output of this layer is a feature tensor of 256 channels. Afterward, the next three stage blocks will undergo the feature transfer in this form. The difference is that the first bottleneck structure of each remaining stage will be expanded by twice the number of channels, resulting in a final feature output channel of 2048. In terms of feature size changes, for the first bottleneck structure of each stage block, the convolution operations will be used to downsample the features transmitted from the shallow layer while maintaining the same feature size within each stage block. Therefore, the corresponding feature sizes within the four stage blocks are $(H/4) \times (W/4)$, $(H/8) \times (W/8)$, $(H/16) \times (W/16)$, and $(H/32) \times (W/32)$.

For the bottleneck structure proposed in this paper, in terms of cross-multi-scale fusion, the segmentation number of $s = 4$ has mainly been adopted to divide the feature channels. In terms of Chunking Attention, it is noted note that Spatial Chunking Attention helps the model capture spatial relationships and local details in images, while Channel Chunking Attention helps the model understand the interaction and importance of different channels. Therefore, the reasonable combination of Spatial Chunking Attention and Channel Chunking Attention can make the model extract more robust semantic features, so as to enhance the recognition ability of the model. In order to effectively integrate the key features obtained by the two modules, it is necessary to design a serial and parallel feature computing structure, as shown in Fig. 6 (b) and Fig. 6 (c). In the serial structure, it is designed to cascade the two modules and obtain the final feature representation through sequential calculation. In parallel architecture, it is designed to fuse the features of the two modules by adding them point by point. Experiments have shown that both structures exhibit excellent performance.

IV. Experiments and Analysis

The experimental environment for the algorithm in this paper is a 64-bit Ubuntu 16.04.1 operating system with an Intel Core i9-10900k processor, 64GB of memory, an NVIDIA GeForce RTX 2080Ti graphics card, and a tensor operation library version of pytorch-1.8.1-cuda-10.1. This chapter will first introduce the collection of ceramic microscopic data, algorithm evaluation indicators, and experimental results, and analyze the results.

A. Ceramic Microscopic Image Dataset

In this paper, a camera of 600 times optical is used to collect microscopic images of 12 pairs of 24 Blue and white pottery tea cups from Jingdezhen and Dehua Fig. 7). After manual filtering of some pictures that are not correctly focused, the final size of this ceramic data set is 1548 pictures. In terms of dataset production, 24 tea cups have been divided into 24 categories, and their data formats were defined according to the ImageNet dataset. The division ratio between the training set and the test set is $7:3$.

To simulate the real ceramic imitation scene, the macroscopic shape of each pair of blue and white porcelain tea cups in the experiment will tend to be consistent. Therefore, if the model correctly classifies all the pictures on this dataset, especially the same pair of Blue and white pottery tea cups can be correctly classified. So, it can be considered that this model has high anti-counterfeiting performance for ceramics and can capture more essential and discriminative features in ceramic microscopic images. The number of samples collected for each pair of ceramic microscopic images is as shown in Fig. 8.

B. Evaluation Indicators

To objectively and fairly evaluate the performance of the proposed model and its performance on ceramic microscope image data, this paper will select evaluation indicators widely used in machine learning to test the effectiveness of the proposed model. Mainly including Accuracy(Acc), Precision(Pre), Recall(Rec), F1-Score(F1) and Kappa(Kap). This paper also calculates mAUC and mAP based on the Receiver Operating Characteristic(ROC) curve and Precision-Recall (PR) curve, respectively.

$$mAUC = \frac{1}{n} \times \sum_{i \in n} AUC_i \tag{17}$$

$$mAP = \frac{1}{n} \times \sum_{i \in n} AP_i \tag{18}$$

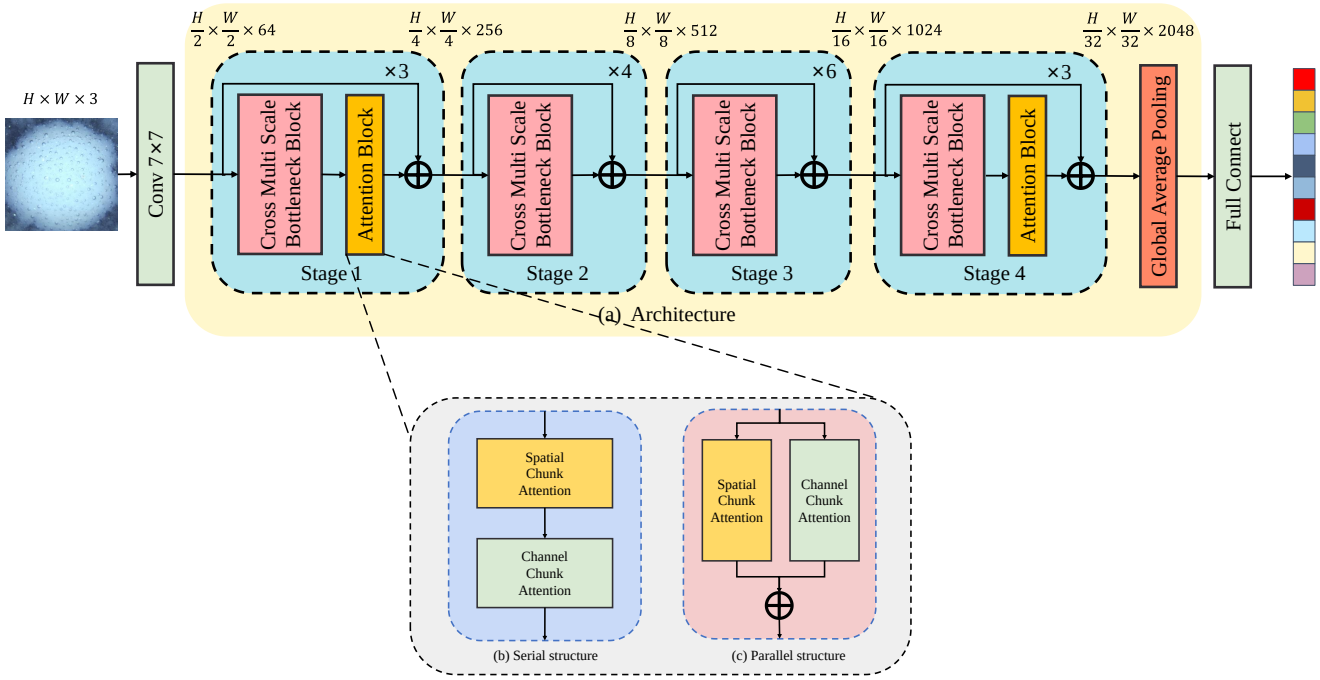


Fig. 6. Flowchart of the backbone network structure.

