A Hybrid CNN-Transformer Framework for Fine-Grained Recognition of Manchu Embroidery Motifs

Zhaoli Wang*, Minghao Liu, Chunlei Cao, Yuting Zhang

Harbin Institute of Information Technology, Harbin, China

*hxciwangzhaoli@126.com

Abstract. Manchu embroidery motifs exhibit strong visual similarity across categories while differing in subtle stitch-level details, making fine-grained recognition a challenging task. To address this issue, this paper proposes a hybrid CNN-Transformer framework that integrates local feature extraction and global context modeling. The CNN backbone captures fine-grained texture information such as stitch direction and density, while the Transformer encoder models long-range dependencies and overall motif structures. A feature fusion module is further designed to combine local and global representations for improved discrimination. Experimental results demonstrate that the proposed method outperforms conventional CNN and Transformer baselines, achieving superior performance in recognizing visually similar embroidery patterns.

Keywords: Manchu embroidery; fine-grained classification; CNN-Transformer; feature fusion; computer vision.

1 Introduction

Manchu embroidery, an important form of traditional Chinese textile art, is characterized by bold compositions, stylized motifs, and strong color contrasts. However, its patterns often exhibit high visual similarity, making them difficult to distinguish using conventional image classification methods. To address this challenge, this paper focuses on fine-grained recognition of Manchu embroidery motifs, where discriminative information lies in subtle local details such as stitch direction and texture patterns. We propose a hybrid CNN-Transformer framework that combines convolutional neural networks for local feature extraction with Transformers for global context modeling, enabling more effective representation of both stitch-level details and overall structural patterns.

Recent studies have explored embroidery image analysis and fine-grained classification using both traditional methods and deep learning approaches. Convolutional neural networks (CNNs) have demonstrated strong capability in extracting local texture features, while Transformer-based models have shown advantages in modeling global contextual relationships. In addition, hybrid CNN-Transformer architectures have achieved promising results in various vision tasks by integrating local and global representations. However, their application to embroidery motif recognition remains

R. Zur et al. (eds.), *Proceedings of the 2026 5th International Conference on Art Design and Digital Technology (ADDT 2026)*, Advances in Computer Science Research 131,
https://doi.org/10.2991/978-94-6239-737-8_26

limited, especially for capturing both stitch-level details and overall structural patterns. This motivates the proposed approach.

2 Literature Review

2.1 Embroidery Image Analysis and Fine-Grained Classification

Existing studies on embroidery image analysis mainly focus on texture classification and traditional pattern recognition. Early methods rely on handcrafted features such as color histograms, SIFT, and local binary patterns to describe visual characteristics. While these approaches achieve reasonable performance on simple textile patterns, they are often insufficient for complex embroidery motifs with high intra-class variation and subtle inter-class differences. In recent years, deep learning-based methods, especially convolutional neural networks (CNNs), have been widely applied to cultural heritage image analysis, significantly improving classification accuracy [1], [2].

Fine-grained image recognition has also attracted increasing attention, aiming to distinguish visually similar subcategories by capturing subtle discriminative features. Methods based on attention mechanisms and part-based models have shown promising results in domains such as bird species and vehicle classification [3]. However, existing approaches rarely consider the unique characteristics of embroidery data, where critical information lies in stitch-level textures rather than semantic object parts. Therefore, these methods fail to jointly capture fine-grained local details and global structural patterns, revealing a gap in embroidery classification tasks and motivating the design of our approach.

2.2 CNN-Transformer Hybrid Models

CNNs are well known for their strong ability to extract local spatial features, making them suitable for capturing fine-grained texture information. However, their limited receptive field restricts their capability to model long-range dependencies and global contextual relationships. Transformer-based models, such as Vision Transformer (ViT) and its variants, address this limitation by leveraging self-attention mechanisms to capture global feature interactions [4], [5]. Nevertheless, Transformers often require large-scale datasets and may overlook fine local details when applied to small or specialized datasets.

To overcome these limitations, recent research has explored hybrid CNN-Transformer architectures that combine the strengths of both paradigms. These models have demonstrated superior performance in various vision tasks by integrating local and global feature representations [6]. However, their application to embroidery motif recognition remains limited, and they do not explicitly consider the characteristics of stitch-level textures. In this paper, we propose a hybrid CNN-Transformer framework tailored for fine-grained classification of Manchu embroidery motifs. By jointly modeling stitch-level textures and overall structural patterns, our approach effectively addresses the above limitations and provides a more suitable solution for this specialized task.

