

<https://doi.org/10.1038/s40494-026-02404-y>

Research on batik image pattern detection based on improved YOLOv11

Yiting Li^{1,2}, Huafeng Quan¹ ✉, Qin Li¹ & Junjie Wang¹

Batik, as an important intangible cultural heritage, embodies profound cultural significance through intricate pattern systems. However, detecting these patterns in complex batik images poses significant challenges due to dense pattern distributions, scale variations, complex backgrounds, and degraded image quality. This paper proposes a robust batik pattern detection model based on improved YOLOv11 architecture that balances detection accuracy with computational efficiency. First, we construct a comprehensive Chinese batik dataset, addressing the critical data scarcity in this domain. Second, we develop an enhanced YOLOv11 model integrating Vision Outlooker (VOLO) attention mechanisms for capturing long-distance spatial dependencies and Fused-MBConv modules for efficient feature extraction. Third, we implement a prototype system that bridges visual detection with cultural knowledge interpretation through batik knowledge graphs. The proposed approach provides a practical and scalable solution for the digital preservation and interpretation of intangible cultural heritage.

Batik, as a time-honored manual resist dyeing technique, is widely distributed across various cultural regions of the world, including China, Indonesia, Malaysia, Singapore, India, Japan, and other countries. Chinese batik has a history of over two thousand years and maintains living traditions among ethnic minority groups in southeastern Guizhou^{1–3}. In 2006, batik was included in the first batch of national-level intangible cultural heritage lists. As the Miao people do not have their own written language, batik images have become a “living fossil” for studying ethnic history, religious beliefs, aesthetic concepts, and social life. However, with accelerated industrialization and transformations in modern lifestyles, batik faces a severe crisis of inheritance.

To protect this precious cultural heritage, scholars have conducted systematic digital research on batik from multiple perspectives. In cultural connotation mining, researchers have analyzed the ethnic cultural information embedded in batik patterns through field investigations and interviews, revealing how these patterns reflect the history, culture, and aesthetic concepts of the Miao ethnic group^{2,4,5}. Knowledge representation methods, including natural language processing and multimodal knowledge graphs, have been developed to systematically organize scattered batik knowledge and intuitively display visual pattern characteristics^{6,7}. For pattern innovation design, various computational methods have been proposed, including fractal geometry-based automatic generation⁸, shape grammar combined with artificial neural networks⁹, interactive genetic algorithms with neural network cognitive models¹⁰, and semiotics¹¹. These approaches have also extended to pattern re-use optimization using particle swarm algorithms

and personalized product recommendation systems^{12,13}. In image feature extraction, researchers have developed techniques such as morphological operations with edge detection for contour extraction¹⁴, Zernike Moments combined with curvelet transform for global and local feature extraction¹⁵, and distance transformation algorithms combined with EFEM for crack feature extraction and simulation¹⁶. These technological innovations have significantly expanded the expression forms of traditional patterns and opened new avenues for the living inheritance of intangible cultural heritage.

Current research on batik patterns is mostly limited to the classification and recognition of individual patterns. For example, Elvitaria et al.¹⁷ proposed a batik pattern classification system based on the integration of deep learning and gray level co-occurrence matrix (GLCM) feature extraction. Dewa et al.¹⁸ used a convolutional neural network to classify Indonesian batik patterns into five categories. To address the class imbalance problem in pattern classification, Beatrice et al.¹⁹ constructed a pattern classification model by introducing oversampling and enhancement operations and using DenseNet169. In our previous research⁵, we also used an improved ResNet34 model to construct a Chinese batik pattern classification model. These studies all segment patterns from the original image for independent feature extraction and classification. Although this method has demonstrated effectiveness in terms of classification accuracy, it has the following limitations: First, this method overlooks the correlation between patterns and the overall composition, ignoring the spatial relationships between patterns; second, the boundary blurring and pattern overlap during image segmentation lead to substantial loss of detailed information. To address this

¹College of Big Data and Statistics, Guizhou University of Finance and Economics, Guiyang, China. ²Key Laboratory of Advanced Manufacturing Technology, Ministry of Education, Guizhou University, Guiyang, China. ✉e-mail: quan_huafeng@163.com

issue, Wu et al.⁷ attempted to use RT-DETR for pattern detection. However, this model is only suitable for ideal situations where the image is complete and the patterns are tiled, and struggles to handle complex situations such as pattern occlusion, deformation, and blurring commonly seen in actual batik images. Therefore, existing research lacks systematic recognition and localization of multiple coexisting patterns in batik under complex conditions, which affects the comprehensive interpretation of batik cultural semantics and the effectiveness of digital preservation.

Object detection technology, capable of simultaneously performing pattern localization and recognition, has demonstrated substantial value in cultural heritage protection. On one hand, it enables automated annotation of extensive batik image collections, significantly enhancing digitization efficiency; on the other hand, by recording spatial distributions of patterns, it supports subsequent research in cultural semantic analysis, stylistic evolution, and association mining.

Among object detection methodologies, the You Only Look Once (YOLO) series has emerged as the dominant paradigm, due to their optimal speed-accuracy trade-off. In heritage applications, researchers have successfully employed YOLO variants for cultural element detection. Gao et al.²⁰ proposed a YOLOv8-based recognition model for Jiangnan private gardens, optimizing model representation and feature extraction to facilitate garden heritage preservation. They constructed a comprehensive dataset capturing various viewing angles and lighting conditions in complex garden environments, achieving 8.7% accuracy improvement over the baseline through data augmentation. In Portugal, where Azulejos tiles cover heritage buildings with intricate patterns, natural weathering and material degradation pose preservation challenges. Karimi et al.²¹ constructed a tile defect dataset, employing YOLO for damage detection and MobileNet for severity classification, achieving over 72% overall accuracy. Darma et al.²² introduced a YOLOv5-based method for detecting Balinese carving patterns on sacred architecture, utilizing dataset expansion and network scaling strategies to enable robust detection in complex environments.

YOLOv11, with its enhanced feature pyramid networks and refined detection heads, provides a solid foundation for pattern recognition tasks. However, its convolutional architecture inherently operates on local receptive fields, limiting its ability to model the long-distance spatial dependencies that define cultural semantics in batik compositions. This architectural constraint presents unique challenges when applying baseline YOLOv11 to batik pattern detection:

First, batik images contain exceptionally dense pattern distributions with numerous patterns co-occurring in a single image, where culturally significant symbols often appear at extremely small scales. This combination of high density and small object size exceeds the detection capabilities of baseline YOLOv11. The model's local receptive fields fail to capture the long-distance spatial relationships that define cultural meanings in pattern combinations, resulting in insufficient detection accuracy for these complex scenarios.

Second, authentic batik samples exhibit substantial quality variations stemming from both historical degradation and collection limitations. Traditional dyeing processes create irregular boundaries through wax-bleeding effects, while age-related fading and surface contamination significantly reduce visual contrast. Additionally, field-collected images suffer from inconsistent lighting conditions and complex textile backgrounds. These factors demand detection models with enhanced robustness and adaptability beyond what baseline YOLOv11 provides. These challenges necessitate architectural innovations that specifically address batik's unique characteristics while maintaining computational efficiency.

To address these challenges, we present a comprehensive solution that combines dataset construction, architectural innovation, and practical application. Our main contributions are summarized as follows:

(1) We have constructed a large-scale batik pattern dataset containing 861 images. This dataset covers 7 major categories of batik patterns, with 9933 bounding boxes. It is currently the largest and most comprehensive Chinese batik detection dataset. It addresses the critical data scarcity bottleneck in batik pattern analysis.

(2) We propose an improved YOLOv11 model with two targeted enhancements: Vision Outlooker (VLO) integration for capturing long-distance feature dependencies with reduced computational complexity, and Fused-MobileNet Bottleneck Convolution (Fused-MBCConv) optimization in C3K2 modules that synergistically combines depth-wise convolutions, inverted residuals, and SE Layers. These modifications reduce network complexity while enhancing feature quality, as validated by ablation studies showing substantial improvements in detection accuracy. The model demonstrates robust performance across diverse batik patterns, effectively addressing the robustness and generalizability challenges inherent in heritage preservation applications.

(3) We have developed an integrated prototype system that bridges pattern detection with cultural knowledge interpretation. By linking our detection model with a batik knowledge graph⁵, the system provides end-to-end functionality from visual recognition to semantic analysis, enabling users to explore pattern meanings, historical contexts, and cultural significance.

The proposed model achieves high accuracy while maintaining minimal hardware dependencies, making it ideally suited for deployment across diverse cultural institutions and educational scenarios.

The remainder of this paper is organized as follows: Section 2 details the improved YOLOv11 model including attention mechanism and convolution optimizations; Section 3 presents experimental results covering dataset construction, ablation studies, comparative experiments, and prototype system implementation; Section 4 concludes the paper.

Methods

Benchmark network

Since batik pattern detection results often serve as inputs for subsequent analysis models, benchmark network selection must balance detection accuracy with computational efficiency. We adopt YOLOv11 as our baseline model, released by Ultralytics in 2024²³. This model has demonstrated superior performance in various computer vision tasks, achieving optimal trade-offs between accuracy and inference speed. Compared to its predecessor YOLOv8, YOLOv11 has mainly improved in three aspects: convolutional unit structure, multi-head attention mechanism, and lightweight detection head. The basic structure is shown in Fig. 1.

First, YOLOv11 introduces the C3K2 architecture in the convolutional unit structure. This architecture is built on the C3K module in the Cross Stage Partial (CSP) structure, and its advantages mainly focus on two aspects: On the one hand, the module divides the features into two parts, one part extracts deeper information from the input features through multiple C3K or Bottleneck structures, and the other part directly passes the features through standard convolution structures, finally concatenating the two parts, balancing parameter reduction and feature extraction effectiveness; On the other hand, C3K2 introduces the idea of multi-scale convolution kernels, allowing designers to freely adjust the size of the convolution kernels to expand the receptive field, enabling the model to handle detection targets of various sizes and improve feature extraction quality in complex scenes or backgrounds. The overall structure of C3K2 is shown in Fig. 2, and its detailed structure is shown in Fig. 2.

