Textile & Leather Review

ISSN 2623-6281| www.tlr-journal.com | 10.31881/TLR

High-Precision Image Recognition Technology in Pattern Analysis of Intangible Cultural Heritage Textiles

Xue Bai

How to cite: Bai X. High-Precision Image Recognition Technology in Pattern Analysis of Intangible Cultural Heritage Textiles. Textile & Leather Review. 2026; 9:1162-1190. https://doi.org/10.31881/TLR.2026.1162

How to link: https://doi.org/10.31881/TLR.2026.1162

Published: 28 April 2026

High-Precision Image Recognition Technology in Pattern Analysis of Intangible Cultural Heritage Textiles

Xue Bai

Academy of Fine Arts, Chongqing University of Education, Chongqing 400067, China

cxx545890@163.com

Article

https://doi.org/10.31881/TLR.2026.1162

Received 10 August 2025; Accepted 14 October 2025; Published 28 April 2026

ABSTRACT

Aiming at the problems such as high consumption of computing resources and insufficient feature extraction when processing complex pattern images in the existing methods, based on VGGNet, ResNet-50, EfficientNetV2 and soft attention mechanism, a new type of recognition and classification method for patterns of intangible cultural heritage textiles is proposed. This method effectively enhances the multi-scale feature extraction ability of the network for complex intangible cultural heritage textile patterns by introducing the improved ReLU activation function and the optimised convolutional block structure, and improves the recognition accuracy and computational efficiency. The findings demonstrated that this method had the highest classification and recognition accuracy rate for Shuizumaweixiu, Xiqinciciu, Hamicixiu, Suxiu, Xiangxiu, Shuxiu and other embroiderings in the textile pattern image classification dataset, while the classification accuracy rate for the types of Yuexiu was relatively the lowest. Overall, the research method achieves an average prediction accuracy rate of over 88% for the eight types of pattern images in the textile pattern image classification dataset. Not only does it outperform other advanced models numerically, but it also performs more evenly across various categories. It is demonstrated that the research method has significant adaptability and generalisation capabilities in the field of pattern recognition of intangible cultural heritage textiles. It can effectively address the limitations of conventional deep learning models in complex pattern classification, thereby providing a scientific foundation and technological support for the digital preservation and inheritance of intangible cultural heritage textiles.

KEYWORDS

high precision, image recognition, convolutional neural network, efficientnet, textiles

INTRODUCTION

Intangible cultural heritage is an important component of human civilisation, carrying rich historical, cultural, and social information. As one of the important carriers of intangible cultural heritage, textiles not only reflect the exquisite craftsmanship of traditional crafts in their pattern design, but also contain profound cultural connotations and ethnic characteristics [1]. However, with the acceleration of modernisation, many traditional textile techniques are in danger of becoming extinct, and the preservation and inheritance of intangible cultural heritage textiles (ICHTs) have become the most immediate issue to be addressed [2]. As computer vision and artificial intelligence technology rapidly develop, high-precision image recognition technology offers a novel solution for the automated analysis of ICHT patterns. High precision image recognition technology is an artificial intelligence and computer vision-based advanced technology. Through deep learning (DL) algorithms, it can extract complex features from massive image data and perform accurate classification and recognition [3].

For example, Sabeenian RS et al. used Pseudo-Convolutional Neural Network (P-CNN) for image preprocessing. Subsequently, the detection and defect classification of various fabric types were achieved through the improved Convolutional Neural Network (CNN). The experimental results show that this method performs excellently in terms of sensitivity, specificity and accuracy, and has high testing accuracy for different fabric types [4]. To establish an objective and quantitative methodology for the diagnosis of the state of preservation of cultural heritage paintings, Eom and Lee developed a technological method for diagnosing the state of preservation. This method involved the analysis of the colour space of Buddhist paintings, the calculation of shape information, the area of damage, and the use of the DL algorithm. The accuracy and time advantage of using a high-precision image recognition technology method were confirmed by a comparative assessment of individual deviations by users [5]. In addition, Yalemisew et al. used DL algorithms to analyse and extract cultural elements conveyed by cultural digital images to fully explore and utilise the rich value information contained in cultural heritage images. The outcomes denoted that this method had high efficiency and accuracy in processing large-scale cultural heritage images, and could quickly identify and classify key features of cultural heritage from thousands of images [6]. The aforementioned studies demonstrate the significant potential for high-precision image recognition technology to be broadly

utilised in the domain of cultural heritage protection. This technology is capable of furnishing substantial practical assistance for the digital protection, pattern analysis and cultural inheritance of ICHTs.

As the DL algorithm further develops, the breakthrough of the EfficientNet model in image recognition has provided new possibilities for cultural relic image analysis. EfficientNet, through composite scaling methods, can significantly raise the accuracy and computational efficiency of the model while balancing network depth, width, and resolution [7]. This feature makes it particularly suitable for processing complex and detail-rich image data. For example, Amit and Prabhat proposed a complex road recognition and classification method using the EfficientNet model as the basic framework and unified spatial channel attention. The findings demonstrated that the accuracy of this method in identifying road surfaces reached 99.39%, demonstrating the powerful ability of the EfficientNet model in high complexity image analysis [8]. Alruwaili and Mohamed proposed a fusion-level image recognition model using EfficientNet-B0, EfficientNet-B2, and ResNet-50 models. The findings demonstrated that the accuracy of the model reached 99.14%, further verifying the superior performance of EfficientNet in multi-model fusion [9]. In addition, Ishaq et al. used gradient weighted class activation mapping to raise the interpretability of the EfficientNet model, achieving an average accuracy of 98.6%, providing higher transparency and reliability for complex image analysis tasks [10]. These studies denoted that the EfficientNet model not only had good performance in accuracy and efficiency, but also had good scalability and interpretability.

In summary, numerous researchers have achieved significant results in intangible cultural heritage image recognition, and these methods have effectively improved the efficiency and accuracy in practical applications. However, ICHT patterns often have complex pattern structures, diverse colour combinations, and unique cultural symbolic meanings. This places higher demands on the robustness and generalisation ability of high-precision image recognition technology. Although the EfficientNet model can, to some extent, handle complex image data and improve recognition accuracy (RA), it still has some limitations when dealing with ICHT patterns. For example, the EfficientNet model consumes a large amount of computational resources when processing high-resolution, multi-scale pattern images, making it difficult to meet real-time requirements. In addition, the diversity and regional differences of ICHT patterns also make it difficult for a single model to adapt to all scenes. The soft attention mechanism (SA) can allocate dynamic and continuous weights to different regions of the input features through weighted averaging, effectively suppressing

irrelevant information and thereby enhancing the recognition ability of DL models for important features of images. The VGGNet is widely used in image recognition tasks due to its simple structure, deep feature layer stacking, and good generalisation ability under the design of small convolutional kernels. The ResNet-50, by introducing residual connections, can effectively alleviate the problems of vanishing and degradation of gradients in deep neural networks, enabling the network to learn complex image features at a deeper level. In view of this, based on the Convolutional Neural Network (CNN), the VGGNet and ResNet-50 networks, the researcher innovatively constructed a high-precision image extraction and recognition model based on the improved CNN, and based on this model, methods such as EfficientNetV2 and SA were introduced for improvement, and a textile pattern classification model based on EfficientNetV2-SA was proposed. Through the above design, the research aims to achieve efficient and accurate identification and classification of the deep-level features of ICHT patterns, improve the RA and robustness of the model on different types and styles of patterns, thereby achieving lower computing resource consumption to support digital protection and inheritance applications.