Fig. 7. Real object images of ceramic microscopic images captured by industrial cameras.

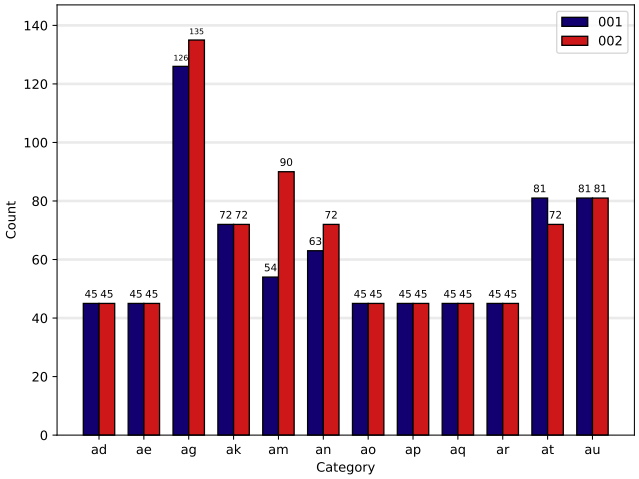


Fig. 8. The quantity distribution of ceramic microscopic image datasets.

In addition, due to the large number of learnable parameter weights in the neural network model used in this paper, including fully connected layers, convolutional layers, etc., its spatial and temporal costs are also worth paying attention to. Therefore, this paper will introduce Params, FLOPs, and inference time to evaluate the running cost of the model. Among them, Params represent the parameter quantity of the model, and FLOPs represent the number of floating-point operations of the model. It is worth noting that due to the presence of memory access costs (MAC), FLOPs cannot be equivalent to inference time. Therefore, we will calculate the average inference time for each ITERATION in the model.

C. Ablation Experiments

This study adopted a classic image classification experimental process. In the pre-processing stage, it is scheduled to randomly cut the original image to 224×224 pixels of standard size and randomly flip it with a probability of 50% to enhance the robustness of the model to interference. For neural network parameter optimization, the AdamW optimizer has been chosen with an initial learning rate of 0.001, a momentum of 0.9, and a weight regularization term of 0.1. To ensure that the model is not heavily influenced by significant updates in the wrong direction during the early stages of training, a linear warm-up learning rate strategy, which gradually increases the learning rate from 0.0001 to the initial value of 0.001 over

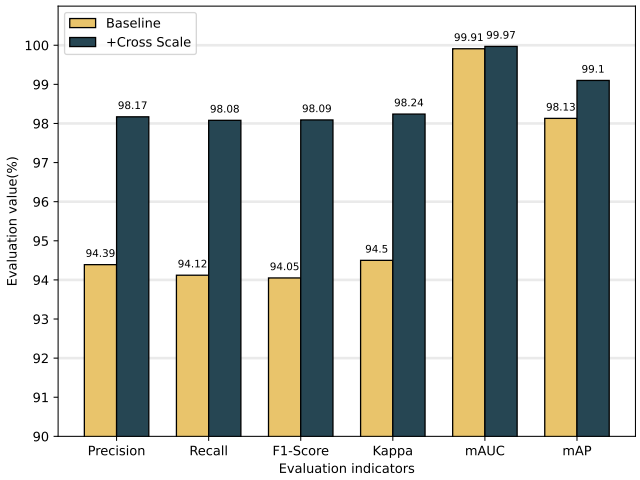


Fig. 9. Evaluation of the multi-scale fusion bottleneck structure in different indicators.

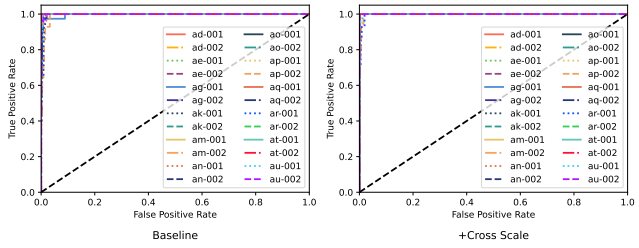
the first 30 epochs has been adopted. Then, for the following 370 epochs, a cosine annealing learning rate decay strategy has been employed. Through this training scheme, a total of 400 epochs, have been trained on the ceramic micro dataset and comprehensively evaluated the performance of the model on various indicators.

1) Effect of the multi-scale fusion bottleneck structure on the experimental results: To verify the effectiveness of the multi-scale fusion bottleneck structure on the ceramic microscope image dataset, the classic Resnet50 [22] has been selected as the benchmark and compared the changes in different evaluation indicators before and after replacing the multi-scale fusion bottleneck structure (see Table I). The results showed that the replacement of the multi-scale fusion bottleneck structure increased the Top1 Accuracy by 3.56%, reaching 98.32%, while the Top2 Accuracy increased by 0.42%, reaching 100%. From the gap between Top1 Accuracy and Top2 Accuracy, it can be seen that the model still has some performance degradation even when factors such as texture are consistent.

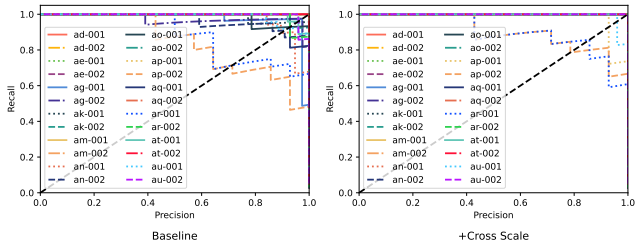
TABLE I. COMPARISON OF PARTIAL RESULTS OF THE MULTI-SCALE BOTTLENECK STRUCTURE UNDER BASELINE

Model	Params (M)	FLOPs (G)	Time (MS)	Top1-acc (%)	Top2-acc (%)
Baseline	23.557	4.109	102.65	94.76	99.58
+Cross Scale	36.1	6.535	76.9	98.32	100.00

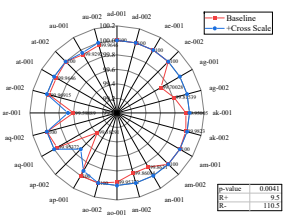
In terms of the number of parameters and the number of floating-point operations, due to the model's use of scale segmentation and the introduction of more convolutional kernels and cross-attention calculations, both indicators have correspondingly increased. It is worth noting that although the computational complexity of the model has increased, the adoption of a multi-scale fusion bottleneck structure with segmented scales enables the model to have a higher parallelism. According to the inference time test conducted on the GPU for the last batch of data in the test sample, the network that replaced this module showed a lower average inference time, reduced by 25.75ms.