Second, YOLOv11 proposes the C2PSA architecture by combining CSP with Pyramid Squeeze Attention (PSA) mechanisms. The CSP structure employs the C2 module, which has a structure similar to C3K. Specifically, the input features are divided into two parts after passing through a convolutional layer. One part is directly passed, and the other part is processed by the PSA module. Then, the two parts of the features are concatenated and another convolution is performed to restore the original number of channels. PSA replaces the original Bottleneck module here. PSA uses multiple convolution kernels (such as 3×3 , 5×5 , 7×7) to extract multi-scale features in parallel, and weights the concatenated feature channels through Squeeze and Excitation (SE) modules. This design enhances the network's attention focus on targets of different scales. The structure of C2PSA is shown in Fig. 3.

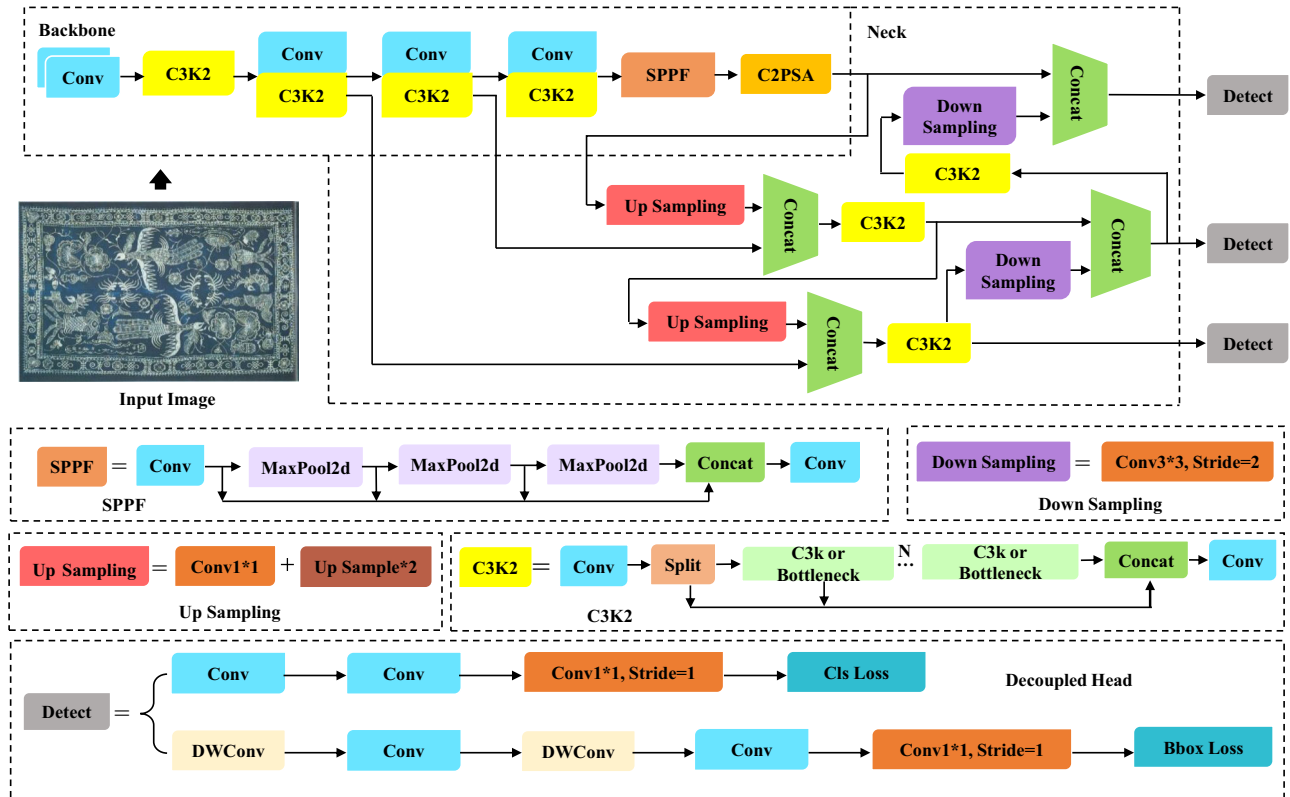
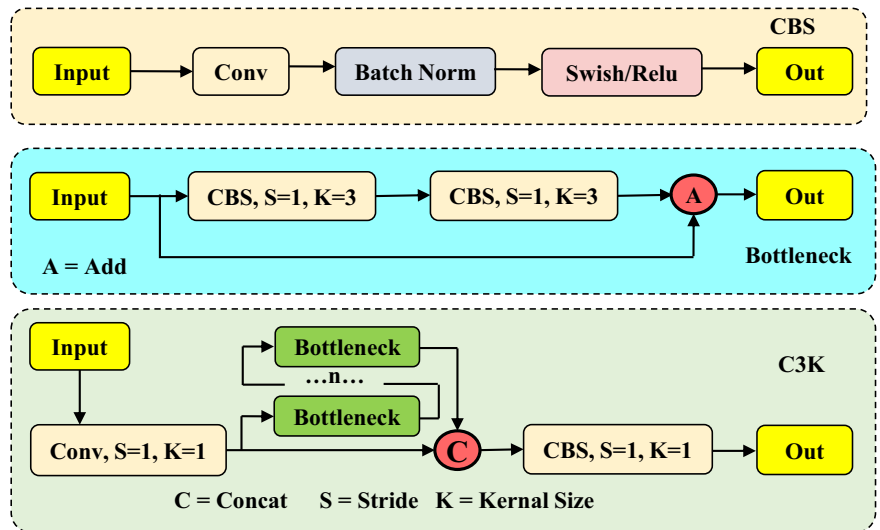


Fig. 1 | Basic structure of YOLOv11. The Conv unit in the figure represents the convolution operation. C3K2, SPPF and other units are module names or operation modes, and some core unit structures are shown in the lower half of the figure.

Fig. 2 | C3K2 module structure diagram. The Conv unit in the figure represents the convolution operation. CBS, Bottleneck and other units are module names or operation methods. Swish/Relu represents the activation function.



Third, YOLOv11 incorporates a depth-wise separable convolution layer in a branch of the decoupled head. Depth-wise separable convolution consists of depth convolution and point convolution. The former processes each input feature channel separately, that is, each channel only focuses on using a single convolution kernel for convolution and does not interact with other channels. This process significantly reduces the number of parameters while improving computational efficiency. Afterwards, to compensate for the lack of cross channel information exchange in deep convolution, a 1 × 1 pointwise convolution layer is used to integrate cross channel feature information and form a complete depth-wise separable convolution

structure. The optimized decoupled head greatly reduces computational complexity and parameter count while maintaining network performance. The schematic diagram of a depth separable structure is shown in Fig. 4.

The loss function of YOLOv11 considers both classification loss and localization loss. The former uses Binary Cross Entropy Loss (BCEL) for classification optimization, while the latter combines Distribution Focal Loss (DFL) and CIoU. The complete loss function can be expressed as:

$$f_{loss} = \lambda_1 f_{cls} + \lambda_2 f_{DFL} + \lambda_3 f_{CIoU} \quad (1)$$

Fig. 3 | C2PSA module structure diagram. The Conv unit in the figure represents the convolution operation. Gap, SE and other units are module names or operation methods. Softmax/Relu represents the activation function.

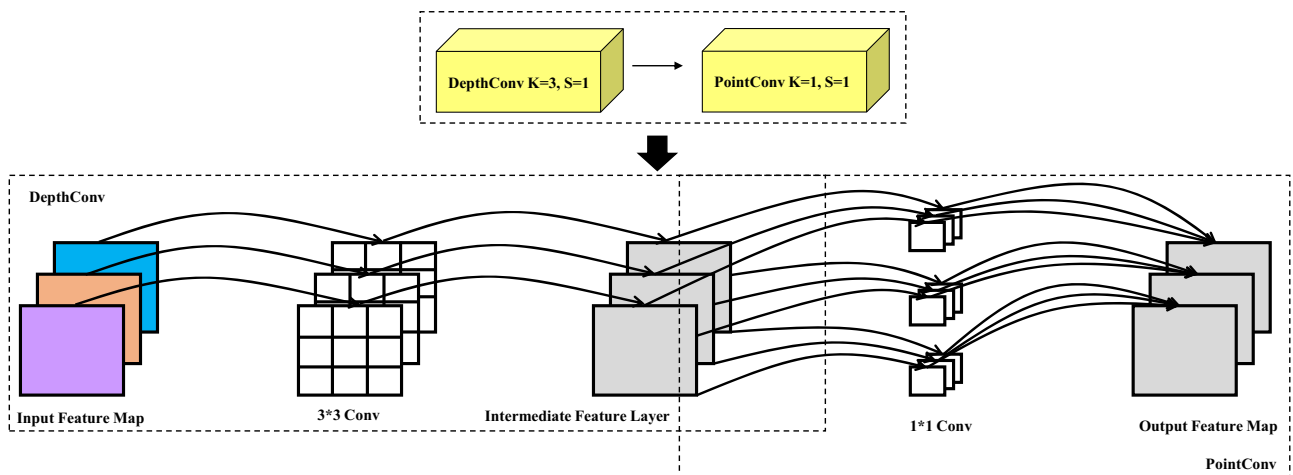
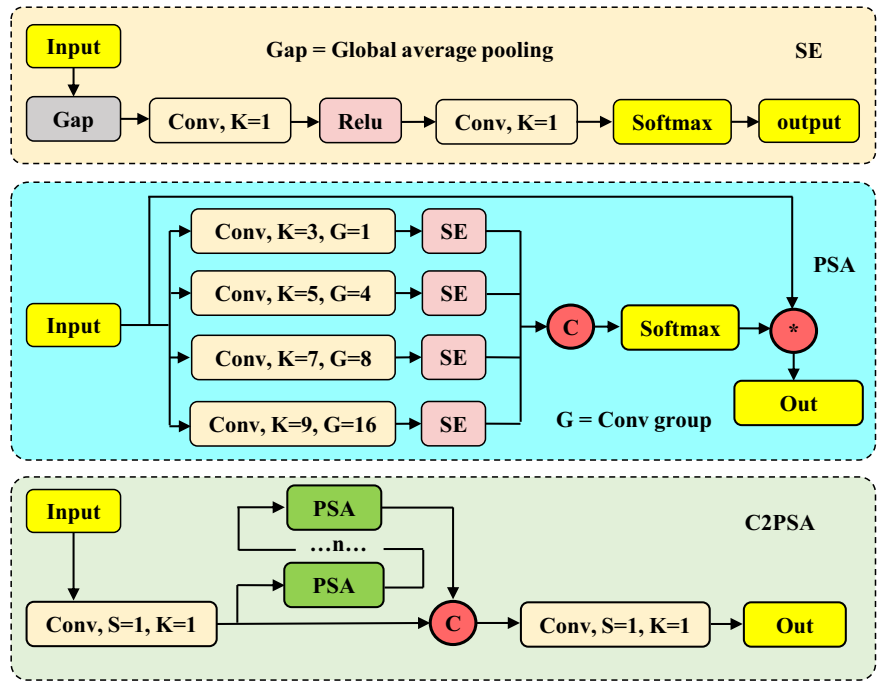