METHODS AND MATERIALS

High Precision Image Extraction and Recognition Model Based on Improved CNN

CNN is a widely used DL model for image recognition tasks. The central tenet of this approach entails the extraction of local features within an image via convolutional layers (CLs). These local features are then subjected to reduction in dimensionality through the utilisation of pooling layers. The ultimate objective is to achieve classification through the application of fully connected layers [11]. The CL can not only extract the low-level features at the bottom layer, but also gradually obtain more abstract and high-level semantic information through layer-by-layer stacking. Pooling operations are mainly used for downsampling. By aggregating the pixel values within a local area (such as maximum pooling or average pooling), the size of the feature map is reduced. The calculation for the two-dimensional convolution is denoted in equation (1).

$$P_j^k = \sum_{i \ni M_j} x_i^{k-1} w_{ij}^k + a_j^k \quad (1)$$

In equation (1), P_j^k denotes the net activation of the j th channel of the CL k, x_i^{k-1} denotes the elements of the convolutional region M_j in the input x, w_{ij}^k and a_j^k represent the elements and bias parameters in the convolutional kernel, respectively. The calculation for the pooling is shown in equation (2).

$$u_j^l = \beta_j^l down\left(y_j^{l-1}\right) + b_j^l \tag{2}$$

In equation (2), u_j^l means the net activation of the j th channel in the pooling layer l, β_j^l means the weight parameter of the pooling layer l, $down(\cdot)$ and b_j^l represent the downsampling function and bias parameter, respectively, and y_j^{l-1} represents the output. The training process of the CNN model includes two stages: forward propagation and backward propagation [12]. The study uses cross-entropy loss (CEL) as the loss function for backpropagation. The method for calculating CEL L is denoted in equation (3).

$$L = -\sum_{i=1}^{k} y_i log\ (p_i) \tag{3}$$

In equation (3), $\mathrm{y_i}$ and $\mathrm{p_i}$ represent the true label and predicted probability, respectively. However, CNN's ability to extract multiscale features is relatively limited, and problems such as gradient vanishing, gradient explosion, and high computational resource consumption are prone to occur during model training [13]. Therefore, the study also proposed an improved CNN structure based on VGGNet and ResNet-50 networks. The schematic diagram of the VGGNet structure adopted by the research is shown in Figure 1.

As shown in Figure 1, the study used VGG-11, VGG-13, and VGG-16 from VGGNet as training networks for pattern recognition of ICHTs. Among them, VGG-11 contains 8 convolutional layers and 3 fully connected layers. The first two convolutional blocks adopt a structure of a single convolutional layer + Max pooling layer (Conv→MaxPool); In the subsequent convolutional blocks, a structure of two consecutive convolutions plus a pooling layer (Conv→Conv→MaxPool) is adopted. VGG-13 consists of 10 convolutional layers and 3 fully connected layers. The structure of the first two convolutional blocks is the same as that of VGG-11, while in the subsequent convolutional blocks, the structure of two consecutive convolutional layers + Max pooling layers (Conv→Conv→MaxPool) is adopted. VGG-16 consists of 13 convolutional layers and 3 fully connected layers. The structure of the first two convolutional blocks is consistent with that of VGG-11 and VGG-13, while

in the subsequent convolutional blocks, a method of consecutive three-layer convolution + Max pooling layer (Conv→Conv→Conv→MaxPool) is adopted.

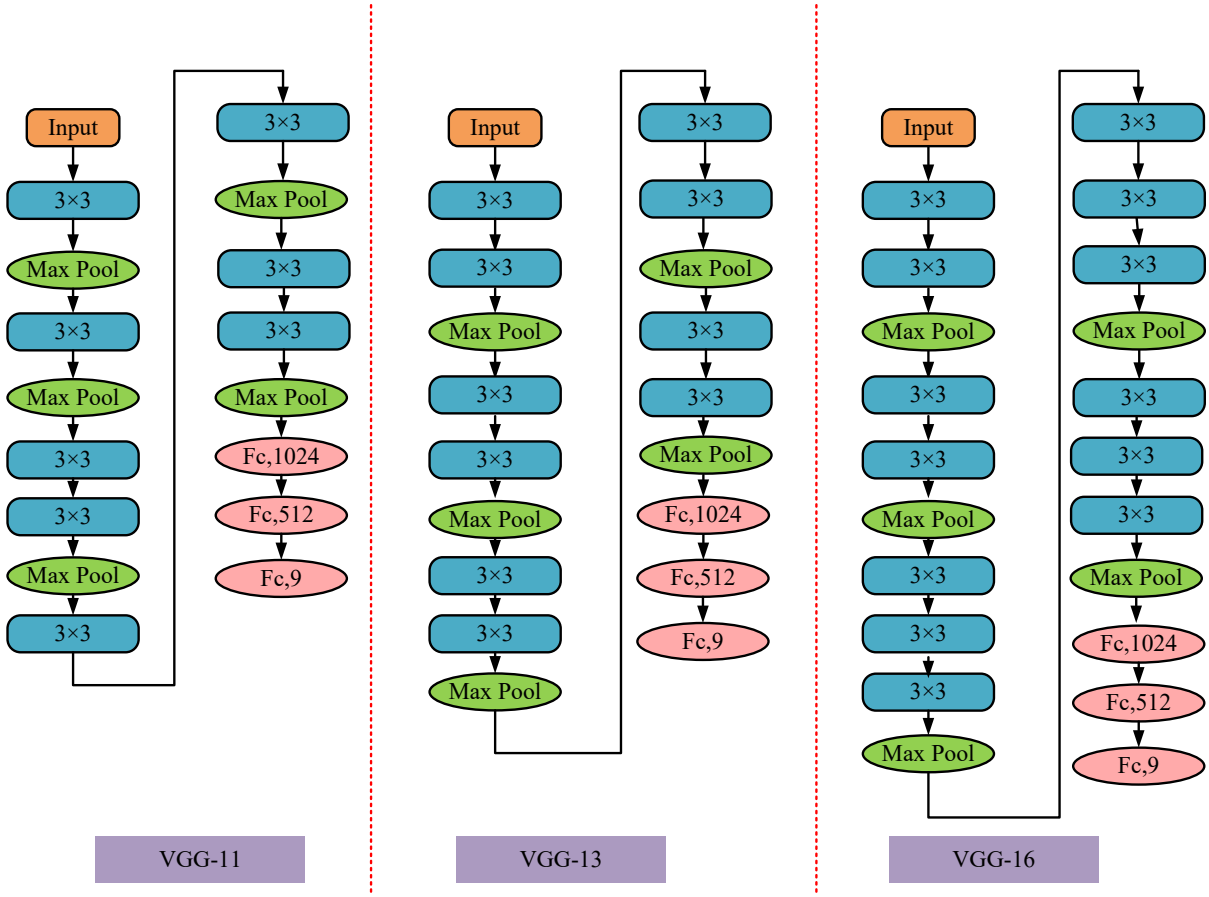

Figure 1. Schematic diagram of the VGG structure

Although VGGNet can effectively improve the effectiveness of CNN models by increasing the number of network layers, the problem of gradient vanishing will gradually worsen with the increase in network layers, leading to a decline in CNN performance [14]. Therefore, the study further introduced ResNet-50 to address this issue. The structures of two residual blocks (RBs) in ResNet-50 are shown in Figure 2.

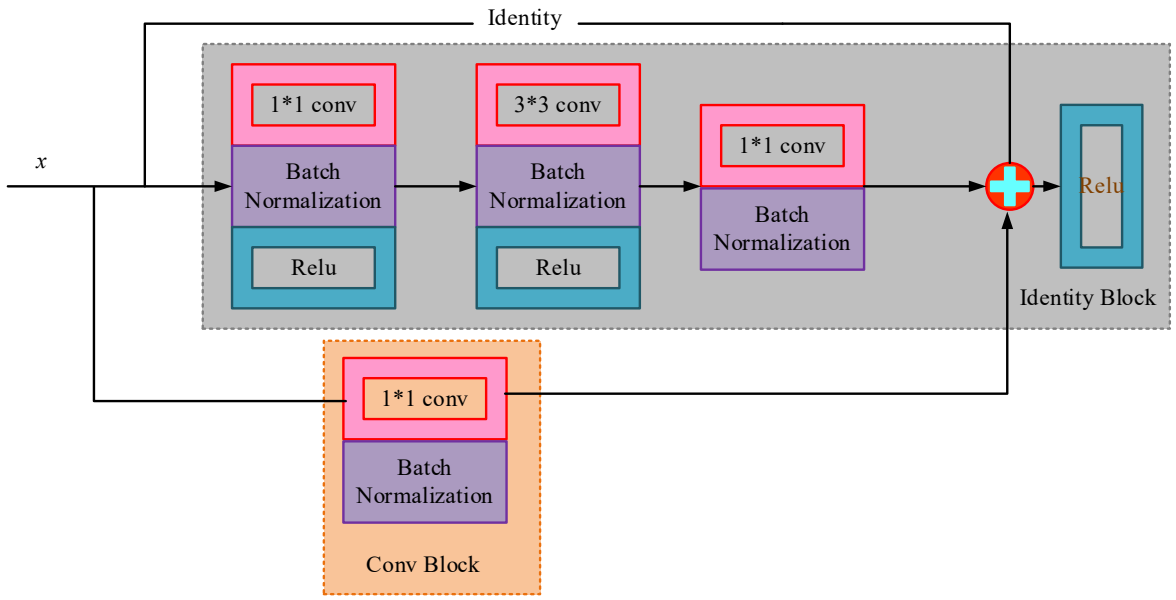

Figure 2. Two RBs' structure of the ResNet-50

In Figure 2, the convolutional RBs and identity RBs of the ResNet-50 are constructed in the order of 1×1 CL, 3×3 CL, and 1×1 CL. Among them, the former is utilised to extract shallow feature information of the image. The latter adds the input information directly to the output information through skip connections, preserving the original input information. The calculation for the ReLU activation function (AF) $f(x)$ is denoted in equation (4).