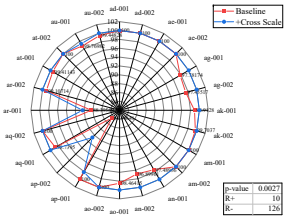
(a) ROC curve plotting for each category.



(b) PR curve plotting for each category.



(c) Comparison of AUC for each category and test.



(d) Comparison of AP for each category and test.

Fig. 10. The evaluation performance of multi-scale fusion bottleneck structure in ROC and PR curves and the area under their curves.

As shown in Fig. 9, the Resnet50 model, which replaces the multi-scale fusion bottleneck structure, has improved all the six indicators in the figure. Among them, the Precision, Recall, F1 Score, and Kappa coefficients have significantly increased, with an increase of about 3%; While the increase of mAUC and mAP is relatively low, to explore the performance results of this module in the Receiver operating characteristic and PR curve, Data and information visualization have been conducted on the evaluation of various categories of ceramic microscopic images:

First, in Fig. 10 (a) ROC curve, the multi-scale fusion bottleneck structure is closer to the upper left corner than the baseline model in most cases, and the baseline model has some area gaps in most categories. Secondly, in Fig. 10 (b) PR curve, it can be intuitively observed that the multi-scale fusion bottleneck structure is generally closer to the upper right corner. Based on the above two points, it indicates that the improved multi-scale fusion bottleneck structure is effective in improving the classification accuracy of various categories.

On the other hand, for each class of ROC curves and PR curves, the AUC and AP metrics can be derived, respectively. As shown in Fig. 10 (c) and Fig. 10 (d), the performance of the baseline model and the multi-scale fusion bottleneck structure for these two metrics in the 24 categories of ceramic microscope image data. It can be observed that the multi-scale fusion bottleneck structure outperforms the baseline model

TABLE II. COMPARISON OF PARTIAL RESULTS OF THE CHUNKING ATTENTION UNDER THE ORIGINAL BOTTLENECK STRUCTURE

Model	Params (M)	FLOPs (G)	Time (MS)	Top1-acc (%)	Top2-acc (%)
Baseline	23.557	4.109	102.65	94.76	99.58
+PA	37.143	6.06	74.80	94.76	99.58
+CA	49.966	6.677	92.42	95.39	99.79
+PCA-Serial	63.551	8.627	81.37	96.44	99.79
+PCA-Parallel	63.551	8.627	107.06	94.97	99.79

in most of the categories, especially on the data with more complex features like ap-002, which also has a higher recognition rate. To further verify whether the multi-scale fusion bottleneck structure is significantly superior to the baseline model in these two indicators, Wilcoxon signed rank test, a non-parametric hypothesis testing method that can compare the overall distribution differences between two paired samples have also been conducted. Here, R+ and R- respectively indicate the sum of ranks where the baseline model is greater than and less than the multi-scale fusion model in paired samples. In the testing process, the minimum value has been mainly chosen between these two as the test statistic. The larger the test statistic is the more significant the difference in the indicators will be. In this paper, hypothesis tests have been conducted at a significance level of $\alpha = 0.05$. For the AUC and AP metrics, the corresponding p-values are 0.0041 and 0.0027, both of which are smaller than the significance level α; therefore, it is necessary to reject the null hypothesis H_0 and accept the alternative hypothesis H_1. This indicates that the multi-scale fusion bottleneck structure exhibits significant differences compared to the baseline model in both of these metrics.

In summary, in the ceramic microscope image recognition task, the multi-scale fusion bottleneck structure of this model is superior to the 3×3 convolution in the baseline bottleneck structure. In the improvement of related tasks, it can be considered to replace it with this module to achieve higher recognition accuracy.

2) Effect of the chunking attention module combined with primitive bottleneck structure on experimental results: To verify the effectiveness of the chunking attention module, Resnet50 has still been used as the baseline model in this section and trained spatial chunking attention (PA) and channel chunking attention (CA), as well as their serial fusion structure (PCA Serial) and parallel fusion structure (PCA Parallel). The changes have also been evaluated in different indicators on the ceramic microscopic image dataset (see Table II). Under the original bottleneck structure, the results showed that the PA module was able to achieve the same accuracy as the baseline model with an average inference time reduction of 27.85ms, while the CA module increased Top1 Accuracy and Top2 Accuracy by 0.63% and 0.21%, respectively, based on an average inference time reduction of 10.03ms. In the serial and parallel structures fused with PA and CA, the Top1 Accuracy has been improved by 1.68% and 0.21%, respectively, but there is a significant difference in time between the two structures.

It is observed that in the parallel structure, Top1 Accuracy decreased by 0.42% compared to the CA module. This difference may be related to the limited ability of 3×3 convolutions in the original bottleneck structure for feature extraction. Therefore, in a parallel structure, the PA and CA modules calculate two feature tensors with significant distribution differences, and adding them up may interfere with the high-quality features extracted by the channel attention module, thereby reducing the recognition performance of the model. On the contrary, cascaded serial structures can further model the relationships between channels based on global spatial modeling, thus being able to identify more significant ceramic microscopic features. Therefore, when the feature extraction ability of bottleneck structure is limited, priority should be given to using serial fusion to avoid performance degradation results from distribution differences.

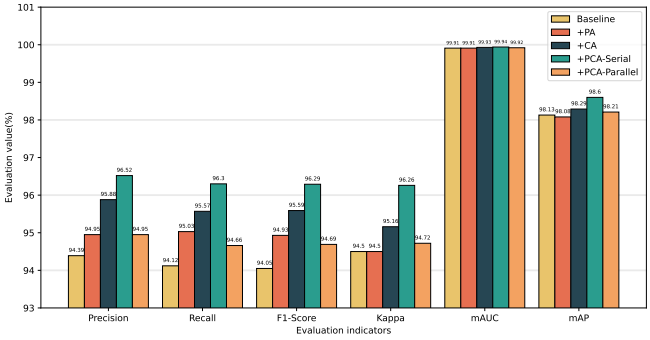


Fig. 11. Evaluation of the chunking attention module for different metrics in the original bottleneck structure.