Fig. 4 | Schematic diagram of depth separable structure. The depth separable structure consists of two parts, DepthConv and PointConv, as shown in the figure. The structure and collaboration of the two units are shown in the lower part of the figure.

where, BCEL and CIoU are standard optimization objectives in the YOLO series. DFL generalizes discrete classification results into continuous values through integration:

$$f_{DFL}(S_i, S_{i+1}) = -((y_{i+1} - y) \log(S_i) + (y - y_i) \log(S_{i+1})) \quad (2)$$

where y_i and y_{i+1} are the values that are close to each other on the left and right sides of the continuous label y , satisfying $y_i < y < y_{i+1}$; S_i and S_{i+1} represent the predicted probabilities for the left and right endpoints, respectively.

This paper uses YOLOv11n, which has the least number of parameters, as the benchmark model for constructing a batik pattern detection model through two enhancements. First, addressing limitations of convolutional neural networks (CNNs) in capturing long-distance feature correlations, we introduce the global attention mechanism from VOLO. This mechanism performs attention computation within each image block (window) while combining local block and cross-block interactions to aggregate information from surrounding blocks, maintaining global perception while reducing computational complexity and achieving fine-grained and efficient

feature encoding. Secondly, we employ Fused-MBCConv to optimize the C3k and Bottleneck modules in C3K2. By integrating convolution strategies, inversed residual structures, and SE Layer coordination, this optimization reduces network branches and layers when using depth-wise separable convolutions, enabling the entire network to have stronger feature engineering quality while preserving efficient inference speed. These improvements enable YOLOv11 to achieve lightweight, efficient, and accurate batik pattern detection in complex backgrounds.

Improvement of attention mechanism

VOLO was proposed by Yuan et al.²⁴, which designed a visual foreground operator attention mechanism aimed at efficiently encoding fine features and realizing efficient interaction between image blocks through self-attention mechanism, thereby modeling global dependencies and achieving more optimal visual feature engineering. Considering the bottleneck problem of object detection algorithms in batik printing patterns, namely the large scale variation, high density, and wide range of data sources, this paper refers to the work of reference²⁴ and introduces the core idea of VOLO to the C2PSA

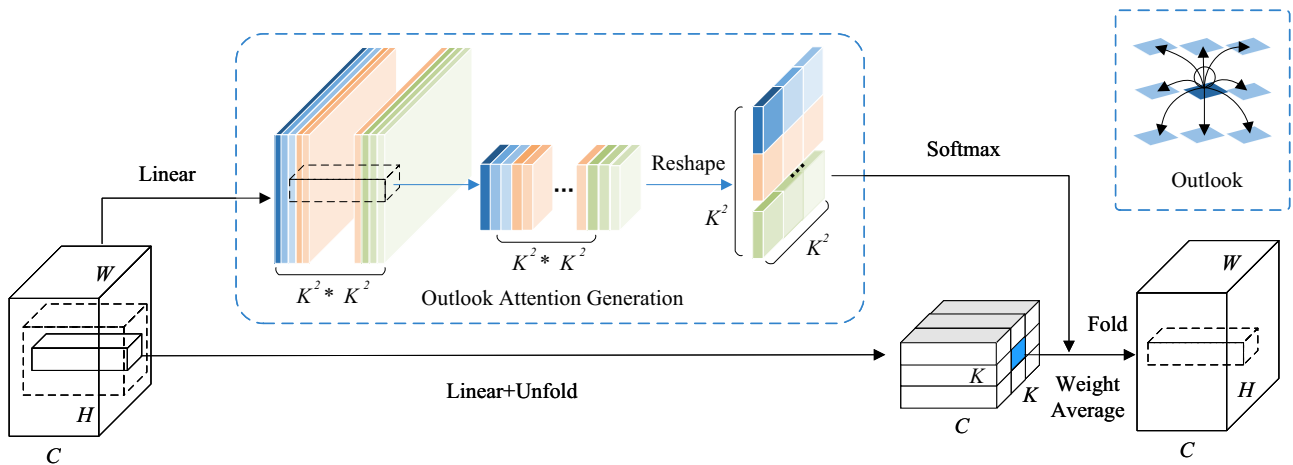


Fig. 5 | Schematic diagram of Outlook structure. The overall structure and operation process of Outlook are shown in detail in the figure, which corresponds to the specific description in the text.

Table 1 | Characteristic parameters of three convolutional structures

Characteristic	Ordinary convolution	MBCnv	Fused-MBCnv
Core operations	Standard convolution	Depth-wise Separable Conv +Inverted Residual	Fusion convolution (standard convolution replaces separation)
High-computational	High	Low	Medium
Applicable scenarios	General computing	Complex feature extraction	Fast feature extraction
Parameter quantity	High	Low	Medium
Typical network	ResNet, VGG	MobileNet, EfficientNet	EfficientNet-V2

module of YOLO v11, replacing the PSA attention mechanism and establishing high-quality relationships between long-distance features of the model, aiming to improve the feature engineering quality of the model and overcome the related problems of complex batik printing pattern detection. Specifically:

First, for input features with width and height of W and H respectively $X \in \mathbb{R}^{H \times W \times C}$, two linear transformation layers with weights of $W_A \in \mathbb{R}^{C \times K^4}$ and $W_V \in \mathbb{R}^{C \times C}$ are used to map each C -dimensional window feature into a similarity matrix $A \in \mathbb{R}^{H \times W \times K^4}$ and value representation $V \in \mathbb{R}^{H \times W \times C}$.

Second, defining $V_{\Delta i,j} \in \mathbb{R}^{C \times K^2}$ as all values within the local window at position (i, j) , it can be expressed as:

$$V_{\Delta i,j} = \{V_{i+p-|k/2|,j+q-|k/2|}\}, 0 \leq p \leq K, 0 \leq q \leq K \quad (3)$$

By rearranging the dimension of the similarity matrix at position (i, j) $A_{i,j} \in \mathbb{R}^{H \times W \times K^4}$ into $A_{i,j} \in \mathbb{R}^{K^2 \times K^2}$, applying the attention mechanism in the form of Eq. (4), and adding the different weighted values of different local windows at the same position, the output is obtained.

$$Y_{\Delta i,j} = \text{MatMul}(\text{Soft max}(\hat{A}_{i,j}), V_{\Delta i,j}) \quad (4)$$

Finally, the results of implementing the Outlook attention mechanism will be fused with the original input features to effectively capture the long-distance feature relationships of the model. This process can be represented by the following equations:

$$\hat{X} = \text{Outlook Attention}(\text{LN}(X)) \quad (5)$$

$$Z = \text{MLP}(\text{LN}(\hat{X})) + \hat{X} \quad (6)$$

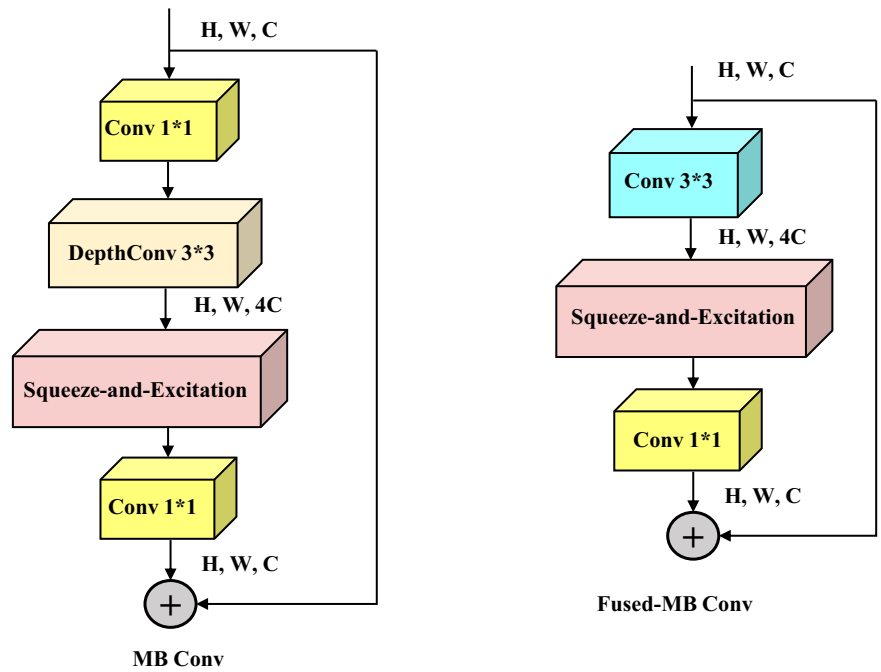
where Outlook Attention is the attention representation in Eq. (4); LN is layer normalization. The visualization of this process are shown in Fig. 5.