$$f(x) = \max(0, x) \tag{4}$$

In equation (4), x represents the output value. To further improve the training effectiveness of the ReLU AF, the ReLU AF in the ResNet-50 network is improved, and the IRELU AF is proposed. The formula for the IRELU AF $f_1(x)$ is shown in equation (5).

$$f_1(x) = \max(0, x) + \min(0, \beta^*(\exp(x/\beta) - 1) \tag{5}$$

In equation (5), the β value is 1. The formula for calculating the CL integration $J(t)$ is denoted in equation (6).

$$J(t) = (f * g)(t) = \int f(x) g(t - x_1)\, dx_1 \tag{6}$$

In equation (6), g represents the function, while t and x_1 represent the weights and input values, respectively. Different from the traditional ReLU activation function, the IReLU activation function introduces a fixed non-zero linear term in the negative half-axis part, thereby avoiding all negative inputs being suppressed to zero. This enhances the gradient propagation ability of the activation function in the early stage of training, helps alleviate the vanishing gradient problem, and improves the training stability and model convergence speed. In the CNN structure, although the fully connected layer has a strong feature mapping ability, it also has shortcomings such as a large number of parameters, heavy computational burden and easy overfitting. To this end, the research introduces a global average pooling layer at the end of the network, replacing the two fully connected layers in VGGNet. The global average pooling layer can map high-dimensional features to low-dimensional representations of channel dimensions by averaging the spatial dimensions of each feature map. The calculation of the global average pooling is denoted in equation (7).

$$Y_{ij} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} X_{h,w} \tag{7}$$

In equation (7), for the feature map, Y_{ij} represents the pooled output value, $X_{h,w}$ represents the pixel value, and H and W mean the height and width. Based on the various improvements mentioned above, a high-precision image extraction and recognition model based on an improved CNN is ultimately proposed. Figure 3 denotes the model's network structure.

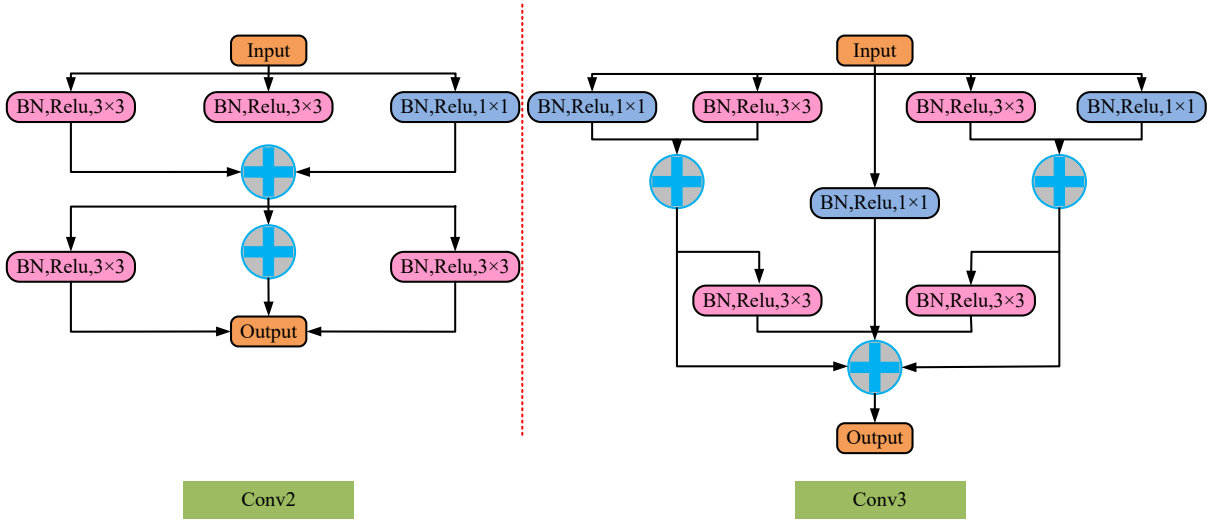

Figure 3. High precision image extraction and recognition model network architecture

In Figure 3, the improved CNN model consists of three different convolutional blocks, which are respectively used for shallow, middle and deep feature extraction. Among them, convolutional block 1 is mainly used to extract low-level texture features, and it is composed of Batch Normalisation (BN) layers, IReLU activation functions, and 3×3 convolutional kernels connected in sequence. Convolutional block 2 is used to capture middle-level features and consists of 4 3×3 convolutional kernels and 1×1 convolutional kernels. The input data is first processed with two 3×3 convolution kernels and one 1×1 convolution kernel. The output results are added element-wise in the channel dimension and then input into the other two 3×3 convolution kernels. The final output is obtained by concatenating all the convolution results in the channel dimension to ensure the richness of mid-level features. Convolutional block 3 is used to extract deep semantic features and consists of 4 3×3 convolutional kernels and 3 1×1 convolutional kernels. The input data passes through one 1×1 convolution kernel, two 3×3 convolution kernels and two 1×1 convolution kernels in sequence. Among them, the output of the first 1×1 convolutional kernel and the output of the first 3×3 convolutional kernel are added together as the input of the third 3×3 convolutional kernel, and the output of the second 3×3 convolutional kernel and the output of the second 1×1 convolutional kernel are added together as the input of the fourth 3×3 convolutional kernel. The final output is composed of the results of the third 3×3 convolutional kernel, the fourth 3×3 convolutional kernel, and the third 1×1 convolutional kernel concatenated in the channel dimension.

Textile Pattern Classification Model Based on EfficientNetV2-SA

The high-precision image extraction and recognition model based on an improved CNN can efficiently extract multi-level features of ICHT patterns, while significantly improving the performance and computational efficiency of the model. However, when handling large datasets, manual annotation of all image samples is both costly and impractical [15]. Consequently, to realise good RA of ICHT pattern images, research has focused on the automatic extraction of discriminative features from images. EfficientNetV2, as an efficient and low-parameter model, performs well in image classification tasks and can effectively solve the above problems [16]. To enhance the categorical features of ICHT pattern images and achieve accurate recognition and classification of pattern images, a textile pattern classification model based on EfficientNetV2-SA was proposed through SA. The SA structure is shown in Figure 4.

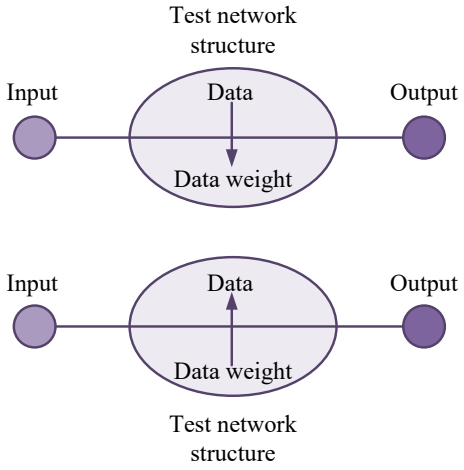

Figure 4. Schematic diagram of the SA structure

As shown in Figure 4, SA mainly focuses on key information and weakens irrelevant information through dynamic weight allocation, while adjusting weights based on input to raise the model's scalability and robustness [17]. The attention weight α_i The calculation is denoted in equation (8).

$$\alpha_i = \frac{\exp\left(s(x_i,q)\right)}{\sum_{j=1}^{N} exp(s(x_j,q))} \tag{8}$$

In equation (8), $s(x_i, q)$ represents the attention scoring function. The process of SA's impact on information is shown in equation (9).