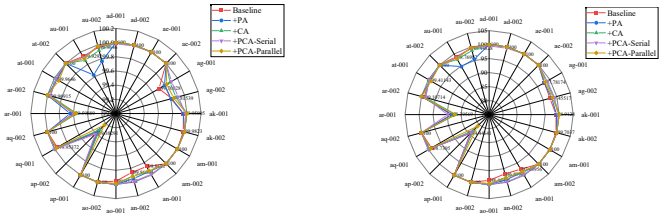
Fig. 11 shows the performance of this module in other indicators, demonstrating the optimal effect by integrating spatial and channel chunking attention through a serial structure. At the same time, the improved model with only Spatial Chunking Attention and Channel Chunking Attention also has some improvement compared to the benchmark model, indicating that this module can further improve the micro recognition ability of the model in the benchmark model with comparatively limited representation ability.

In addition, when evaluating mAUC and mAP, the performance of different categories has been examined separately. From Fig. 12, it can be seen that the baseline structure with the addition of the PCA-Serial module exhibits the optimal level of classification ability for all categories. Table III shows the Wilcoxon signed rank test results for the baseline model and ablation module. Under the two evaluation indicators of AUC and AP, it can be concluded that the PCA-Serial module is significantly superior to the benchmark model, thus proving the effectiveness of the block-based attention module proposed in this paper.

3) Effect of the chunking attention modules combined with multi-scale fusion bottleneck structure on experimental results: To verify the effectiveness of the proposed chunking attention module in feature extraction modules with strong representation capabilities, this paper has been devoted to replacing the original bottleneck structure of Resnet50 with a multi-scale fusion bottleneck structure and using this as a baseline model for ablation experiments of spatial chunking attention (PA) and channel chunking attention (CA). According to the results in Table IV and Fig. 13, the model with the addition of the CA module showed significant improvement in various

TABLE III. THE WILCOXON SIGNED-RANK TEST OF THE CHUNKING ATTENTION UNDER THE ORIGINAL BOTTLENECK STRUCTURE

Baseline	AUC				AP			
vs.	R+	R-	P-value	Sig.	R+	R-	P-value	Sig.
+PA	38.5	66.5	0.3792	No	40	65	0.4326	No
+CA	34.5	70.5	0.2583	No	38	67	0.3627	No
+PCA-Serial	15.5	120.5	0.0066	Yes	35	118	0.0494	Yes
+PCA-Parallel	38	67	0.3624	No	37	68	0.3305	No



(a) Comparison of AUC for each category.

(b) Comparison of AP for each category.

Fig. 12. Comparison of metrics AUC and AP per category in the original bottleneck structure for the chunking attention module.

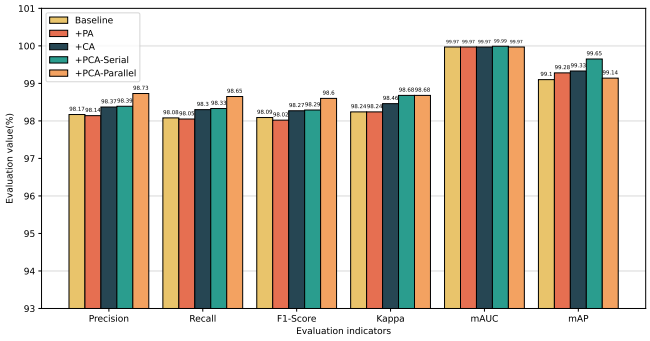
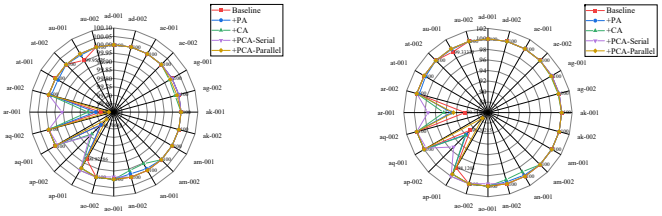


Fig. 13. Evaluation of the chunking attention module for different metrics in the multi-scale fusion bottleneck structure.



(a) Comparison of AUC for each category.

(b) Comparison of AP for each category.

Fig. 14. Comparison of metrics AUC and AP per category in the multi-scale fusion bottleneck structure for the chunking attention module.

indicators, while the PA module had almost no improvement in the model's recognition ability and even had inhibitory effects on certain indicators. It is believed that this suppression effect may be due to the modeling of multi-scale fusion structure and PA modules in the spatial dimension, resulting in mutual redundancy and suppression, exacerbating noise in the spatial dimension, and ultimately losing the recognition ability of some key features. However, in the module that integrates PA and CA, the channel chunking attention introduces much more robust channel information, which offsets the aforementioned suppression effect. Therefore, it further improves the recognition performance in both serial and parallel fusion structures. In particular, the parallel fusion structure further optimizes the inference time and achieves optimal results in all metrics compared to adding PA and CA modules separately.

TABLE IV. COMPARISON OF PARTIAL RESULTS OF THE CHUNKING ATTENTION UNDER THE MULTI-SCALE FUSION BOTTLENECK STRUCTURE

Model	Params (M)	FLOPs (G)	Time (MS)	Top1-acc (%)	Top2-acc (%)
Baseline	36.1	6.535	76.9	98.32	100.00
+PA	49.685	8.485	95.59	98.32	100.00
+CA	62.508	9.102	96.83	98.53	99.79
+PCA-Serial	76.094	11.052	107.95	98.74	100.00
+PCA-Parallel	76.094	11.052	91.46	98.74	100.00

V. DISCUSSION AND CONCLUSIONS

In addition, it was observed in Fig. 14 that the model with PCA-Parallel showed a decrease in AUC and AP in very few ceramic samples, but according to the test results in Table V, the difference between these two indicators was not so significant compared to the baseline model. Based on the performance of various indicators, PCA-Parallel is still the best choice for ceramic microscope image recognition.

1) Effect of different block attention embedding structures on experimental results: Considering the potential impact of PA and CA on different stage blocks of the backbone model in the ceramic microscope image dataset, a series of combined experiments with different embedding structures have been designed to determine the optimal neural network architecture. As shown in Table VII, four embedding methods have been demonstrated: Version A represents embedding chunking attention modules into each stage block, which results in the maximum space and time overhead of the model. Version B only embeds chunking attention in stages like stage-1 and stage-4. It is believed that this structure can effectively establish the remote dependencies between low-level detail features and high-level semantic features. Version C means that all stages except stage-1 are embedded with chunking attention. This way will abandon the modeling of low-level details and focus on the expression of semantic information at different levels of abstraction. The D version only embeds chunking attention in the stage-4 stage, and compared to versions a, b, and c, the D version has the smallest space and time overhead.