Improvement of convolution method

Similar to the core idea embodied in YOLOv11 in the decoupled head, we apply depth-wise separable convolution to the C3K2 module. At the same time, we refer to the Fused-MBCnv structure proposed in EfficientNetV2²⁵ and make the following considerations: first, we use deep convolution (channel wise convolution) and point wise convolution (1x1 convolution) to significantly reduce computational and parameter complexity. For example, for a standard 3 x 3 convolution kernel, assuming its input channel is C_{In} and output is C_{Out} , the number of standard convolution parameters is $3 \times 3 \times C_{In} \times C_{Out}$, while the number of depth-wise separable convolution parameters is only $3 \times 3 \times C_{In} + 1 \times 1 \times C_{In} \times C_{Out}$. Second, the introduction of inversed residual structure and SE Layer further enhances feature engineering. We perform a 1x1 convolution to increase the dimensionality (by expanding the number of channels), then perform a deep convolution, and enhance the quality of feature extraction through an SE layer, then use a 1x1 convolution to reduce the dimensionality of the features. This structure preserves more feature information while reducing computational complexity, known as the MBCnv structure. Finally, if the deep convolution and pointwise convolution in MBCnv are combined into a single standard convolution to reduce branching operations, it can further accelerate the efficiency of feature extraction. In summary, the features of standard convolution, MBCnv, and Fused-MBCnv are represented in Table 1. The structural analysis of MBCnv and Fused-MBCnv is shown in Fig. 6.

Fused-MBCnv offers the following advantages: On the one hand, conventional depth-wise separable convolution separates spatial feature extraction (depth-wise convolution) from channel feature fusion (pointwise convolution), and the decomposed operation may not fully fuse feature information, especially when dealing with high complexity tasks (such as the complex background of batik patterns in this paper), which may lead to insufficient modeling ability for complex spatial channel relationships and may result in decreased accuracy. In addition, this structure usually needs to be compensated for by increasing the network width or depth, thereby partially offsetting its efficiency advantage. Fused-MBCnv combines the

Fig. 6 | Schematic diagram of MB Conv/Fused MB Conv structure. The specific structure of MB Conv is shown on the left side of the figure, and the specific structure of Fused MB Conv is shown on the right side. Conv, DepthConv, etc. are module names or operation methods.



deep convolution and pointwise convolution in MBConv into a single standard convolution, reducing network branches and layers, which to some extent combines the advantages of both. On the other hand, Fused-MBConv improves the quality of feature extraction through the SE attention mechanism. SE is a lightweight and efficient channel attention mechanism that can explicitly model the dependencies between channels, adaptively learn the weights of different channels, enhance the feature responses of important channels, and suppress the contributions of irrelevant or noisy channels. It works by compressing the spatial dimension into channel descriptors through global average pooling, capturing global contextual information, avoiding the limitations of local receptive fields, and enhancing the flexibility of the model's expressive ability by dynamically adjusting the importance of channel features. Therefore, considering the complexity of the batik modes, which requires the detection model to have better robustness and adaptability due to historical degradation and collection limitations of real batik samples, this paper uses Fused MB Conv to replace the C3k and Bottleneck modules in C3K2. This method has to some extent improved the feature engineering quality of the model, effectively solving the inherent challenges of robustness and universality in the application of batik heritage protection.

So far, the improved network structure is shown in Fig. 7. In the figure, the “Δ” mark represents the location where VOLO is added. The “✱” mark represents the location where Fused-MBConv was added. It can be seen that VOLO has replaced the attention mechanism of some PSA Blocks in the original C2PSA module. Fused-MBConv is integrated into the C3K2 module. The two complement each other, enabling the entire network to have stronger feature engineering quality while maintaining efficient inference speed.

Results

Dataset and its preprocessing

The intangible cultural heritage of China covers traditional skills, folk festivals, music and dance, oral literature, etc., and is the “living gene” accumulated over five thousand years of Chinese civilization. Protecting intangible cultural heritage helps prevent cultural homogenization and maintain the diversity and vitality of Chinese culture. Among them, the batik patterns of the Miao ethnic group in Guizhou, with their mysterious style, beautiful patterns, clear intentions, profound cultural connotations, and rich themes, have been passed down from generation to

generation and have become one of the most representative intangible cultural heritages of China. Focusing on the detection of pattern elements and analysis of meanings in complex batik patterns, we have conducted on-site visits, collected public information, and consulted authoritative books (from libraries in multiple parts of Guizhou Province, China). We have compiled 861 complete complex batik patterns for constructing batik pattern detection models. Overall, the constructed dataset includes 7 pattern categories: butterfly, fish, dragon, bird, drum, flower, and fruit. Partial patterns are shown in Fig. 8a. Due to the difficulty of collecting the dataset, the wide range of channels for obtaining it, and the uneven quality of the patterns involved, there are detection challenges including complex backgrounds, large variations in lighting, diverse pattern forms, and concentrated pattern density in the images. Each complete batik image contains multiple patterns of various types. The data contains a large amount of pattern overlap and combination, which increases the complexity of the detection process. The dataset contains a total of 9933 manually annotated bounding boxes, as shown in Fig. 8b. From Fig. 8c, it can be seen that the scale of the anchors in the batik patterns varies greatly, with most of them concentrated in the 50 × 50 to 250 × 250 pixel range. This scale variation requires the model to have strong multi-scale feature extraction capabilities. From the perspective of small classes, the number of flower class anchors is 2486, while the number of drum class anchors is only 422, which emphasizes the high requirement for model robustness to better handle the problem of imbalanced samples between classes.

The partitioning and processing method of the dataset in this paper is as follows: First, 734 complex batik images are randomly partitioned as the training set and 127 images as the testing set. Second, the resolution of the original patterns is mostly concentrated within 640×640 pixels, with a few patterns exceeding this range (up to 1435 × 2292 pixels). To balance the efficiency and effectiveness of the training process, the patterns are normalized to 640×640 pixels to preserve their features. Third, due to the difficulty in collecting complex batik patterns, the sample size is limited. To ensure the feasibility of establishing deep learning models and reduce the risk of overfitting, the dataset is augmented in the experiment. Specifically, it includes Blur, Advanced Blur, ToGray, CLAHE, Flip, Mosaic, Translation, and Erasing. The specific meanings and parameter settings are shown in Table 2, where *p* represents the probability of using each data augmentation operation.

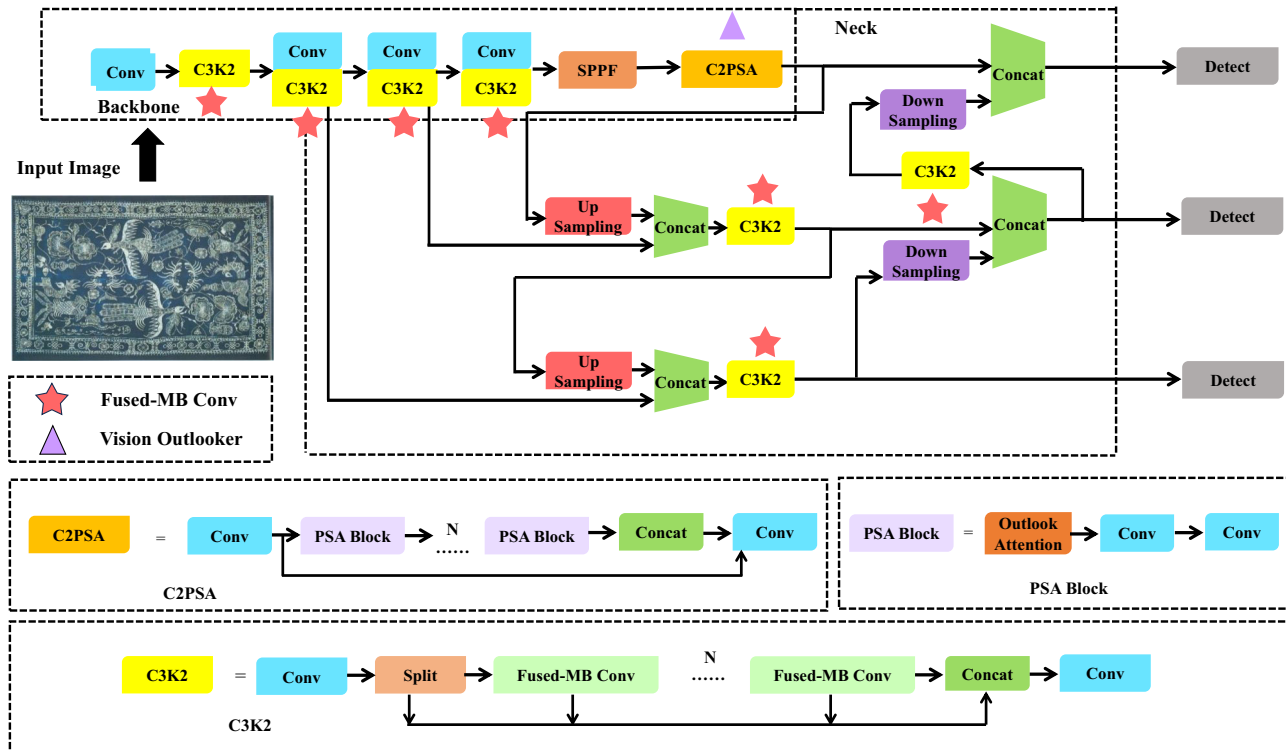


Fig. 7 | Structure diagram of batik detection model. The Conv unit in the figure represents the convolution operation. C3K2, SPPF and other units are module names or operation modes, and some core unit structures are shown in the lower half of the figure. Triangle and star markings represent positions where the structure has been improved.

During the model training process, the specific experimental hardware, software, and parameter settings are shown in Table 3. The deep learning model frameworks used in the experiment are PyTorch1.9.1 and Torchvision 0.10.1. To ensure the fairness of various models in ablation experiments and comparative experiments, the following two experimental settings are made: first, the training epoch of all models is set to 300. 300 epochs ensure that the model achieves its optimal performance, and this parameter has been validated through repeated experiments. Second, the parameter initialization process of all models does not use pre-trained weights and uses the same optimizer (SGD) to control experimental variables. In addition, considering the universality of the model in reality and reducing its dependence on hardware, we optimize and validate it using YOLOv11n, the smallest model in the YOLOv11 series.