$$SA(x_{1:N}, q) = \sum_{i=1}^{N} a_i \, x_i \tag{9}$$

In equation (9), $x_{1:N}$ represents input feature information, and q represents query information related to the current task of the model. The EfficientNetV2 model consists of three parts: stem, blocks, and head [18]. Among them, Convolutional Block 2 is alternately composed of the MBConv module and the Fused MBConv module, which jointly undertake the task of extracting hierarchical features in the image. The MBConv module reduces the computational complexity by introducing depth-separable convolution while retaining the fine-grained extraction ability of spatial features [19]. The Fused MBConv module uses ordinary

convolution instead of depth-separable convolution, which can better capture local spatial information [20]. The Fused MBConv module is mostly used in the early or middle stage of the network, focusing on improving the convergence speed and training stability of the network. The MBConv module is mainly deployed in the middle and later sections of the network and is used to extract more abstract, advanced semantic features. Through the reasonable combination of these two types of modules, EfficientNetV2 can achieve efficient extraction of multi-scale features while maintaining a low number of parameters, providing sufficient feature support for the subsequent attention weighting mechanism and classification tasks [21]. The EfficientNetV2-SA textile pattern classification model proposed by the research incorporates the SA mechanism after the feature extraction part (stem and blocks) of the EfficientNetV2 model. The specific process is as follows: the input image is first subjected to preliminary feature extraction through the stem section, generating a low-level feature map. Subsequently, the blocks section further extracts high-level features through multiple Mobile Inverted Bottleneck Convolution (MBConv) and Fused-MBConv modules. After feature extraction is completed, the SA mechanism reshapes and weights the extracted feature matrix to strengthen key features. The input of SA is the feature matrix, which is weighted by calculating attention weights to generate a weighted feature matrix. The weighted feature matrix is used as the input for the head part, and classification is achieved through global average pooling and fully connected layers. The schematic diagram of the MBConv and Fused-MBConv module structures is shown in Figure 5.

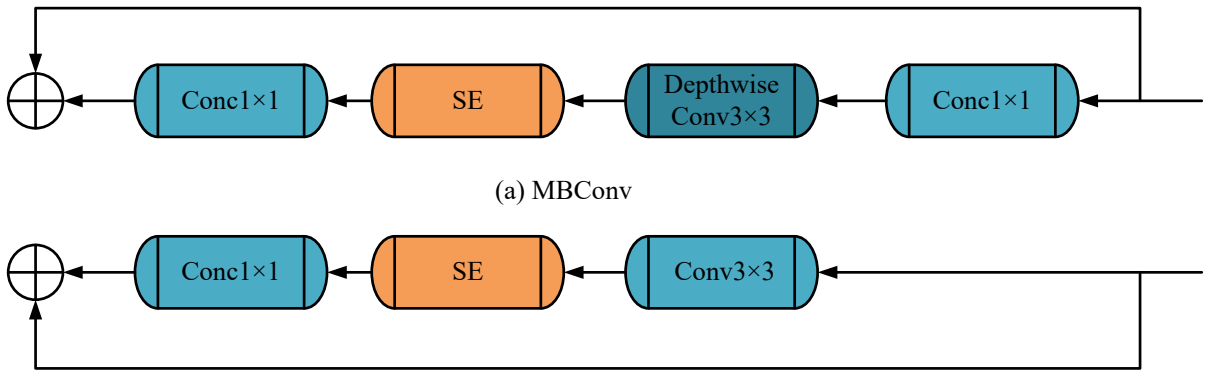

Figure 5. Schematic diagram of the MBConv and Fused-MBConv module structure

In Figure 5 (a), the MBConv module contains two 1×1 CLs, a Squeeze and Excitation (SE) module, and a 3×3 depth separable convolution, and includes residual connections. The input features are first processed through a 1×1 CL, followed by an SE module to expand the number of channels to 4C. Then, a 3×3 depth separable convolution is performed, and finally, the number of channels is restored to C through a 1×1 CL. As shown in Figure 5 (b), the Fused-MBConv module is similar to the MBConv module, but its 3×3 convolution is a regular convolution rather than a depthwise separable convolution, and also includes the SE module and residual connections. The input features are first passed through a 1×1 CL, then expanded to 4C through the SE module, followed by a 3×3 convolution, and finally restored to C. The process structure of the EfficientNetV2-SA textile pattern classification model is shown in Figure 6.

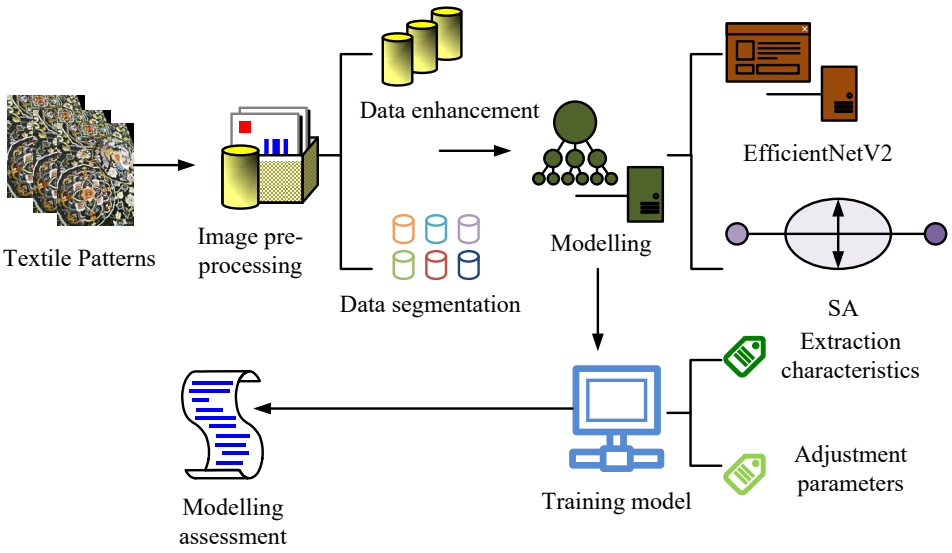

Figure 6. EfficientNetV2-SA model textile pattern classification process

According to Figure 6, the EfficientNetV2-SA model for textile pattern classification consists of six main steps. Firstly, the collected images of ICHT patterns are segmented and denoised to highlight the categorical features of the patterns. The segmentation operation separates the pattern areas in the image from the background, while the denoising operation removes noise from the image through Gaussian filtering to ensure the quality of the input data. To enhance the model's generalisation ability, a data augmentation method based on extension augmentation was adopted to expand the pattern data of ICHTs. Subsequently, based on the EfficientNetV2 model, SA was introduced to weight the extracted feature matrix and reshape

it. Next, the network parameters of the EfficientNetV2-SA model were adjusted. Then, the Neural Architecture Search (NAS) technology was adopted to adjust the network parameters of the EfficientNetV2-SA model. The NAS search space includes the size of the convolutional kernel, the number of convolutional layers, the number of channels, the pooling method, the type of activation function, and the number of neurons in the fully connected layer. During the search process, the controller network generates candidate network structures and conducts performance evaluations on the subset training data. Through iterative update strategies, it finds the network structure that performs best on the validation set. To balance search efficiency and accuracy, NAS adopts a hierarchical search strategy: first, it optimizes the number of convolutional layers and channels in the shallow convolutional blocks, then optimizes the number of convolutional layers, the size of convolutional kernels and the pooling method in the deep convolutional blocks, and finally fine-tuning the number of neurons and the type of activation function in the fully connected layers to achieve adaptive optimisation of the network structure for the pattern recognition task. The preprocessed and enhanced data are input into the EfficientNetV2-SA model for training. During the training, the CEL function is used as the optimisation objective, and the network parameters are updated through the backpropagation algorithm. Meanwhile, validation sets are used to optimise the model and prevent overfitting. Finally, the effectiveness of the EfficientNetV2-SA model is comprehensively assessed using the classification accuracy of ICHT patterns as the main evaluation indicator. The formula for calculating classification accuracy is denoted in equation (10).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

In equation (10), TP and TN denote the truly predicted positive and negative samples, while FP and FN stand for the falsely predicted negative and positive samples. The F1 value is an indicator that combines Precision and Recall, and is suitable for evaluating the performance of models when there is an imbalance in categories. The formula for calculating the F1 value is shown in equation (11).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

In equation (11), Precision is the proportion of samples that are actually positive among those predicted as positive by the model; Recall refers to the proportion of actual positive samples that are correctly predicted as positive by the model. Based on the above various designs, the research ultimately proposed a new method for identifying and classifying patterns of intangible cultural heritage textiles. This method first uses an improved CNN network to extract the shallow, middle and deep features of the image. Subsequently, the extracted features were input into the EfficientNetV2-SA model. Through the attention mechanism, the key pattern information was strengthened, achieving high-precision recognition and classification of the patterns on intangible cultural heritage textiles.