3 Methodology

3.1 CNN-Based Local Feature Extraction

As illustrated in Fig. 1, we propose a hybrid CNN-Transformer framework for fine-grained recognition of Manchu embroidery motifs. The model aims to jointly capture local stitch-level textures and global structural patterns for distinguishing highly similar motifs. The overall pipeline consists of four stages: CNN-based feature extraction, feature tokenization, Transformer-based global modeling, and feature fusion.

In the first stage, fine-grained local features are extracted from embroidery images using ResNet-50 as the convolutional backbone, which consists of four residual stages with bottleneck blocks. The output feature map is taken from the last convolutional stage (C5) with a stride of 32 to capture detailed texture representations. Given the dense and directional stitch patterns in embroidery images, where discriminative information lies in subtle variations of orientation and texture, preserving spatial information is essential.

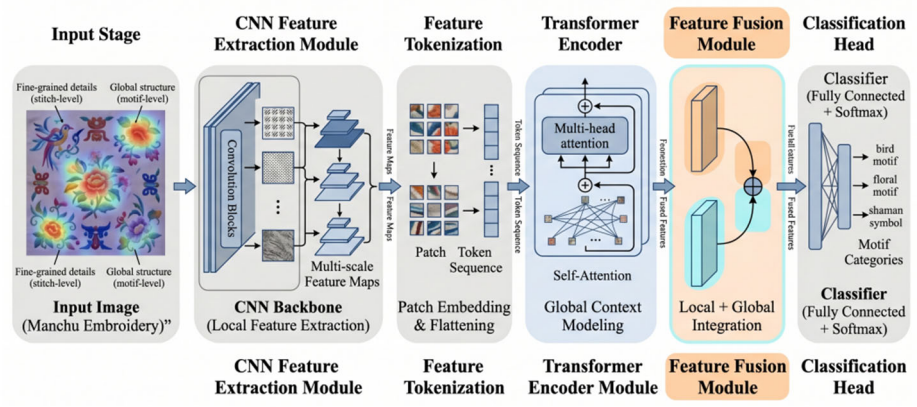

Fig. 1. System Architecture of the CNN-Transformer Framework.

Given an input image, the CNN backbone generates high-dimensional feature maps that encode local patterns such as stitch direction, density, and color transitions. To further enhance feature representation, we introduce a multi-scale feature extraction strategy by aggregating intermediate features from different stages. This allows the network to simultaneously capture micro-level stitch details and mid-level structural cues. The resulting feature maps serve as the foundation for subsequent global modeling.

3.2 Feature Tokenization and Embedding

To enable global context modeling, the extracted CNN feature maps are transformed into a sequence of tokens. As illustrated in Fig. 1, the 2D feature maps are partitioned into non-overlapping patches of size P×P, and each patch is flattened into a vector

representation. This process converts spatial features into a sequence format compatible with the Transformer architecture.

Formally, a feature map of size H×W×C is reshaped into a sequence of tokens, where each token corresponds to a local region of the image. Each patch is flattened into a vector of dimension $P^2 \cdot C$, and a learnable linear projection is applied to map these tokens into a unified embedding space with dimension D=256. In addition, positional encoding is incorporated to preserve spatial relationships among patches, ensuring that structural information of embroidery motifs is not lost during transformation.

This tokenization step bridges the gap between CNN-based local feature extraction and Transformer-based global modeling, enabling the model to operate on both representations effectively.

3.3 Transformer-Based Global Context Modeling

While CNNs excel at capturing local textures, they are limited in modeling long-range dependencies. To address this limitation, we employ a Transformer encoder to learn global contextual relationships across different regions of the embroidery image. As shown in Fig. 1, the token sequence is processed through multiple Transformer layers. Specifically, the Transformer encoder consists of L = 4 layers, each with 8 attention heads, and the embedding dimension is set to 256.

Each layer consists of a multi-head self-attention mechanism and a feed-forward network with a hidden dimension of 512. The self-attention mechanism allows the model to dynamically evaluate the importance of each token with respect to others, thereby capturing global dependencies. The attention operation is defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{1}$$

Through this process, the Transformer effectively models global structures such as symmetry, layout, and motif composition, which are crucial for distinguishing visually similar embroidery patterns.