Ablation experiment

The advantages of the batik pattern detection model designed in this paper mainly lie in two aspects: At the data processing level, considering the limited amount of complex batik pattern data, we have designed a data augmentation method to prevent overfitting. In terms of model structure, we integrate the core ideas of VOLO, MBCConv, and Fused-MBConv into the baseline model to enhance its ability to extract global features and model complex spatial channel relationships. To fully analyze the effectiveness of the above two aspects for batik pattern detection tasks, we designed an ablation experimental process, as shown in Table 4. The difference between models IV, V, and VI is that model IV completely uses MBCConv to replace the benchmark network’s C3K2 module; model V completely replaces the C3K2 module of the benchmark network with Fused-MBConv; model VI replaces the C3K2 module of the benchmark network with a mixture of MBCConv and Fused-MBConv.

To evaluate model performance from multiple dimensions, widely used indicators in the field of object detection are used here: precision rate (P), recall rate (R), average precision (AP), mean average precision (mAP),

frames per second (FPS). We quantitatively explore the advantages and disadvantages of the model based on the number of parameters and model size. The performance of the six models mentioned above is shown in Table 5 and Fig. 9.

Overall, the introduction of data augmentation, VOLO module, and Fused-MBConv module has gradually improved the performance of the benchmark model detection. Especially in complex batik patterns where data is precious and scarce, increasing the training data reasonably can to some extent avoid overfitting of the model. In addition, improvements in attention mechanisms and convolutional structures have enhanced the feature engineering quality of the baseline network. Under the implementation of the above measures, mAP increased from 0.716 in benchmark model I to 0.748 in model V, with an improvement rate of 4.47%. At the same time, although the parameter count and size of the model have slightly increased, the FPS is comparable to the benchmark model, with a model size of only 6.962MB. This means that the designed network structure effectively improves detection accuracy while ensuring detection efficiency.

From the perspective of individual modules, the benchmark model II, which has undergone data augmentation, obtained the optimal precision rate among all models, but there is a problem of low recall rate, which is prone to missed detection; After adding the VOLO module, the model has been optimized in both feature extraction ability and structure, significantly improving mAP while only slightly increasing the number of parameters. The MBCConv and Fused-MBConv modules balance detection accuracy and detection speed, respectively. For example, Model IV, which completely uses MBCConv to improve the C3K2 module in the benchmark network, has the optimal number of parameters and model size, but the detection results are slightly lower than Model III. Model V, which completely replaces the C3K2 module with the Fused-MBConv module, achieved the optimal recall rate and mAP, but slightly increased the number of model parameters. Model VI, which combines two modules, actually exhibits a loss of accuracy.

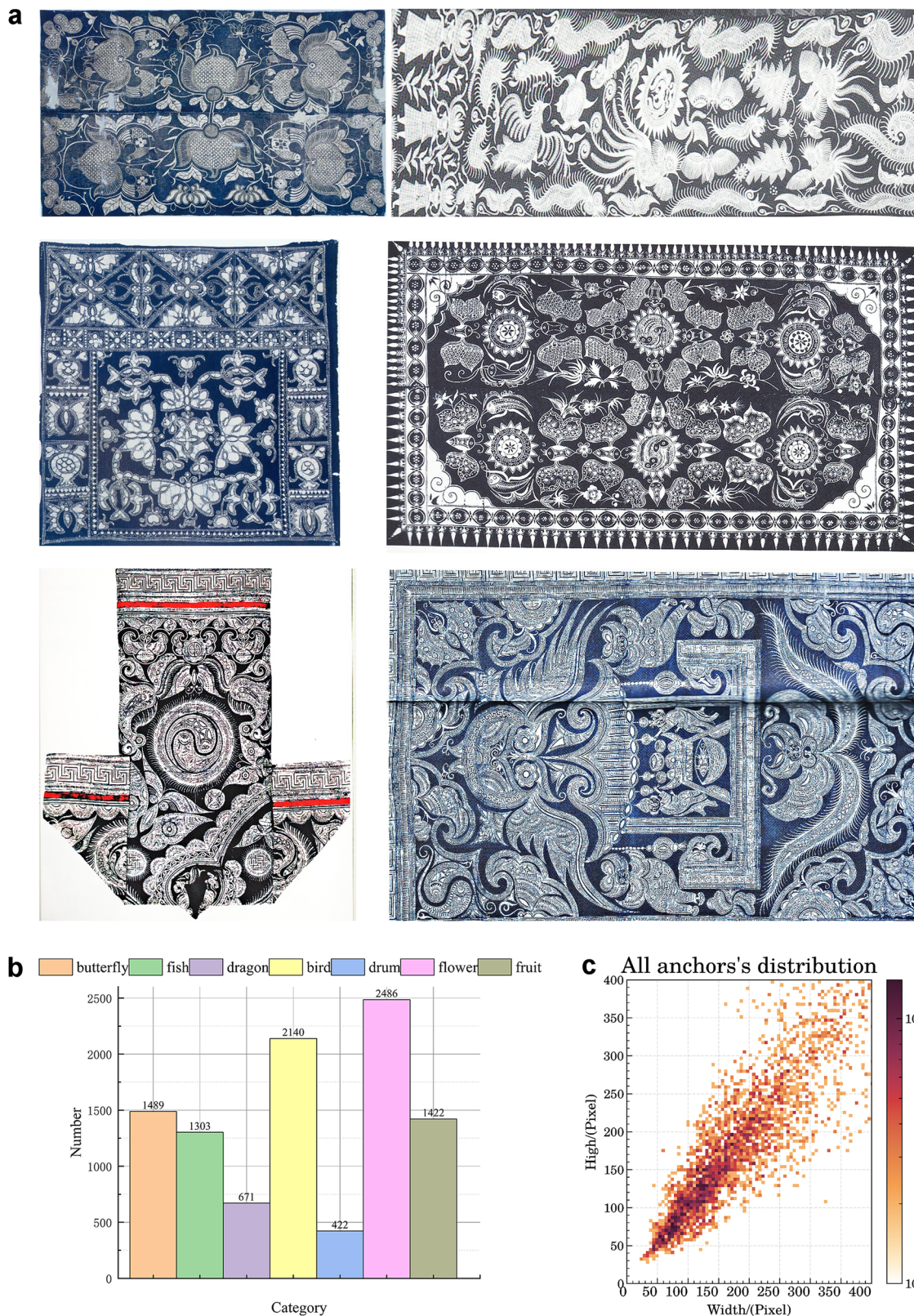


Fig. 8 | Schematic diagram of dataset and labels. The upper part of the figure shows typical batik patterns in the dataset. The lower left part shows the distribution of pattern types in the dataset. The lower right part shows the distribution of targets

sizes in the dataset. **a** Schematic diagram of batik printing pattern. **b** Number distribution of anchors. **c** Size distribution of anchors.

From the perspective of individual categories, models III, IV, and V have all shown impressive performance in multiple categories. For example, Model III has strong detection capabilities for Dragon and Flower categories, Model IV has better detection performance for Butterfly and Drum

categories, but ultimately Model V has the best comprehensive detection accuracy. This indicates that the improvement mechanism for the benchmark model is feasible and effective. An example of partial detection results is shown in Fig. 10.

Table 2 | Parameter setting

Parameters	Setup	Specific operation
Blur	$\rho = 0.1$	Blur the input image using a random-sized kernel
Advanced Blur	$\rho = 0.1$	Blur the input image using a generalized normal filter with a randomly selected parameters
ToGray	$\rho = 0.1$	Convert the image to grayscale
CLAHE	$\rho = 0.1$	Apply Contrast Limited Adaptive Histogram Equalization to the input image
Flip	$\rho = 0.5$	Flip the image up and down
Mosaic	$\rho = 1.0$	Combine four images into one
Translation	$\rho = 0.1$	Image translation
Erasing	$\rho = 0.4$	Randomly erase certain areas of the image
Close mosaic	Last 10 epochs	Time to close Mosaic

Table 3 | Training parameter setting

Environment	Parameter		
CPU	Intel(R) Core(TM) i9-12900K@3.19 GHz	Batch Size	16
GPU	GeForce RTX 3090	Optimizer	SGD
VRAM	24GB	Initial Learning Rate	0.01
RAM	32GB	Weight Decay	0.0005
CUDA Version	11.6	NMS IoU	0.7
Torch Version	1.91	Warmup Epochs	3
Epoch	300	Warmup Momentum	0.8

From Fig.10, it can be seen that the robustness of Model V can meet the requirements of practical intangible cultural heritage application scenarios, whether in printed patterns or clear vector images. However, in detection tasks with dense targets and complex backgrounds, the model inevitably suffers from missed or false detections. For example, in Fig.10a, flower and butterfly patterns alternate and incomplete flower patterns appear, which greatly increases the difficulty of detection and leads to missed detections. Due to the complex background composition (as shown in Fig. 10d), multiple pattern categories (fruits, birds, butterflies) are present. Moreover, the three types of patterns of fruits, fish, and flowers have certain morphological similarities, leading to false positives.

Comparative experiment

In order to make the model deployment universal, we have chosen representative models of lightweight networks: YOLO series and SSD series

Table 4 | Model structure in ablation experiment

Model code	Basic network	Data enhancement	VOLO	MB Conv	Fused-MBConv
I	✓	✗	✗	✗	✗
II	✓	✓	✗	✗	✗
III	✓	✓	✓	✗	✗
IV	✓	✓	✓	✓	✗
V	✓	✓	✓	✗	✓
VI	✓	✓	✓	✓	✓

Table 5 | Results of ablation experiment.