RESULTS

Analysis of Dataset Characteristics

The Nantong Blueprint Pattern Recognition Dataset and Suzhou Silk Pattern Database were used as test data sources. The training set and test set were divided in the ratio of 8:2. Among them, the Nantong Blue Print Pattern Recognition Dataset is sourced from the public data of Nantong Museum. It contains approximately 4,500 high-resolution images, which are classified into 30 typical patterns. It includes geometric patterns (such as the loop pattern, rhombus pattern, bagua pattern, etc.), plant patterns (such as plum, orchid, bamboo, chrysanthemum, lotus, grapevine, etc.), animal patterns (such as fish, phoenix, deer, crane, etc.) and totem elements with folk cultural characteristics (such as Tai Chi, longevity and happiness patterns, door gods, etc.), which have strong regional characteristics and cultural value. The Suzhou Silk Pattern Database is sourced from the public data of the Suzhou Silk Museum. It includes images of silk fabric patterns from multiple historical periods, with a data volume of approximately 5,100 pieces, covering various weaving types such as Song brocade, Yun brocade, satin, jacquard, and decorative patterns. The images include various styles such as realistic flowers, birds, beasts, auspicious birds, religious themes, and mythological stories. The patterns are fine and complex, and the colour layers are rich, fully demonstrating the aesthetic characteristics and technical depth of traditional Chinese silk weaving art. Some of the images in the database are also accompanied by corresponding historical tag information, such as dynasty, use, weaving process, etc., which facilitates further research on the evolution and classification of pattern styles. Some of the dataset images are shown in Figure 7.

(a) Plant pattern (b) Animal pattern (c) Geometric pattern

Figure 7. Dataset image (Source from: Picture (a): https://colorhub.me/photos/alGBm; Picture (b): https://colorhub.me/photos/e7RVB; Picture (c): https://colorhub.me/photos/xGXKo)

During the experiment, all images underwent standardised preprocessing. First, size normalisation processing was performed, uniformly adjusting all images to 224×224 pixels to ensure the consistency of the input data. Subsequently, the image was adjusted for grayscale and normalised for colour. By scaling the pixel values to the [0, 1] interval, the model could converge more quickly during the training and the impact of brightness differences between different images could be mitigated. In addition, to highlight the structural features of the patterns, the study applied Gaussian filtering denoising to the images to remove random noise generated by shooting or scanning, thereby preserving the main texture information of the patterns themselves. Meanwhile, to enhance the generalisation ability and anti-interference performance of the model, the study also carried out data augmentation processing on the training set images, including random cropping, horizontal flipping, ±15° rotation, and mild scaling. For instance, for an image of a Shuxiu flower pattern, random cropping can generate multiple sub-images of the centre or edge area of the flower, enabling the model to capture local features such as petal texture, leaf texture and embroidery lines. Rotation and flipping operations can turn the originally right-facing branch and leaf structure to the left, helping the model maintain its recognition ability for symmetrical or directionally changing patterns on both sides. Scaling can enhance the model's ability to recognise patterns of different scales.

The selection of the model should be based on the inherent characteristics of the data. After constructing the dataset, the study first conducted quantitative and qualitative analyses of the key visual features of the pattern images of intangible cultural heritage textiles to verify the rationality of the model architecture.

Intangible cultural heritage patterns (such as the gold plate embroidery of Cantonese embroidery and the scattered stitch technique of Suzhou embroidery) contain a large number of fine and irregular local texture patterns. It indicates that the model requires a powerful local feature extraction capability, verifying the rationality of choosing VGGNet and ResNet-50 as the infrastructure. Meanwhile, the pattern images also include macroscopic layouts (such as the overall shapes of dragons and phoenixes) and microscopic details (such as the lustre and direction of the silk threads). This feature requires the model to have multi-scale perception capabilities. The EfficientNetV2 model can efficiently process multi-scale images by uniformly scaling the depth, width and input resolution of the network through compound scaling, and thus has been selected as the backbone network. In addition, different types of patterns (such as the realistic animal patterns of Xiang embroidery and Shu embroidery) may be similar on a macro level, but their distinctive features often exist in specific local areas (such as the direction of the stitch and the transition of colours). This uneven distribution of features requires the model to adaptively focus on key areas. Therefore, it is necessary to introduce an attention mechanism. SA can dynamically assign weights to different image regions, highlighting the most important features for classification. From the perspective of data scale, the dataset adopted in the research is of medium scale. Directly using ultra-large models with too many parameters (such as ViT-Huge) is very likely to lead to overfitting. Therefore, the research selects EfficientNetV2 as the network architecture and compensates for the relative insufficiency of data volume through data augmentation. This is a model selection decision driven by data scale.

Performance Testing of High-Precision Image Extraction and Recognition Model Based on Improved CNN

A proper test environment was in place for the verification of the model's performance. The experimental environment and model parameter Settings are shown in Table 1.

Table 1. Experimental environment and model parameter settings

Serial number	Type	Settings
1	Operating system	Windows 10, Python 3.8
2	CPU	Intel Core i7
3	GPU	NVIDIA GeForce
4	Memory	32GB
5	Learning framework	TensorFlow2.3.0

6	Batch size	50
7	Learning rate	0.001
8	Maximum number of iterations	400
9	Loss function	Cross entropy
10	L2 regularisation	0.0002
11	Optimizer	Adam

In Table 1, the study first conducted ablation testing with RA as the indicator, and the test outcomes are shown in Figure 8.

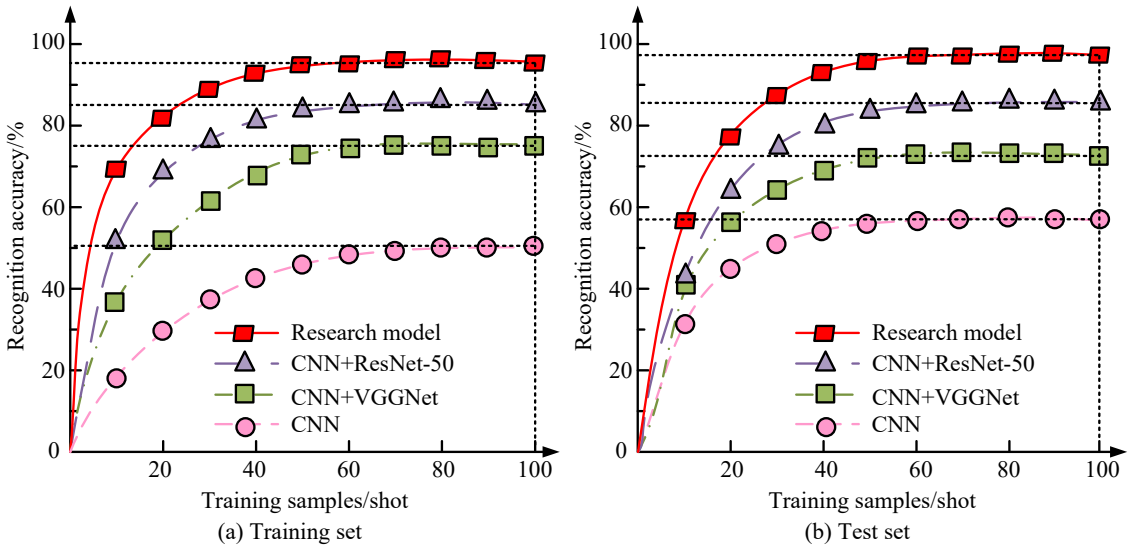

Figure 8. The ablation test results of the research model

Figures 8 (a) and 8 (b) respectively show the recognition accuracy performance curves of each module of the high-precision image extraction and recognition model based on the improved CNN in the training set and the test set. In Figure 8, in contrast to the basic CNN model, the enhanced CNN model introduced by VGGNet increased the RA of the training and testing sets by 25.2% and 15.0%. After further introducing ResNet-50 and IRELU AFs for improvement, the RA of the final model training set and test set arrived at 94.63% and 98.04%, which were 44.6% and 36.5% higher than the basic CNN model. The obtained data indicates that the improvement of various parts of the CNN model has enhanced the model's recognition ability for various textile patterns in the Nantong Blue Print Pattern Recognition Dataset and Suzhou Silk Pattern Dataset, proving the performance of the raised method. In addition, the study also brought in advanced models of

the same type as comparison models, namely Dense Connected Convolutional Networks (DenseNet), Lightweight Convolutional Neural Network (LCNN), and Efficient Convolutional Neural Network (ECNN). Performance tests were conducted based on computational efficiency. The study conducted 20 independent calculations in both the training set and the test set. The main reasons for choosing 20 calculations are as follows. Firstly, through pre-experiments, it was found that fewer than 10 calculations would cause the calculation time to be significantly affected by accidental factors (such as system load fluctuations, GPU resource occupation, etc.), making it difficult to stably reflect the model performance. Secondly, although increasing the number of calculations can further reduce accidental fluctuations, the calculation cost was relatively high and had a limited impact on the conclusion. Therefore, considering the experimental accuracy and computational cost comprehensively, the study ultimately adopted 20 independent calculations to ensure the reliability and repeatability of the results. The test results are shown in Figure 9.