As shown in Table VI, this section of the experiments was conducted on the three best-performing models from previous experiments. These models include the one based on

TABLE V. THE WILCOXON SIGNED-RANK TEST OF THE CHUNKING ATTENTION UNDER THE MULTI-SCALE FUSION BOTTLENECK STRUCTURE

Baseline	AUC				AP			
vs.	R+	R-	P-value	Sig.	R+	R-	P-value	Sig.
+PA	10	18	0.4990	No	19	36	0.3862	No
+CA	11	25	0.3264	No	17	28	0.5147	No
+PCA-Serial	3	18	0.1148	No	13	42	0.1381	No
+PCA-Parallel	12	9	0.7532	No	16.5	19.5	0.8334	No

TABLE VI. COMPARE THE RESULTS OF DIFFERENT EMBEDDING METHODS

Model	Version	Params(M)	FLOPs(G)	Time(MS)	Acc(%)	Pre(%)	Rec(%)	F1(%)	Kap(%)
+PCA-Serial	A	86.125	16.066	118.39	97.27	97.29	97.26	97.25	97.14
	B	63.551	8.627	81.37	96.44	96.52	96.30	96.29	96.26
	C	85.521	13.925	91.12	97.06	97.21	96.94	96.93	96.92
	D	62.947	6.485	85.67	96.02	95.71	95.51	95.54	95.82
+Cross Scale +PCA-Serial	A	98.668	18.492	142.79	98.32	98.31	98.25	98.23	98.24
	B	76.094	11.052	107.95	98.74	98.39	98.33	98.29	98.68
	C	98.063	16.35	120.57	97.90	97.70	97.62	97.55	97.80
	D	75.489	8.91	99.46	98.74	98.36	98.35	98.30	98.68
+Cross Scale +PCA-Parallel	A	98.668	18.492	142.58	97.90	97.67	97.61	97.62	97.80
	B	76.094	11.052	91.46	98.74	98.73	98.65	98.60	98.68
	C	98.063	16.35	116.27	97.69	98.03	97.91	97.88	97.58
	D	75.489	8.91	102.64	98.53	98.38	98.24	98.24	98.46

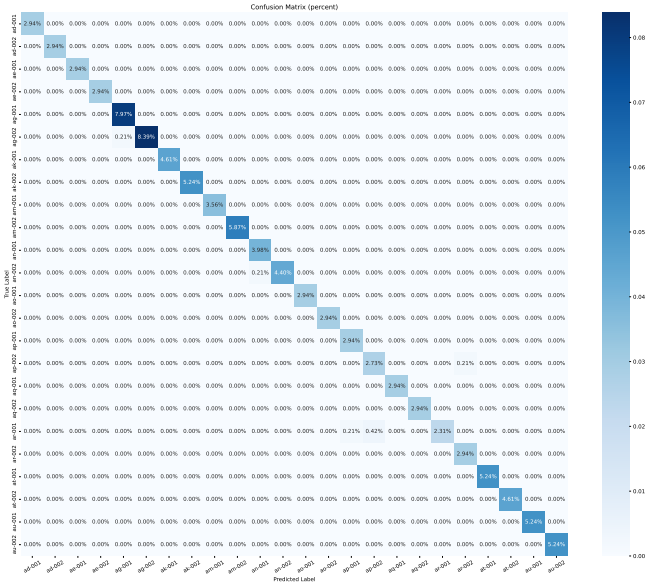


Fig. 15. The confusion matrix of this model on the ceramic microscopic dataset.

TABLE VII. DIFFERENT EMBEDDING METHODS FOR CHUNKING ATTENTION MODULES

Version	Stage-1	Stage-2	Stage-3	Stage-4
A	✓	✓	✓	✓
B	✓	✗	✗	✓
C	✗	✓	✓	✓
D	✗	✗	✗	✓

the original bottleneck structure with the addition of PCA-Serial and the one based on a multi-scale fusion bottleneck structure with the addition of PCA-Serial and PCA-Parallel, respectively. The experimental results are shown as follows:

In the original Resnet50 with the addition of PCA-Serial, the embedding structure of version A performed the best. From the evaluation results of different embedding structures, it can be found that the results of different indicators also show an overall upward trend as the number of chunking attention embeddings increases. This is consistent with the motivation of this study to explore the insufficient recognition ability of basic bottleneck structures. Therefore, chunking attention can effectively alleviate this defect and improve the performance of the model.

In the model with the addition of PCA-Serial and replacement of the multi-scale fusion bottleneck structure, chunked attention did not show a significant ability to improve, with versions B and D performing the best, with version D showing the best level of performance across a wide range of metrics. It is believed that the global modeling capabilities of the chunking attention fusion module and the multi-scale fusion bottleneck structure in a serial structure are equivalent. In some features, there is a coupling relationship between the recognition of these two structures. Therefore, simply modeling high-level semantic features can improve the effectiveness. However, in indicators sensitive to positive and negative samples, there is still some room for improvement in this structure. On the other hand, compared to the original Resnet50 model, this model has achieved an improvement of approximately 1% to 3%, further demonstrating the superiority of the multi-scale fusion bottleneck structure.