Classification	Indicators	I	II	III	IV	V	VI
Butterfly	P	0.874	0.867	0.818	0.819	0.809	0.872
	R	0.781	0.773	0.832	0.811	0.811	0.777
	AP	0.863	0.866	0.867	0.874	0.863	0.842
Fish	P	0.573	0.547	0.564	0.519	0.574	0.576
	R	0.369	0.369	0.409	0.443	0.407	0.396
	AP	0.498	0.413	0.423	0.460	0.442	0.435
Dragon	P	0.677	0.670	0.685	0.688	0.704	0.690
	R	0.574	0.637	0.674	0.605	0.655	0.590
	AP	0.625	0.686	0.706	0.661	0.692	0.673
Bird	P	0.803	0.839	0.815	0.859	0.827	0.872
	R	0.803		0.828	0.803	0.779	0.804
	AP	0.855	0.853	0.871	0.868	0.867	0.872
Drum	P	0.788	0.911	0.892	0.835	0.813	0.882
	R	0.860	0.800	0.900	0.880	0.868	0.940
	AP	0.896	0.918	0.913	0.940	0.934	0.938
Flower	P	0.753	0.829	0.721	0.750	0.737	0.777
	R	0.719	0.662	0.755	0.729	0.758	0.742
	AP	0.765	0.802	0.809	0.793	0.801	0.808
Fruit	P	0.654	0.714	0.609	0.642	0.664	0.596
	R	0.464	0.495	0.545	0.495	0.612	0.505
	AP	0.513	0.573	0.573	0.536	0.636	0.521
All	P	0.732	0.768	0.729	0.730	0.733	0.752
	R	0.653	0.645	0.706	0.681	0.699	0.679
	mAP	0.716	0.729	0.737	0.733	0.748	0.727
	FPS /f.s-1	375	375	347	357	370	385
	Parameters/million	2.584	2.584	2.592	2.477	3.379	3.053
	Model size/MB	5.379	5.379	5.509	5.232	6.962	6.344

The bolded part in the table represents the optimal value of the indicator.

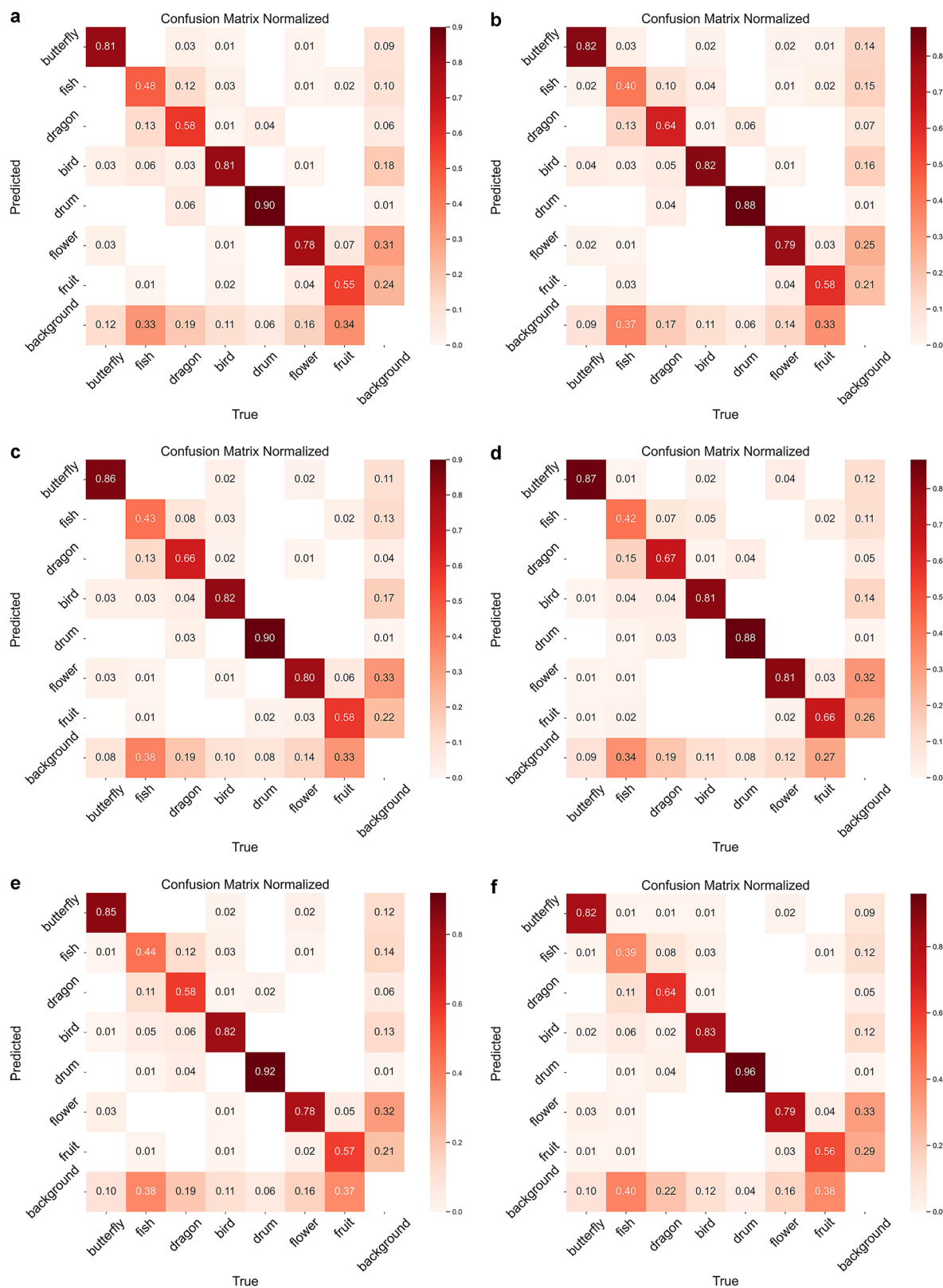


Fig. 9 | The confusion matrix effect of each model. a Model I confusion matrix. **b** Model II confusion matrix. **c** Model III confusion matrix. **d** Model IV confusion matrix. **e** Model V confusion matrix. **f** Model VI confusion matrix.

models as comparison targets. Both are classic One-stage object detection systems that have the advantage of balancing accuracy and speed. Among them, the YOLO series has selected YOLOv8n^{26,27}, YOLOv9t²⁸, YOLOv10n²⁹ and PP-YOLOE-Lite³⁰, which are relatively

mature and innovative applications, as well as the improved networks Ghost YOLO (based on YOLOv8n, optimized with Ghost convolution)³¹ and MobileViT YOLO (based on YOLOv7s, optimized with MobileViT)³² that have performed well in the engineering

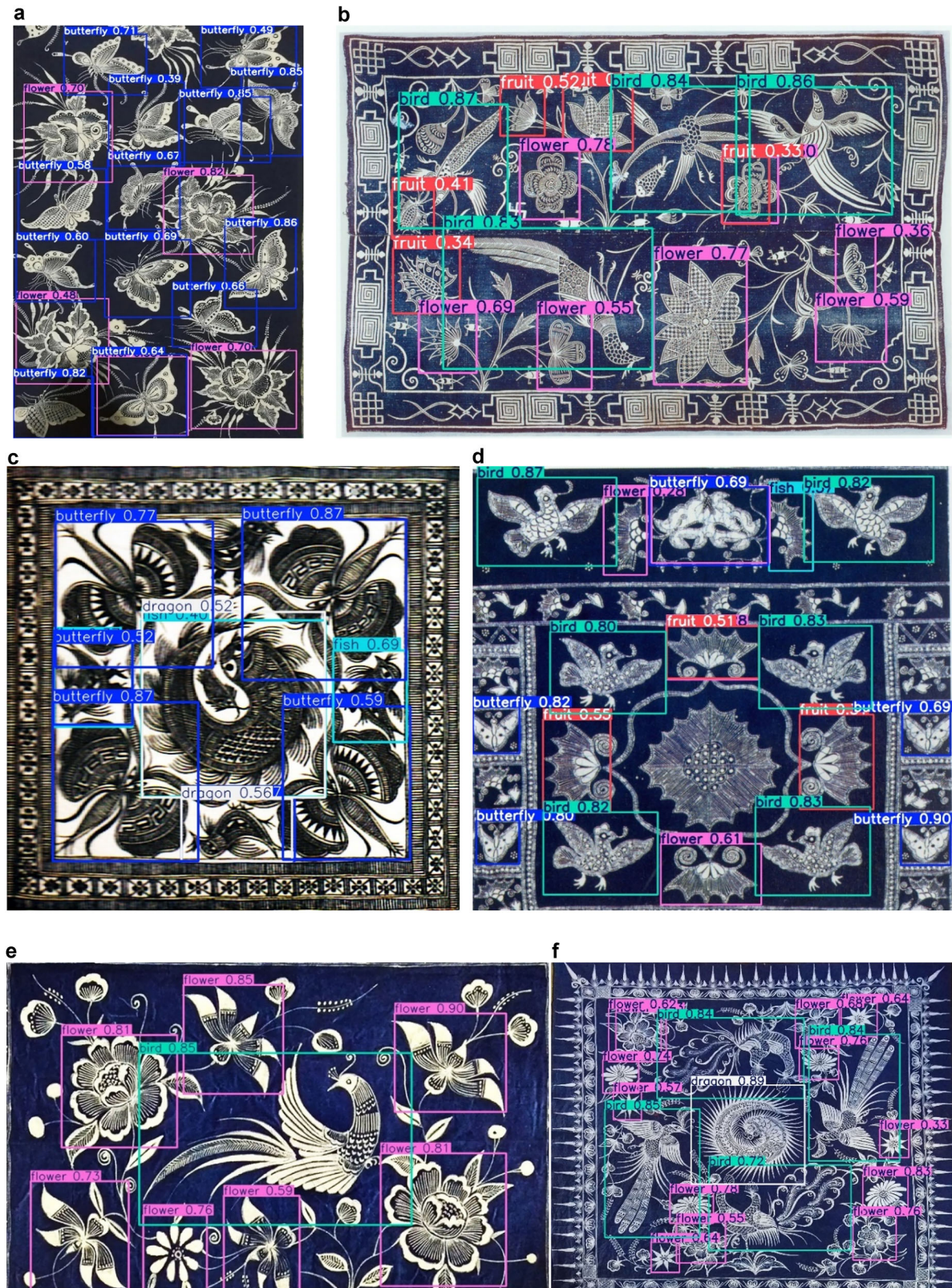


Fig. 10 | Example of detection effect. **a** Example Fig. 1. **b** Example Fig. 2. **c** Example Fig. 3. **d** Example Fig. 4. **e** Example Fig. 5. **f** Example Fig. 6.

field.SSD uses the most widely used classic structure MobileNetv2-SSD^{33,34}. To ensure the fairness of the comparison process, the experiment follows the following settings: first, all models in the comparison experiment were conducted in the hardware environment shown in Table 4, and the training process parameters were set

consistently. Second, the evaluation criteria are the same as those used in the ablation experiment, and no pre-training weights are used for training. The experimental results are shown in Table 6 and Fig. 11.