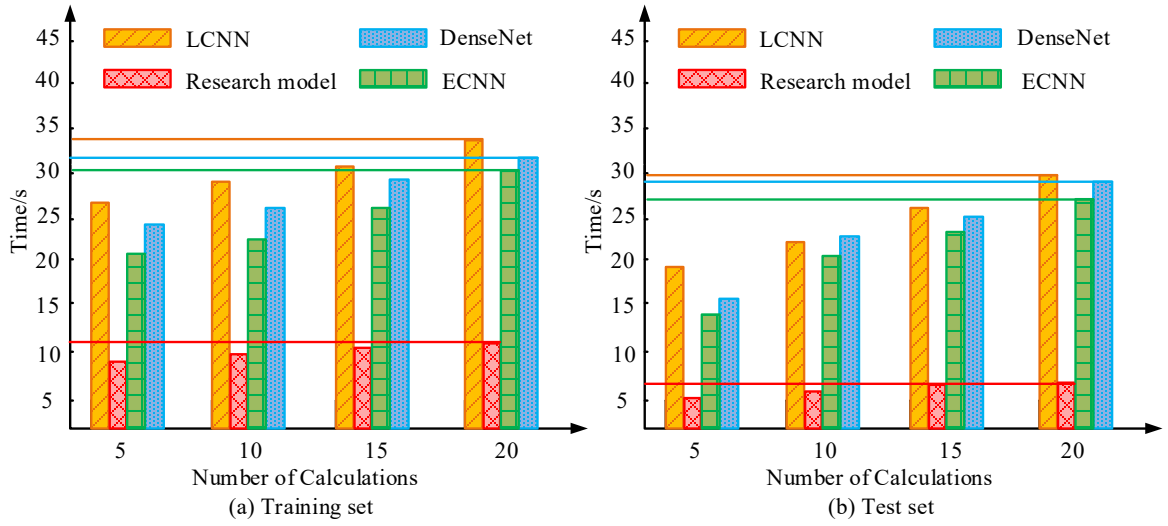

Figure 9. Comparison results of the computational efficiency of different models

Figures 9(a) and 9(b) illustrate the results of a comparative analysis of the computational efficiency of different image recognition models in the training and test sets, respectively. In Figure 9, as the calculation times increased, it was demonstrated that there was a gradual upward trend in the duration taken by all four models to complete their calculations. Following the completion of 20 calculations, the computational times of the LCNN, DenseNet, ECNN, and the research model in the training set amounted to 34.6 s, 32.2 s, 30.4 s,

and 11.3 s, respectively. In the test set, they were 29.9s, 28.2s, 27.1s, and 7.0s, respectively. From this, in both sets, the high-precision image extraction and recognition model proposed by the study had the highest computational efficiency. The above experimental data indicate that the research model can effectively reduce computational overhead while ensuring RA, and has strong practical value and application potential. To verify the effectiveness of IReLU, the study also conducted comparative experiments with standard ReLU and common variants (Leaky ReLU, PReLU, ELU), and the test results are shown in Table 2.

Table 2. Comparison results of different ReLU variants

Activation function	Training Accuracy (%)	Testing Accuracy (%)	Convergence Epoch	Training Time (s)
ReLU	88.12	85.46	75	12.4
Leaky ReLU	90.37	87.25	68	13.1
PReLU	91.64	88.93	63	13.7
ELU	92.25	89.54	61	14.2
IReLU	95.08	92.37	52	11.6

As can be seen from Table 2, under the same experimental environment, different activation functions have a significant impact on the performance of the model. The accuracy rates of traditional ReLU in the training set and the test set are 88.12% and 85.46% respectively. Although Leaky ReLU and PReLU have certain improvements in accuracy, their convergence speed and computational efficiency are still limited. ELU has enhanced its nonlinear feature expression ability, with the test accuracy rate increasing to 89.54%. However, its training time has relatively increased. In contrast, the IReLU proposed by the research institute is significantly superior to other activation functions in both training accuracy and test accuracy, reaching 95.08% and 92.37% respectively. Moreover, it also performs the best in convergence speed and training time. Convergence can be completed in just 52 epochs, with an average training time of 11.6 seconds. This indicates that IReLU not only enhances the recognition accuracy of the model on complex pattern images but also has advantages in training efficiency and stability, thereby verifying its application value and rationality in improving CNN models.

EfficientNetV2-SA Textile Pattern Classification Model Simulation Test

Due to the limitations of existing datasets in terms of image quality, category richness, and annotation completeness, the study adopted an active collection strategy. Using OpenCV technology, eight different categories of ICHT pattern images were collected from the internet, covering images of China's four famous embroidery patterns and unique embroidery patterns of ethnic minorities. To construct a comprehensive dataset, the study also conducted detailed data preprocessing on these images, including data cleaning, data normalisation, classification, and augmentation. After expansion, the dataset finally contained 10025 images, which were then divided into training and testing sets in an 8:2 ratio to support model training and performance evaluation. The study first conducted ablation tests on the proposed final model, and the test results are shown in Table 3.

Table 3. Ablation test results

Model	Accuracy/%	F1/%	Inference time/s
CNN	74.35	71.82	0.067
CNN+VGGNet	83.21	80.57	0.091
CNN+VGGNet+ResNet-50	88.62	86.93	0.104
CNN+VGGNet+ResNet-50+IReLU	92.15	91.02	0.108
EfficientNetV2	94.26	93.44	0.089
EfficientNetV2-SA (Final model)	97.30	96.65	0.092

From Table 3, with the gradual enhancement of the model structure, significant improvements have been achieved in both classification accuracy and F1 value. Among them, after the introduction of VGGNet, the accuracy rate of the model increased from 74.35% to 83.21%, indicating that the optimisation of the shallow structure has a positive effect on the extraction of basic image pattern features. After further combining the ResNet-50 residual structure, the accuracy rate increased to 88.62%, indicating that cross-layer information fusion can significantly enhance the deep semantic understanding ability of the model. The addition of the IReLU activation function made the gradient transfer of the model more stable during the training process, and the accuracy rate and F1 value increased to 92.15% and 91.02% respectively. Subsequently, after replacing the model backbone with EfficientNetV2, the model further improved the RA while maintaining a high computational efficiency. Finally, after introducing the SA, the model achieved the optimal

performances of 97.30% and 96.65% in terms of accuracy and F1 value, respectively. Meanwhile, the reasoning time was controlled within 0.092 seconds, indicating that this mechanism can effectively enhance the model's perception ability of the key features of complex patterns, and almost no additional computational overhead was added. The validity and efficiency of the EfficientNetV2-SA model in the task of image classification of ICHTs were verified. Besides, the study also introduced the LCNN, ECNN, and Compressed Convolutional Neural Network (GIST) models, and conducted performance tests using classification accuracy, recall, and specificity as indicators. The test outcomes are denoted in Table 4.