Finally, in the model that added PCA-Parallel and replaced the multi-scale fusion bottleneck structure, version B showed the best performance among all versions. Therefore, it is believed that the method of modeling remote dependencies based on both low-level and high-level semantic features proposed in this paper is effective. However, from other versions of this chunking attention fusion method, it can also be found that this method is more sensitive to the recognition ability of other

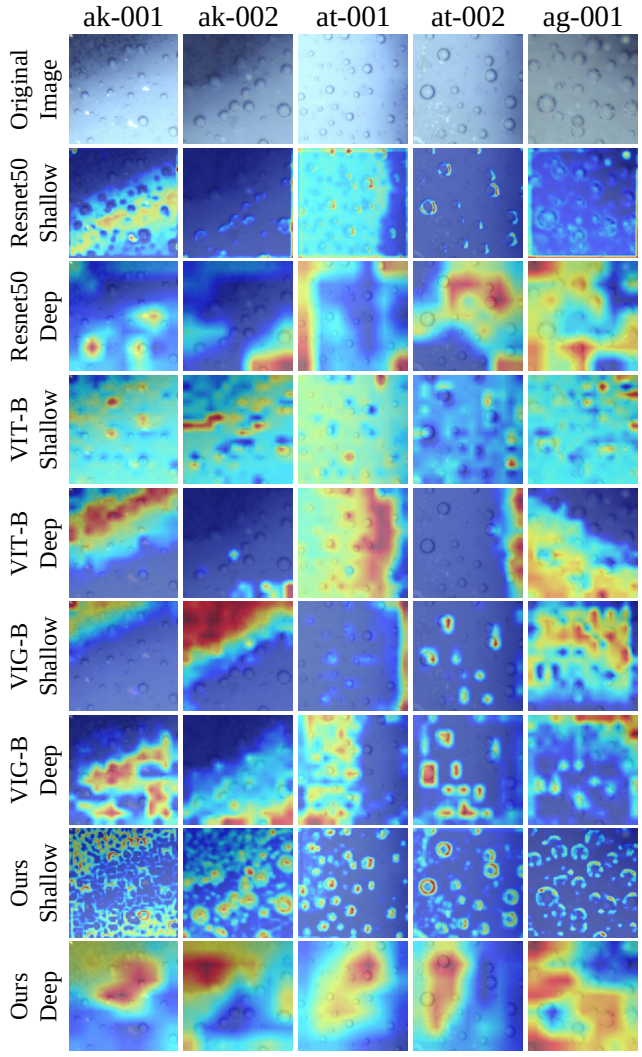


Fig. 16. Compare the Grad-CAM visualization results of deep and shallow modules in different models.

modules. For most application scenarios, there are serial fusion structures have better adaptability. Therefore, when selecting a model, specific data characteristics should be considered.

In summary, by integrating spatial and channel chunking attention modules into the network in different ways, it is very likely to maximize the advantages of each stage block of the model and enhance its performance in semantic feature extraction and recognition. These research results have further expanded the understanding of the attention mechanism of ceramic microscopic image data and provided valuable references for future similar map image classification tasks.

A. Comparative Experiments

In this section, the proposed model has been proposed with the most classic and advanced backbone models in different fields of neural networks in recent years: VGG11[33], WResnet50[30], Resnext50[31], Densenet121[34], SEResnet50[24], Vision Transformer(VIT-B)[3], Conformer[35], MLP-mixer[36], Resnest50[32], Vision GNN (VIG-B)[38] and Hornet[37].

The detailed comparison results are as shown in Table VIII, indicating that the model proposed in this paper has achieved the best performance among all evaluation indicators. Fig. 15 shows the Confusion matrix of this model. It can be seen that all types of models show accurate prediction ability, and the matrix is approximately a diagonal matrix, which indicates that this model has good feature recognition ability. The first four Resnet-type models have shown relatively cutting-edge performance in ceramic microscope image recognition. In recent years, the improvement direction has mainly focused on attention mechanisms. The algorithm of combining chunking attention with multi-scale fusion bottleneck structure in this paper has refreshed the performance of various indicators of this type of model, providing a new baseline for subsequent research.

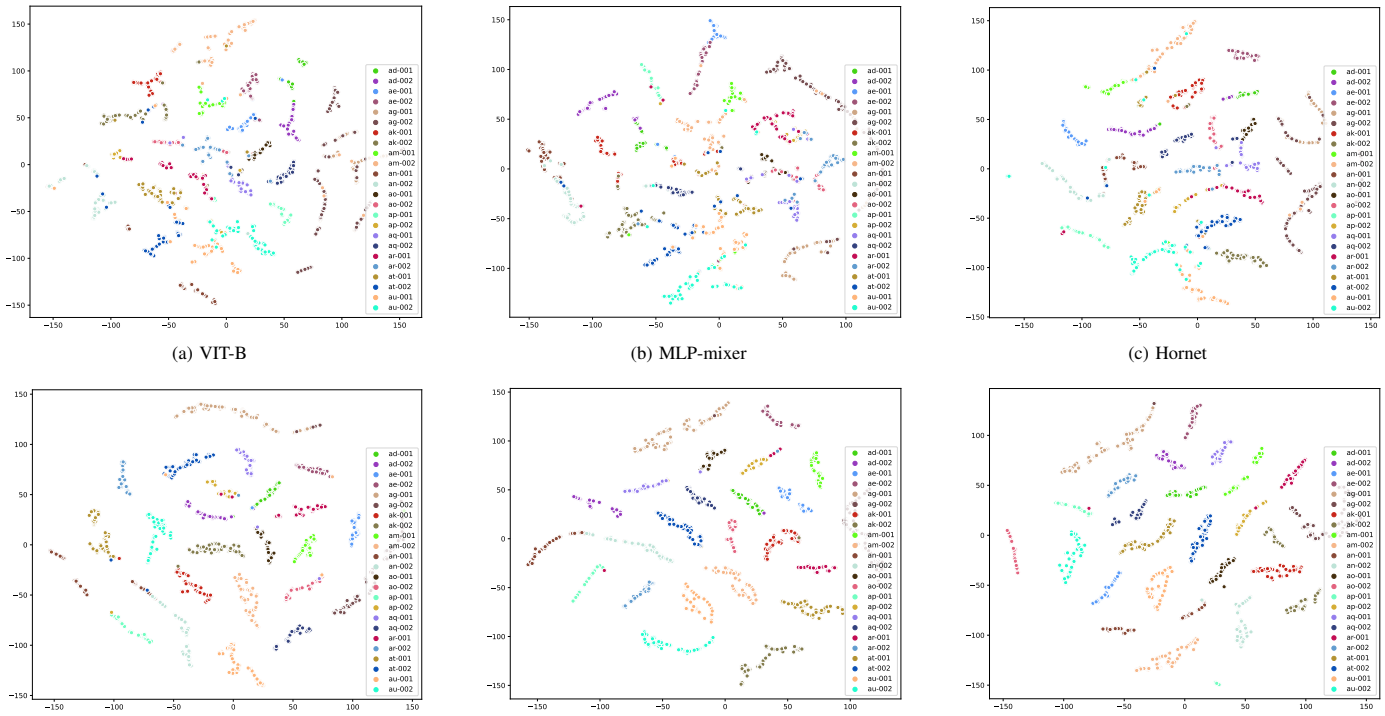
On the other hand, the results show that most models of the non-Resnet type models have a significantly lower recognition ability than this type of model. This is because compared to Resnet-type models, the extraction ability of other models is insufficient when modeling the relationship between global and local features, especially in the complex microstructure of ceramics, which can be amplified. Unlike this, models such as convolutions typically have stronger feature extraction capabilities. For example, convolutional-dependent networks such as Densenet121 and Conformer perform similarly to Resnet-type networks, and our improvement direction only needs to overcome the local dependencies of convolutional operations. Therefore, when dealing with ceramic micro recognition tasks, especially fine-grained image classification problems, priority should be given to neural network models such as convolution.