Summarizing the data in Table 6 and Fig. 11, the performance of each model is as follows:

Table 6 | Comparison of experimental results

Indicators	MobileNetv2-SSD	YOLOv8n	YOLOv9t	YOLOv10n	PP-YOLOE-Lite	Ghost YOLO	MobileVit YOLO	Ours
P	0.773	0.747	0.731	0.701	0.784	0.688	0.536	0.733
R	0.198	0.679	0.669	0.639	0.671	0.610	0.581	0.699
mAP	0.500	0.726	0.720	0.680	0.720	0.661	0.584	0.748
FPS /f.s-1	146	453	286	450	247	476	157	370
Parameters/million	3.926	2.686	1.731	2.700	8.364	1.394	8.580	3.379
Model size/MB	18.283	5.526	4.100	5.655	16.788	3.083	17.097	6.962

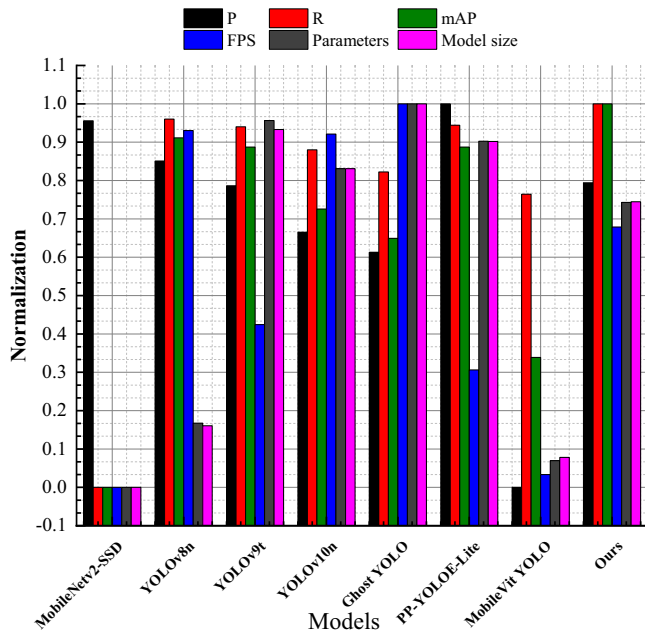


Fig. 11 | Normalization effect of comparative experimental indicators. The figure shows the specific performance of 8 models in 6 indicators. All indicators have been normalized.

First, although MobileNetv2-SSD is one of the most classic lightweight models, it can be seen that it no longer has the advantage of real-time performance. Excluding the improved MobileVit YOLO, the YOLO series models have comprehensively surpassed MobileNetv2-SSD in FPS, Parameters, and Model size metrics. In the object detection task, MobileNetv2-SSD achieved the best performance in the P index, reaching 0.773. However, the R index of the model is only 0.198, indicating that the model has ignored a large number of targets to be detected in its work. The model can only detect a small number of targets with significant features, resulting in obvious missed detections. Therefore, MobilNetv2-SSD is not suitable for detecting patterns in batik patterns with complex and dense backgrounds.

Second, the improved MobileVit YOLO and Ghost YOLO based on YOLOv7s and YOLOv8n also did not achieve good detection results. The former used more parameters in an attempt to improve the feature engineering quality of the model, but the values of P, R, and mAP were the lowest among the YOLO series models used. The latter used Ghost convolution to make the model lightweight, resulting in Ghost YOLO having the best lightweight performance, but the detection accuracy is not satisfactory. The situation of missed detections for both is still not optimistic.

Third, YOLOv8n, YOLOv9t, YOLOv10n, and PP-YOLOE-Lite achieved relatively good performance. All four models have reduced the probability of missed detections. YOLOv8n’s R and mAP metrics are second only to our proposed model, and its FPS performance is impressive.

YOLOv9t has only 0.227 million more parameters than Ghost YOLO, but its FPS metric is not outstanding. The performance of YOLOv10n in model accuracy is actually inferior to YOLOv9t, with an mAP index of only 0.680. Although PP-YOLOE-Lite performs the best in the P index, it has a large number of parameters and a relatively large model, resulting in average overall performance. This indicates that the generational differences of the YOLO series in the general dataset may not necessarily apply to the detection scenario in this paper.

The model proposed in this paper is the best in terms of detection accuracy indicators. Especially compared to other models, R shows a significant improvement, which enables the model to accurately detect target patterns in complex batik backgrounds. Compared to the five comparison models in Table 6, mAP increased by 49.6%, 3.89%, 10.00%, 13.16%, and 28.74% respectively. From the three lightweight indicators of FPS, Parameters, and Model size, the model still demonstrates relatively superior capabilities, indicating that it has the best overall performance. In summary, our proposed model combines high-precision and high-efficiency detection capabilities in batik pattern detection tasks, with low hardware dependence, and can be deployed at a lower cost in the vast majority of cultural dissemination applications and scenarios.

Batik detection and analysis system

To apply deep learning technology to practical applications in the cultural field, we have developed a prototype system for batik image detection and analysis. This system integrates an improved YOLOv11 pattern detection model and the batik pattern knowledge graph we previously constructed, achieving comprehensive functionality from image recognition to cultural analysis. The core interface of the system is shown in Fig. 12, primarily consisting of two functional modules: image detection and knowledge graph.

The image detection module (Fig. 12a) consists of three components: “Batik Image Loading”, “Detection Results”, and “Statistics”. Users upload batik images through the input interface, whereupon the improved YOLOv11 model performs automatic pattern detection. The visualization area displays detected patterns with category annotations, while the statistics panel provides quantitative analysis of pattern distribution and frequency.

The knowledge graph module (Fig. 12b) enables semantic exploration of detected patterns. Upon selecting a detected pattern, users access three integrated views: (1) Pattern-related nodes displaying semantic relationships from the knowledge graph, (2) Similar pattern examples retrieved from our dataset, and (3) Cultural information including prototype sources, symbolic meanings, and associated worship consciousness. This multi-dimensional presentation facilitates comprehensive understanding of pattern semantics.

Our system can provide practical tools for various heritage preservation needs. In digital archiving, it enables systematic documentation of patterns within batik items, automatically detecting and recording multiple pattern types with their positions and frequencies—replacing labor-intensive manual annotation. For cultural education and exhibition, users can photograph batik images and instantly access cultural connotations of detected patterns, such as symbolic meanings, worship consciousness, and

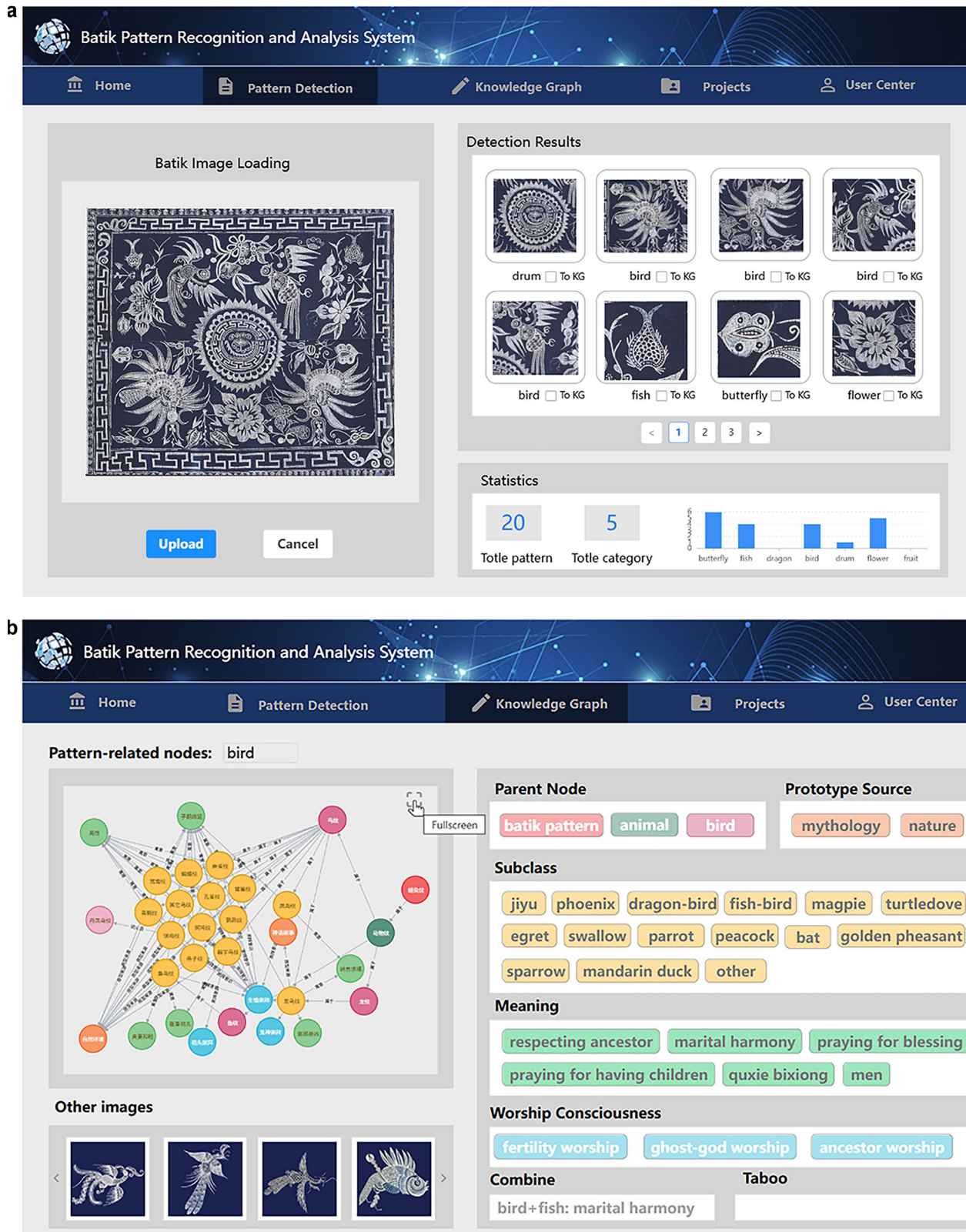


Fig. 12 | Prototype system for batik image detection and analysis. **a** Batik image detection interface. **b** The correlation aspect of the knowledge graph of batik patterns.

cultural taboos. For restoration work, the system can help identify damaged patterns and retrieve intact references.