Table 4. Multi-metric performance test results for different models

Style	Model	Precision/%	Recall/%	Specificity/%
Suxiu	LCNN	73.3	75.5	80.2
	ECNN	77.2	79.3	81.6
	GIST	80.3	88.2	83.4
	Research model	95.4	97.8	95.2
Xiangxiu	LCNN	65.2	66.3	89.6
	ECNN	68.7	70.4	90.2
	GIST	70.2	71.6	91.6
	Research model	97.3	94.2	99.7
Yuexiu	LCNN	67.6	69.6	53.8
	ECNN	76.3	80.4	60.5
	GIST	81.6	79.5	70.8
	Research model	95.1	90.8	95.6
Shuxiu	LCNN	68.8	70.5	55.6
	ECNN	72.5	78.3	58.3
	GIST	80.4	83.4	60.7
	Research model	92.6	96.5	90.4
Hamicixiu	LCNN	64.5	66.0	75.3
	ECNN	68.2	70.1	77.5
	GIST	72.4	74.6	79.8
	Research model	94.5	96.2	93.6
Xiqinciciu	LCNN	63.8	65.5	71.2
	ECNN	67.1	69.0	73.5
	GIST	70.9	73.2	75.8
	Research model	93.8	95.1	92.7
Shuizumaweixiu	LCNN	66.0	67.2	68.4
	ECNN	69.8	71.5	70.9
	GIST	74.3	76.4	73.6

Other embroidery	Research model	95.0	96.8	94.2
	LCNN	60.5	62.0	65.7
	ECNN	64.1	66.5	68.2
	GIST	68.2	70.1	71.4
	Research model	91.5	93.7	92.3

According to Table 4, the EfficientNetV2-SA textile pattern classification model proposed by the research showed significant advantages in the classification tasks of Xiangxiu, Yuexiu, and Shuxiu patterns. Specifically, the accuracy, recall, and specificity of this model in the classification of Xiangxiu patterns reached 97.3%, 94.2%, and 99.7%, respectively. Not only was it numerically superior to other CNN models, but its performance among different categories was also more balanced. This result indicated that EfficientNetV2-SA had strong generalisation ability and robustness in processing the classification task of ICHT patterns, effectively improving the adaptability and classification performance of the model in complex pattern recognition tasks. The confusion matrix obtained on the textile pattern image classification dataset, pre- and post-improvement, is shown in Figure 10.

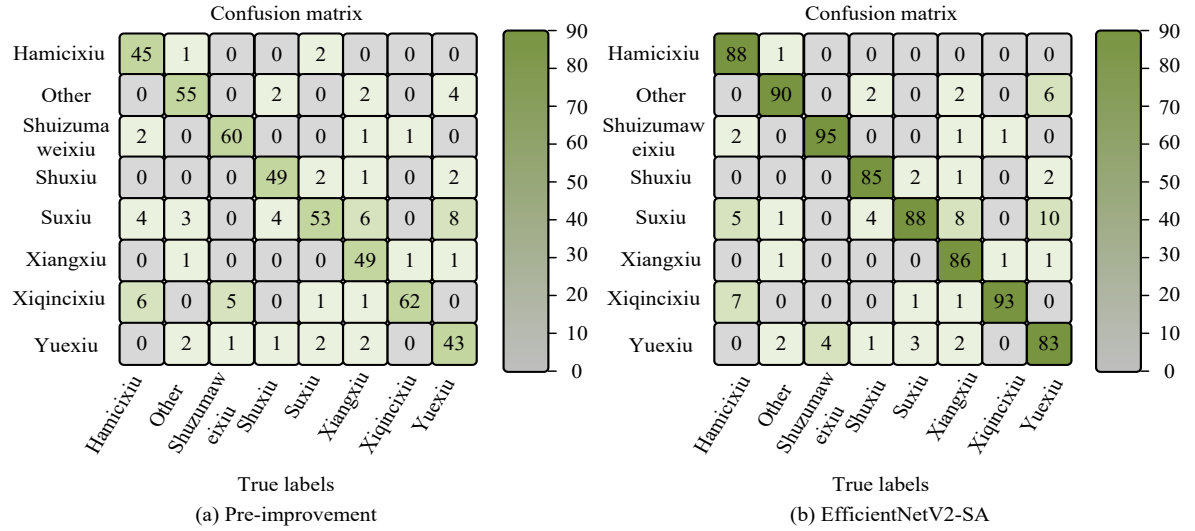

Figure 10. Confusion matrix plots before and after model improvement

Figures 10 (a) and 10 (b) show the confusion matrix before and after model improvement, respectively. From Figure 10, the EfficientNetV2-SA model had the highest classification and RA for Shuizumaweixiu, Xiqinciciu, Hamicixiu, Suxiu, Xiangxiu, Shuxiu, and other embroidery types in the textile pattern image classification

dataset, while the classification accuracy for Yuexiu types was relatively low. Overall, the EfficientNetV2-SA model achieved an average prediction accuracy of over 88% for 8 types of pattern images in the textile pattern image classification dataset. The reason for this is that Yuexiu patterns often use multi-level colour overlay and delicate needlework, resulting in more complex texture features. Compared to other embroidery categories, it was challenging for this model to reliably extract stable key features. From this, the EfficientNetV2-SA model had good performance overall, but there was still room for optimisation when dealing with Yuexiu patterns. The study made a comparison of the classification outcomes obtained by the four models when operating under varying levels of interference within the textile pattern image classification dataset. The outcomes of this comparison are presented in Table 5.

Table 5. Comparative experimental study of various components in the enhanced model

Interference types	LCNN	ECNN	GIST	Research model
Adding random noise	52.3%	60.1%	65.3%	92.3%
Adding salt-and-pepper noise	52.3%	60.2%	65.3%	94.0%
Adding Gaussian noise	52.5%	50.3%	54.3%	92.6%
Histogram equalisation	52.6%	60.0%	64.4%	92.6%
Wiener filtering	52.3%	59.9%	65.8%	94.0%
Median filtering	51.9%	59.3%	65.6%	92.4%
Sharpening	51.6%	59.4%	63.8%	92.8%
Wavelet compression	52.8%	60.1%	64.3%	94.1%
Resampling and subsequent difference	52.8%	60.2%	65.5%	94.2%
Shearing 1/8	52.6%	60.3%	65.4%	94.4%

According to Table 5, under ten various types of interference, the EfficientNetV2-SA textile pattern classification model was still able to maintain over 92% classification accuracy. The RA of the LCNN, ECNN, and GIST models decreased significantly under the same conditions. To further verify the practical effectiveness and reliability of the proposed EfficientNetV2-SA textile pattern classification model, the study invited five experts with backgrounds in ICHT research and rich experience in embroidery techniques to form a review panel to conduct a manual review of the model's prediction results. The review process first required experts to independently assess the matching degree between the model's predicted labels and the image content. Then, they scored the prediction results based on the historical, regional, and craftsmanship characteristics of the patterns (out of 5 points, with 5 indicating a complete match and 1 indicating a

complete mismatch). Finally, they calculated the consistency between the expert scores and the model's prediction results. The accuracy rate of expert review and Cohen's Kappa coefficient were calculated. The relevant test results are shown in Table 6.

Table 6. Expert review consistency statistics

Types	Expert scoring	Accuracy rate of expert review/%	Cohen's Kappa
Suxiu	4.6	92.0	0.88
Xiangxiu	4.4	90.8	0.85
Yuexiu	4.5	91.4	0.86
Shuxiu	4.5	89.6	0.83
Hamicixiu	4.3	88.7	0.82
Xiqinciciu	4.2	87.1	0.80
Shuizumaweixiu	4.3	85.0	0.81
Other embroidery	4.2	88.4	0.78

As shown in Table 6, the average scores given by experts to various types of images were all above 4.2 points, the accuracy rate of expert review was between 85% and 92%, and the Cohen's Kappa coefficient was between 0.78 and 0.88. This indicated that the model's prediction results were highly consistent with the experts' judgments and could meet the application requirements of pattern analysis and classification of ICHTs.

In summary, the research method can effectively extract and analyse the complex structures, multi-scale features and local detail differences existing in the patterns of ICHTs, ensuring the integrity and expressiveness of the pattern features. Moreover, when processing pattern images, both texture information and geometric shape features are taken into account, which can adapt to the minor changes and local deviations that occur during the manual production of ICHTs, thereby improving the accuracy and robustness of recognition. Finally, through comparison with traditional methods, it can be seen that the research method has obvious advantages in terms of accuracy, recall rate and comprehensive performance indicators, fully demonstrating its applicability and practical value in the recognition of patterns on ICHTs. Overall, the research method not only meets the technical requirements for the identification of intangible cultural heritage patterns but also provides reliable technical support for related protection and digital research.