In addition, it is worth noting that VIG-B, a modeling method based on the relationship between Tokens, still has certain competitiveness in classification. The feature space belonging to this method is different from the Euclidean space adapted by traditional attention mechanisms. The relationship graph structure can make the extracted features more robust, so this is also a research direction worth exploring in future work.

B. Visualization Analysis

To further demonstrate the superiority of this model, this model has been selected, Resnet50, VIT-B, and VIG-B, and five images have been randomly selected from the dataset for Grad-CAM class activation map visualization. The gradient thermal maps have been produced for both shallow and deep modules of each model.

From Fig. 16, it can be observed that in the shallow module section, this model captures the details of ceramic micro images more comprehensively than other models, and can effectively identify fine-grained details, such as bubble features and texture features on the ceramic micro surface. In the task of ceramic anti-counterfeiting recognition, the accurate recognition of this information directly affects the identification results. In contrast, although Resnet50 can capture individual details to some extent, its level of attention is limited. This also reflects the effectiveness of the chunking attention module and multi-scale fusion bottleneck structure in this model. Although both VIG-B based on graph structure and VIT-B based on self-attention can also cover some features, there are problems



(a) VIT-B (b) MLP-mixer (c) Hornet

(d) VIG-B (e) Resnet50 (f) Ours

Fig. 17. Compare the clustering diagrams of ceramic microscopic data in the t-SNE algorithm using different models.

TABLE VIII. COMPARISON OF RESULTS BETWEEN DIFFERENT MODELS

Model	Year	Params(M)	FLOPs(G)	Time(MS)	Acc(%)	Pre(%)	Rec(%)	F1(%)	Kap(%)
WResnet50[30]	2016	66.883	11.425	94.57	95.81	95.76	95.73	95.68	95.60
Resnext50[31]	2016	23.029	4.257	104.94	96.86	96.78	96.61	96.61	96.70
SEResnet50[24]	2018	26.088	4.117	92.41	95.39	95.68	95.63	95.51	95.16
Resnest50[32]	2022	25.483	5.4	101.87	97.90	97.69	97.69	97.65	97.80
VGG11[33]	2014	132.096	7.605	96.49	84.28	84.22	83.59	83.54	83.49
Densenet121[34]	2017	6.978	2.865	78.28	98.11	97.74	97.63	97.58	98.02
VIT-B[3]	2020	85.817	17.582	95.17	76.31	79.61	76.92	76.93	75.13
Conformer[35]	2021	81.226	23.401	115.39	93.08	93.81	93.42	93.38	92.73
MLP-mixer[36]	2021	59.13	12.62	89.47	72.96	72.30	72.24	71.51	71.61
Hornet[37]	2022	86.256	15.583	91.41	73.79	75.67	74.26	74.32	72.47
VIG-B[38]	2022	85.841	17.681	140.29	90.99	91.76	91.66	91.51	90.53
Ours	2023	76.094	11.052	91.46	**98.74**	**98.73**	**98.65**	**98.60**	**98.68**

such as scattered and incomplete focus areas, which affect subsequent feature recognition. In the deep module section, due to the establishment of remote dependency and multi-scale fusion features in this model, it extracts more comprehensive semantic information. At the same time, our model reduces the attention to noise bubbles in some images, which is a key difference from models such as VIG-B, thereby improving the recognition accuracy of the model.

In addition, experiments have also been conducted by using the t-SNE feature clustering algorithm to cluster and visualize the model. It is designed to mainly compare the effects of VIT-B, VIG-B, Resnet50, MLP-mixer, and Hornet models. From Fig. 17, it can be observed that each model achieved certain results in partitioning 24 clusters. In comparison, convolution-based models are more effective in increasing the distance between different categories, whereas non-convolutional frameworks such as VIT-B and MLP-mixer are limited by their lower recognition accuracy and less clear boundaries between the output features. Compared to Resnet50, this model can reduce the spacing within the same cluster, which helps the model better complete feature classification tasks. This result further proves the effectiveness of the chunking attention module and multi-scale fusion bottleneck structure in this model.

Based on Resnet50, this study has proposed a segmented attention module and a multi-scale fusion bottleneck structure to improve the existing network model and applied it to the ceramic microscope image classification task for ceramic anti-counterfeiting. It is found that the current popular universal visual recognition deep learning model has certain limitations in complex ceramic micro feature recognition, and the recognition performance of Token-based models is not as good as that of convolutional-based models. However, the

Convolutional neural network model also has the problem of a limited Receptive field. Therefore, the two improved modules proposed in this study can break through this limitation to a certain extent and further enhance the recognition effect based on a Convolutional neural network.

After experimental verification, this model has improved recognition accuracy by 3.98% compared to the baseline model, and has also shown similar improvements in indicators such as recall rate. In the collected ceramic microscopic image dataset, this model has achieved a recognition accuracy of 98.74%, surpassing the recognition accuracy of mainstream models such as Vision Transformer by more than 20%. This result further confirms the viewpoint that convolution is more suitable for ceramic microscope image recognition tasks.

In summary, this improved model has demonstrated certain advantages in ceramic microscope image recognition and anti-counterfeiting tasks. In future work, it will note that ceramics also contain textual modal information such as place of origin, which may also play a certain role in ceramic recognition. However, effectively integrating data from different modalities and achieving consistent expected results is a relatively challenging challenge in this field. In the next step of our work, we plan to explore how to fuse features of different modalities to further improve the recognition accuracy of the model. We will focus on studying how to effectively integrate text information and image information to achieve more accurate ceramic recognition and anti-counterfeiting targets. At the same time, we also plan to explore the intrinsic characteristics of ceramics to further enhance the level of anti-counterfeiting technology. These works will provide certain assistance and promotion for the development and application of the ceramic anti-counterfeiting field.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of Fujian Province of China (No. 2021J011007), Fujian Provincial Department of Education Undergraduate Education and Teaching Research Project (No. FBJY20230083), Principal's Foundation of Minnan Normal University (KJ19015), the Program for the Introduction of High-Level Talent of Zhangzhou and the National Natural Science Foundation of China (No. 61702239).