Through developing this prototype, we establish a technical framework that connects visual recognition with cultural interpretation. The integration of pattern detection and semantic knowledge graphs preserves both the

visual forms and cultural meanings of batik patterns. By transforming pattern identification into comprehensive cultural analysis, the system offers researchers, cultural workers, and educators a practical tool for documenting and transmitting batik heritage knowledge, contributing to the sustainable preservation of this intangible cultural heritage.

Discussion

This paper presents an improved YOLOv11-based model for batik pattern detection that effectively balances detection accuracy with computational efficiency. The proposed method addresses the inherent challenges of detecting culturally significant patterns in complex batik images, where dense pattern distributions, scale variations, and degraded image quality pose substantial obstacles to conventional detection algorithms.

Our main contributions include: (1) construction of a comprehensive batik dataset with detailed pattern annotations, providing an essential resource for future research; (2) development of an improved YOLOv11 model integrating VOLO attention mechanisms and Fused-MBConv modules for improved feature representation; and (3) implementation of a prototype system bridging visual detection with cultural knowledge interpretation. Experimental results demonstrate the effectiveness of our approach, achieving significant improvements in detection accuracy while maintaining real-time performance, particularly for small-scale and densely distributed patterns.

While our work has made meaningful progress, there are still two main limitations that require further research. First, the current model has been trained and evaluated exclusively on Chinese batik patterns. Future work should assess its generalization capability across different batik traditions, including Indonesian, Malaysian, and Indian styles, to establish a cross-cultural detection framework that respects each tradition's unique characteristics while leveraging shared visual features. Second, although we have developed a prototype system, systematic usability validation remains to be conducted. We plan to gather feedback from heritage professionals, museum curators, and cultural researchers to refine the system's interface and functionality for practical deployment in heritage institutions.

This work contributes to the growing body of research on computer-assisted cultural preservation, demonstrating that computer vision techniques can effectively support the documentation and interpretation of intangible cultural heritage while respecting its complexity and cultural significance.

Data availability

The data supporting this study are available from the corresponding author upon reasonable request.

Code availability

The code supporting this study is available from the corresponding author upon reasonable request.

Received: 29 June 2025; Accepted: 22 February 2026;

Published online: 09 March 2026

References

- Chen, Z., Ren, X. & Zhang, Z. Cultural heritage as rural economic development: Batik production amongst China's Miao population. *J. Rural Stud.* **81**, 182–193 (2021).
- Han, D. & Cong, L. Miao traditional patterns: the origins and design transformation. *Vis. Stud.* **38**, 425–432 (2023).
- Dong, B. et al. Spatial distribution and tourism competition of intangible cultural heritage: take Guizhou, China as an example. *Herit. Sci.* **11**, 64 (2023).
- Bo, H. Study on the batik patterns and crafts of the Miao costumes in northwestern Guizhou Province. In 2nd International Conference on Advances in Social Science, Humanities, and Management (ASSHM-14), 250–254. <https://doi.org/10.2991/asshm-14.2014.69> (Atlantis Press, 2014).
- Zhennan, L. Y. U. & Yahaya, S. R. An aesthetic study on traditional batik design of Miao ethnicity in China. *Fakulti Seni, Komputeran dan Industri Kreatif, Universiti Pendidikan Sultan Idris, Tanjong Malim.* <https://doi.org/10.37134/kupasseni.vol9.2.2.2021> (2021).
- Quan, H., Li, Y., Liu, D. & Zhou, Y. Protection of Guizhou Miao batik culture based on knowledge graph and deep learning. *Herit. Sci.* **12**, 202–223 (2024).
- Wu, X. et al. Image-driven batik product knowledge graph construction. *npj Herit. Sci.* **13**, 20 (2025).
- Tian, G. et al. Auto-generation system based on fractal geometry for batik pattern design. *Appl. Sci.* **9**, 2383 (2019).
- Hu, T., Xie, Q., Yuan, Q., Lv, J. & Xiong, Q. Design of ethnic patterns based on shape grammar and artificial neural network. *Alexandria Eng. J.* **60**, 1601–1625 (2021).
- Lv, J. et al. Interactive genetic algorithm oriented toward the novel design of traditional patterns. *Information* **10**, 36 (2019).
- Quan, H. et al. Satisfying consumer preferences: the packaging design of Guizhou red sour soup. *Foods* **13**, 3806 (2024).
- Ding, N., Lv, J. & Hu, L. Research on national pattern reuse design and optimization method based on improved shape grammar. *Int. J. Comput. Intell. Syst.* **13**, 300–309 (2020).
- Ding, N., Lv, J. & Hu, L. Application of improved collaborative filtering algorithm in recommendation of batik products of Miao nationality. In IOP Conference Series: Materials Science and Engineering, Vol. 677, 022038. <https://doi.org/10.1088/1757-899x/677/2/022038> (IOP Publishing, 2019).
- Chen, D. & Cheng, P. A method to extract batik fabric pattern and elements. *J. Text. Inst.* **112**, 1093–1099 (2021).
- Yuan, Q., Xu, S. & Jian, L. A new method for retrieving batik shape patterns. *J. Assoc. Inf. Sci. Technol.* **69**, 578–599 (2018).
- Liu, S. & Chen, D. Computer simulation of batik printing patterns with cracks. *Text. Res. J.* **85**, 1972–1984 (2015).
- Elvitaria, L. et al. A proposed batik automatic classification system based on ensemble deep learning and GLCM feature extraction method. *Int. J. Adv. Comput. Sci. Appl.* **15**. <https://doi.org/10.14569/ijacsa.2024.0151058> (2024).
- Meranggi, D. G. T., Yudistira, N. & Sari, Y. A. Batik classification using convolutional neural network with data improvements. *JOIV: Int. J. Inform. Vis.* **6**, 6–11 (2022).
- Filia, B. J. et al. Improving batik pattern classification using CNN with advanced augmentation and oversampling on imbalanced dataset. *Procedia Comput. Sci.* **227**, 508–517 (2023).
- Gao, C. et al. Applying optimized YOLOv8 for heritage conservation: enhanced object detection in Jiangnan traditional private gardens. *Herit. Sci.* **12**. <https://doi.org/10.1186/s40494-024-01144-1> (2024).
- Karimi, N., Mishra, M. & Lourenço, P. B. Deep learning-based automated tile defect detection system for Portuguese cultural heritage buildings. *J. Cult. Herit.* **68**, 86–98 (2024).
- Darma, I. W. A. S., Suciati, N. & Siahaan, D. CARVING-DETC: A network scaling and NMS ensemble for Balinese carving motif detection method. *Vis. Inform.* **7**, 1–10 (2023).
- Khanam, R. & Hussain, M. YOLOv11: an overview of the key architectural enhancements. arXiv preprint <https://doi.org/10.48550/arXiv.2410.17725> (2024).
- Yuan, L., Hou, Q., Jiang, Z., Feng, J. & Yan, S. VOLO: Vision Outlooker for Visual Recognition, IEEE Trans. Pattern Anal. Mach. Intell. **45**, 6575–6586. <https://doi.org/10.1109/TPAMI.2022.3206108> (2023).
- Tan, M. & Le, Q. V. EfficientNetV2: smaller models and faster training. arXiv preprint <https://doi.org/10.48550/arXiv.2104.00298> (2021).
- Redmon, J., Bochkovskiy, A. YOLOv8: A New Generation of Real-Time Object Detection. arXiv preprint arXiv:2301.00045. (2023)
- Fan, Q. S. et al. LUD-YOLO: A novel lightweight object detection network for unmanned aerial vehicle. *Inf. Sci.* **686**, 121366. <https://doi.org/10.1016/j.ins.2024.121366> (Elsevier, 2025).
- Wang, C. Y., Yeh, I. H. & Liao, H. Y. M. YOLOv9: learning what you want to learn using programmable gradient information. In European Conference on Computer Vision, 1–20. 8-3 -031-72751-1_1 (Springer, Cham, Milan, Italy, 2024).
- Wang, A. Chen, H. Liu, L. Chen, K. Lin, Z. Han, J. & Ding, G. YOLOv10: real-time end-to-end object detection. In *Proc. 38th International*

- Conference on Neural Information Processing Systems (NIPS), 107984–108011. <https://doi.org/10.52202/079017-3429> (2024).
30. Chen, G. et al. PP-YOLOE: An Efficient and Accurate Object Detector for Remote Sensing Images. arXiv preprint arXiv:2203.16250. (2022).
 31. Cui, J. et al. Detection method of effective tillering of rice in field based on lightweight Ghost-YOLOv8 and smart phone. *Smart Agric* **6**, 98–107 (2024). SA202407012.
 32. Uddin, A. et al. MobileViT and YOLOv8: improving bone fracture detection and classification through deep learning. In 2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON). BECITHCON64160.2024.10962615. (IEEE, Dhaka, Bangladesh, 2024).
 33. Sandler, M. et al. MobileNetV2: inverted residuals and linear bottlenecks. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4510–4520. <https://doi.org/10.1109/CVPR.2018.00474> (IEEE, Salt Lake City, UT, USA, 2018).
 34. Wei, L. et al. SSD: single shot MultiBox detector. In European Conference on Computer Vision, 21–37. (Springer, Cham, Amsterdam, Netherlands, 2016).

Acknowledgements

This research was supported by Guizhou Provincial Basic Research Program (Natural Science) under grant Nos. [2024] Youth185, Joint Open Fund of Guizhou Provincial Department of Education Nos. [2022]436, Youth Science and Technology Talent Growth Project of Guizhou Provincial Department of Education Nos. qianjiaoji [2024]74, Guizhou Provincial Basic Research Program (Natural Science) MS[2025] 087.

Author contributions

Y.L. and H.Q. designed the study, performed the experiments, analyzed the results, and wrote the main manuscript. Q.L. provided valuable insights and suggestions on the methodology. H.Q. and J.W. prepared the dataset used in the study. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Huafeng Quan.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026