CONCLUSION

From ancient silk to traditional embroidery, textile patterns reflect the social styles, aesthetic concepts, and craftsmanship levels of different historical periods. However, due to the fragility and vulnerability of textile cultural relics, the protection and research of their patterns face enormous challenges. In view of this, the research proposed a novel method for the patterns of ICHTs by improving CNN and introducing the EfficientNetV2 model. To verify the validity of the model, the study used the Nantong blue and white printed cloth pattern recognition dataset (approximately 4500 images) and the Suzhou silk pattern database (approximately 5100 images) as data sources. Through active collection and expansion, a comprehensive dataset was constructed, totalling 10025 images, for model training and testing.

Due to their simple structure and limited sensing field, lightweight models such as LCNN and ECNN have difficulty effectively capturing the complex textures and detailed features across scales in intangible cultural heritage patterns. The improved CNN model effectively alleviates the vanishing gradient problem and enhances the model's ability to capture the detailed features of patterns by integrating the deep feature extraction ability of VGGNet and the residual connection mechanism of ResNet-50. The ablation test results show that after introducing VGGNet and ResNet-50, the model accuracy has increased by 25.2% and 15.0% respectively, proving the importance of deep network structure in feature extraction. Traditional feature extraction methods, such as GIST, have insufficient representation ability and are prone to losing key discriminative information when dealing with fine-grained pattern images with small inter-class differences and large intra-class differences. Furthermore, these models generally lack an adaptive feature optimisation mechanism, resulting in a significant decline in generalisation performance under complex backgrounds, noise interference and other conditions. EfficientNetV2 balances network depth, width and resolution through a composite scaling strategy, enabling it to efficiently handle multi-scale pattern features. The introduction of SA further enhanced the model's ability to focus on key regions. With almost no increase in computational overhead, the classification accuracy was improved by 3.04%. The accuracy rate of the EfficientNetV2-SA model in the classification task of Xiang embroidery patterns reached 97.3%, while the recall rate and specificity reached 94.2% and 99.7% respectively. It indicates that the SA mechanism can effectively capture the discriminative local features (such as the direction of stitches and colour transitions) in intangible cultural heritage patterns, thereby enhancing the discriminative ability of the model.

Furthermore, under the circumstances of 10 different types of interference, the classification accuracy of the EfficientNetV2-SA model can still remain above 92%, demonstrating its stability and adaptability in complex environments.

However, textile patterns from different regions and technological backgrounds vary significantly in terms of form, colour and texture features. While maintaining high precision, the model still needs to further enhance its adaptability to cross-category and complex patterns. Future research can be optimised in the following directions: (1) incorporating multi-scale feature fusion mechanisms to enhance the recognition performance for complex patterns; (2) integrating transfer learning and domain adaptation methods to enable the model to quickly adapt to new pattern categories under limited labeled data, thereby improving its generalisation ability across regions and crafts; (3) introducing graph neural networks or Transformer architectures into the model to capture the underlying geometric relationships and long-range dependencies of patterns, thereby strengthening the modeling of spatial distribution and symmetrical structures; and (4) applying explainable artificial intelligence methods to visualize and analyze the model's decision basis for different pattern categories, which not only enhances the model's credibility but also provides technical support for the cultural interpretation of intangible heritage.

Author Contributions

Xue Bai Participated in the experimental design, written the draft and revised the manuscript. Completed experiments, written the draft manuscript, analyzed the experimental data.

Conflicts of Interest

The author declares no conflict of interest.

Funding

The research is supported by the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No.KJQN202501621); Supported by the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No.KJQN202401625); Supported by Chongqing Municipal Education Commission's Youth Program for Humanities and Social Sciences Research (Grant No.25SKGH273); The Phased Achievements of First-Class Undergraduate Course Construction at

Chongqing University of Education (Grant No.xylkc2317hh); Stage Achievements of First-class Undergraduate Course of Chongging Universities (Grant No.sylkc2311hh); 2024 Higher Education Teaching Reform Research Project of Chongqing University of Education (Grant No.JG202401).

REFERENCES

[1] Radmila BJ. A comparison of methods for image classification of cultural heritage using transfer learning for feature extraction. Neural Computing and Applications. 2023; 36(20):11699-11709. doi:10.1007/s00521-023-08764-x

[2] Xiang C, Yang Y, Zhou T, Wang T. Digital reconstruction of historical cultural landscapes based on image recognition technology. Traitement du Signal. 2024; 41(3):55-63. doi:10.18280/ts.410336

[3] Ju F. Mapping the knowledge structure of image recognition in cultural heritage: A scientometric analysis using citespace, vosviewer, and bibliometrix. Journal of Imaging. 2024; 10(11):272-276. doi:10.3390/jimaging10110272

[4] Sabeenian RS, Paul E, Prakash C. Fabric defect detection and classification using modified VGG network. The Journal of the Textile Institute. 2023; 114(7):1032-1040. doi:10.1080/00405000.2022.2105112

[5] Eom TH, Lee HS. A study on the diagnosis technology for conservation status of painting cultural heritage using digital image analysis program. Heritage. 2023; 6(2):1839-1855. doi:10.3390/heritage6020098

[6] Yalemisew A, Renato SR, Japesh M, Gerda K, Amelie D. A methodology for semantic enrichment of cultural heritage images using artificial intelligence technologies. Journal of Imaging. 2021; 7(8):121-128. doi:10.3390/jimaging7080121

[7] Ou Y, Sun C, Yuan R. High-frequency workpiece image recognition model integrating multi-level network structure. Sensors. 2024; 24(6):12-16. doi:10.3390/s24061982

[8] Amit C, Prabhat V. Improving freedom of visually impaired individuals with innovative efficientnet and unified spatial-channel attention: A deep learning-based road surface detection system. Tehnički Glasnik. 2025; 19(1):17-25. doi:10.31803/tg-20231018184747

[9] Alruwaili M, Mohamed M. An integrated deep learning model with efficientnet and resnet for accurate multi-class skin disease classification. Diagnostics. 2025; 15(5):551-555. doi:10.3390/diagnostics15050551

[10] Ishaq A, Ullah MUF, Hamandawana P. Improved efficientnet architecture for multi-grade brain tumor detection. Electronics. 2025; 14(4):710-716. doi:10.3390/electronics14040710

[11] Zeng Z, Sun S, Sun J. Constructing a mobile visual search framework for Dunhuang murals based on fine-tuned CNN and ontology semantic distance. The Electronic Library. 2022; 40(3):121-139. doi:10.1108/EL-09-2021-0173

[12] Roland R, Angelica C, Diputra AJ. CNN classifier parameter optimisation with genetic algorithms: A case study of indonesian batik patterns. International Journal of Computational Intelligence and Applications. 2024; 23(02):851-862. doi:10.1142/S1469026824500044

[13] Shigeki K, Kenichiro K. Individual model identification of waste digital devices by the combination of CNN-based image recognition and measured values of mass and 3D shape features. Journal of Material Cycles and Waste Management. 2024; 26(4):2214-2225. doi:10.1007/s10163-024-01961-3

[14] Maalek R, Maalek S. Automatic recognition and digital documentation of cultural heritage hemispherical domes using images. Journal on Computing and Cultural Heritage (JOCCH). 2023; 16(1):21-28. doi:10.1145/3528412

[15] Liu Y, Cheng P, Li J. Application interface design of Chongqing intangible cultural heritage based on deep learning. Heliyon. 2023; 9(11):e22242-e22247. doi:10.1016/j.heliyon.2023.e22242

[16] Dubinsky Y. Country image, cultural diplomacy, and sports during the COVID19 pandemic: Brand America and Super Bowl LV. Place Branding and Public Diplomacy. 2022; 19(3):249-265. doi:10.1057/s41254-021-00257-9

[17] Xi S, Robin C, Shiry G. Spatially-consistent feature matching and learning for heritage image analysis. International Journal of Computer Vision. 2022; 130(5):1325-1339. doi:10.1007/s11263-022-01576-x

[18] Mokayed H, Quan TZ, Alkhaled L, Sivakumar V. Real-time human detection and counting system using deep learning computer vision techniques. Artificial Intelligence and Applications. 2023; 1(4):221-229. doi:10.47852/bonviewAIA2202391

[19] Bond-Taylor S, Leach A, Long Y, Willcocks CG. Deep generative modelling: A comparative review of VAEs, GANs, normalising flows, energy-based and autoregressive models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2022; 44(11):7327-7347. doi:10.1109/TPAMI.2021.3116668

[20] Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2022; 44(7):3523-3542. doi:10.1109/TPAMI.2021.3059968

[21] Cheng H, Zhang M, Shi JQ. A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2024; 46(12):10558-10578. doi:10.1109/TPAMI.2024.3447085