REFERENCES

[1] C. Niu and M. Zhang, "Using image feature extraction to identification of ancient ceramics based on partial differential equation," *Advances in Mathematical Physics*, vol. 2022, p. 3276776, Jan 2022.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008, similarity Matching in Computer Vision and Multimedia.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 012–10 022. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html

[5] Q. Li-Ying and W. Ke-Gang, "Kernel fuzzy clustering based classification of ancient-ceramic fragments," in *2010 2nd IEEE International Conference on Information Management and Engineering*, 2010, pp. 348–350.

[6] T. Mu, F. Wang, X. Wang, and H. Luo, "Research on ancient ceramic identification by artificial intelligence," *Ceramics International*, vol. 45, no. 14, pp. 18 140–18 146, 2019.

[7] J. Li, H. Huang, F. Hu, and Y. Ou, "Classification of ceramics based on improved alexnet convolutional neural network," in *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 2022, pp. 1–8.

[8] J. H. Yi, W. Kang, S.-E. Kim, D. Park, and J.-H. Hong, "Smart culture lens: An application that analyzes the visual elements of ceramics," *IEEE Access*, vol. 9, pp. 42 868–42 883, 2021.

[9] A. Chetouani, T. Debroutelle, S. Treuillet, M. Exbrayat, and S. Jesset, "Classification of ceramic shards based on convolutional neural network," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 1038–1042.

[10] O. Chaowalit and P. Kuntitan, "Using deep learning for the image recognition of motifs on the center of sukhothai ceramics," *Current Applied Science and Technology*, vol. 22, no. 2, Jan 2022.

[11] S. Wang, Z. Chen, F. Qi, C. Xu, C. Wang, T. Chen, and H. Guo, "Fractal geometry and convolutional neural networks for the characterization of thermal shock resistances of ultra-high temperature ceramics," *Fractal and Fractional*, vol. 6, no. 10, 2022.

[12] B. Min, H. Tin, A. Nasridinov, and K.-H. Yoo, "Abnormal detection and classification in i-ceramic images," in *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2020, pp. 17–18.

[13] J. D. Hogan, L. Farbaniec, M. Shaeffer, and K. T. Ramesh, "The effects of microstructure and confinement on the compressive fragmentation of an advanced ceramic," *Journal of the American Ceramic Society*, vol. 98, no. 3, p. 902–912, Mar 2015.

[14] A. Aprile, G. Castellano, and G. Eramo, "Classification of mineral inclusions in ancient ceramics: comparing different modal analysis strategies," *Archaeological and Anthropological Sciences*, vol. 11, no. 6, pp. 2557–2567, Jun 2019.

[15] E. Odelli, F. Volpintesta, S. Raneri, Y. Lefrais, D. Beconcini, V. Palleschi, and R. Chapoulie, "Digital image analysis on cathodoluminescence microscopy images for ancient ceramic classification: methods, applications, and perspectives," *The European Physical Journal Plus*, vol. 137, no. 5, p. 611, May 2022.

[16] G. Wan, H. Fang, D. Wang, J. Yan, and B. Xie, "Ceramic tile surface defect detection based on deep learning," *Ceramics International*, vol. 48, no. 8, pp. 11 085–11 093, 2022.

[17] H. Zhang, L. Peng, and G. Lei, "Saliency detection for surface defects of ceramic tile," *Ceramics International*, vol. 48, no. 21, pp. 32 113–32 124, 2022.

[18] Y. Qi, M.-Z. Qiu, H.-Z. Jing, Z.-Q. Wang, C.-L. Yu, J.-F. Zhu, F. Wang, and T. Wang, "End-to-end ancient ceramic classification toolkit based on deep learning: A case study of black glazed wares of jian kilns (song dynasty, fujian province)," *Ceramics International*, vol. 48, no. 23, Part A, pp. 34 516–34 532, 2022.

[19] J. Y. Byeon and K. Y. Jung, "Dual luminescence optimization of ho3+/yb3+/eu3+-doped gd2o3 phosphor prepared by spray pyrolysis for anti-counterfeiting application," *Ceramics International*, vol. 48, no. 23, Part A, pp. 34 837–34 847, 2022.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html

[21] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

[23] X. Yan, Y. Zhang, and Q. Jin, "Chemical process fault diagnosis based on improved resnet fusing cbam and spp," *IEEE Access*, vol. 11, pp. 46 678–46 690, 2023.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018, pp. 7132–7141. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html

[25] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2021.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[27] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018, pp. 7794–7803. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Non-Local_Neural_Networks_CVPR_2018_paper.html

[28] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, "Efficient content-based sparse attention with routing transformers," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 53–68, 02 2021.

[29] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "Davit: Dual attention vision transformers," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 74–92.

[30] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference (BMVC)*, E. R. H. Richard C. Wilson and W. A. P. Smith, Eds. BMVA Press, September 2016, pp. 87.1–87.12.

[31] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1492–1500. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Xie_Aggregated_Residual_Transformations_CVPR_2017_paper.html

[32] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "Resnest: Split-attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2022, pp. 2736–2746. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022W/ECV/html/Zhang_ResNeSt_Split-Attention_Networks_CVPRW_2022_paper.html

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[34] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4700–4708. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html

[35] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local features coupling global representations for visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 367–376. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Peng_Conformer_Local_Features_Coupling_Global_Representations_for_Visual_Recognition_ICCV_2021_paper.htm

[36] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "Mlp-mixer: An all-mlp architecture for vision," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 24 261–24 272. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Paper.pdf

[37] Y. Rao, W. Zhao, Y. Tang, J. Zhou, S. N. Lim, and J. Lu, "Hornet: Efficient high-order spatial interactions with recursive gated convolutions," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 10 353–10 366. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/436d042b2dd81214d23ae43eb196b146-Paper-Conference.pdf

[38] K. Han, Y. Wang, J. Guo, Y. Tang, and E. Wu, "Vision gnn: An image is worth graph of nodes," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 8291–8303. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/3743e69c8e47eb2e6d3afaea80e439fb-Paper-Conference.pdf