3.4 Feature Fusion and Classification

To fully leverage both local and global features, we design a feature fusion module, as shown in Fig. 1. This module integrates CNN-derived local features and Transformer-derived global features into a unified representation. Specifically, the local feature vector $f_t \in \mathbb{R}^{d_c}$ and global feature vector $f_t \in \mathbb{R}^{d_t}$ are concatenated and passed through a fully connected layer:

$$f = \sigma(W[f_c; f_t] + b) \tag{2}$$

where $[\cdot;\cdot]$ denotes concatenation, $W \in \mathbb{R}^{d\times(d_c+d_t)}$ and b are learnable parameters, and σdenotes the ReLU activation function.

The fused feature is then fed into a fully connected classification head with a Softmax layer to predict motif categories. The overall objective function of the proposed model is defined as the cross-entropy loss:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{3}$$

where y_i and $\hat{y}_i$ denote the ground truth label and predicted probability, respectively. By integrating both fine-grained texture cues and global structural information, the proposed framework achieves improved performance in distinguishing highly similar Manchu embroidery motifs.

4 Experiments and Evaluation

4.1 Experimental Setup

To evaluate the proposed method, a Manchu embroidery motif dataset was constructed from museum archives, online repositories, and curated sources. The dataset contains approximately 1,800 images across 6 categories, with 250–350 samples per class, exhibiting significant intra-class variation and inter-class similarity. All images are resized to 224×224 and augmented using random cropping, horizontal flipping, and color jittering. The dataset is split into training and testing sets with a ratio of 8:2.

The model is implemented on a GPU platform and compared with representative CNN-based and Transformer-based baselines. Performance is evaluated using accuracy, precision, recall, and F1-score, providing a comprehensive assessment of fine-grained recognition performance.

4.2 Main Experimental Results

The main experimental results are summarized in Table 1, where the proposed method is compared with representative baseline models, including a CNN-based model (ResNet-50) and a Transformer-based model (ViT). In addition, an ablation comparison is conducted to evaluate the contribution of different components in the proposed framework.

Table 1. Comparison and ablation results of different models

Model	Description	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet-50	CNN only	84.3	83.7	82.9	83.3
ViT	Transformer only	86.1	85.4	84.8	85.1
CNN + Transformer	w/o fusion	88.0	87.5	87.0	87.2
Proposed Method	full model	90.5	89.8	89.2	89.5

The CNN-only model shows lower performance due to limited global context modeling, while the Transformer-only model captures global information but lacks fine-grained local details. Their combination improves performance, demonstrating the complementarity of local and global features. Furthermore, the proposed method

with the fusion module achieves the best results across all metrics, confirming its effectiveness in enhancing discriminative capability for fine-grained embroidery motif recognition.

4.3 Discussion

The experimental results demonstrate that the proposed hybrid CNN-Transformer framework improves fine-grained recognition performance for Manchu embroidery motifs by effectively combining local texture extraction and global structural modeling. Specifically, CNN captures discriminative stitch-level details, while the Transformer models global relationships, enabling accurate distinction of visually similar motifs. However, the model is still affected by the limited scale and diversity of the dataset, and the fusion module introduces additional computational cost. Future work will focus on expanding the dataset, designing lightweight models, and improving interpretability through attention visualization.

5 Conclusion

This paper proposes a hybrid CNN-Transformer framework for fine-grained recognition of Manchu embroidery motifs, integrating CNN-based local feature extraction with Transformer-based global modeling. The proposed method effectively captures both stitch-level details and overall structural patterns, achieving superior performance over conventional CNN and Transformer models. The results demonstrate the effectiveness of combining local and global representations for cultural pattern recognition tasks. Future work will focus on improving dataset scale, model efficiency, and exploring advanced attention mechanisms.

References

1. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR), 2021. https://openreview.net/forum?id=YicbFdNTTy
2. Khan S, Naseer M, Hayat M, Zamir S W, Khan F S, Shah M. Transformers in vision: A survey. ACM Computing Surveys, 2022, 54(10s): 1-41. https://doi.org/10.1145/3505244

3. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 87-110. https://doi.org/10.1109/TPAMI.2022.3152247
4. Khan A, Rauf Z, Sohail A, Rehman A, Asif H, Asif A, et al. A survey of the vision transformers and their CNN-transformer based variants. Artificial Intelligence Review, 2023, 56(S3): 2917-2970. https://doi.org/10.1007/s10462-023-10595-0
5. Dai Y, He K, Hu Q, Huang K. Review of CNN-Transformer hybrid model in computer vision. Modeling and Simulation, 2023, 12(4): 3657-3672. https://doi.org/10.12677/mos.2023.124336
6. Zhang Y, Liu H, Wang X. FcaFormer: Forward cross attention in hybrid vision transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 12345-12354.