Deep learning-based recognition of Miao ethnic costumes via YOLOv5s: A step toward digital cultural preservation

DOI: 10.35530/IT.077.02.2025156

RUI GUO
TING CHEN
YAN HONG
XIANYI ZENG

ABSTRACT – REZUMAT

Deep learning-based recognition of Miao ethnic costumes via YOLOv5s: A step toward digital cultural preservation

Miao ethnic costumes, celebrated for rich diversity, intricate craftsmanship, and distinctive patterns, represent an important aspect of China's cultural heritage and the broader realm of intangible cultural heritage. In response to the growing need for digital preservation, this study proposes a deep learning-based approach to recognise and document Miao costumes effectively. While traditional costume recognition methods face challenges such as high computational costs and limited analytical capacity, the YOLOv5s framework offers automatic feature extraction and improved scalability. However, its standard form struggles to adequately focus on critical visual features, reducing recognition performance accuracy. To overcome this, we introduce the YOLOv5s-SED model, which incorporates a Squeeze-and-Excitation (SE) attention mechanism and Deformable convolution (DCNv2) into YOLOv5s to enhance feature representation and improve the recognition of fine details. A dedicated dataset of 4,468 annotated images was compiled, and the model was refined through hyperparameter tuning and comparative experiments. The results demonstrate notable performance gains, with precision increasing from 97.1% to 97.6%, recall from 99.3% to 99.8%, and mean Average Precision (mAP) from 70.7% to 71.5%. These outcomes highlight the model's strong generalisation ability in complex environments and its potential to support the digital preservation and promotion of Miao ethnic costumes.

Keywords: *Miao ethnic costumes, cultural heritage preservation, deep learning, YOLOv5s-SED, image recognition*

Recunoașterea costumelor tradiționale Miao pe baza învățării aprofundate prin intermediul YOLOv5s: un pas către conservarea culturală digitală

Costumele tradiționale Miao, apreciate pentru diversitatea bogată, măiestria complexă și modelele distinctive, reprezintă un aspect important al patrimoniului cultural al Chinei și al patrimoniului cultural imaterial în sens larg. Ca răspuns la nevoia tot mai mare de conservare digitală, acest studiu propune o abordare bazată pe învățarea aprofundată pentru recunoașterea și documentarea eficientă a costumelor Miao. În timp ce metodele tradiționale de recunoaștere a costumelor se confruntă cu provocări precum costurile de calcul ridicate și capacitatea analitică limitată, cadrul YOLOv5s oferă extragerea automată a caracteristicilor și o scalabilitate îmbunătățită. Cu toate acestea, forma sa standard se confruntă cu dificultăți în a se concentra adecvat asupra caracteristicilor vizuale critice, reducând precizia performanței de recunoaștere. Pentru a depăși această problemă, a fost introdus modelul YOLOv5s-SED, care încorporează un mecanism de atenție Squeeze-and-Excitation (SE) și convoluție deformabilă (DCNv2) în YOLOv5s pentru a îmbunătăți reprezentarea caracteristicilor și recunoașterea detaliilor fine. A fost compilat un set de date dedicat de 4.468 de imagini adnotate, iar modelul a fost perfecționat prin reglarea hiperparametrilor și experimente comparative. Rezultatele demonstrează îmbunătățiri notabile ale performanței, precizia crescând de la 97,1% la 97,6%, recall-ul de la 99,3% la 99,8%, iar precizia medie (mAP) de la 70,7% la 71,5%. Aceste rezultate evidențiază puternica capacitate de generalizare a modelului în medii complexe și potențialul său de a sprijini conservarea digitală și promovarea costumelor tradiționale Miao.

Cuvinte-cheie: *costume tradiționale Miao, conservarea patrimoniului cultural, învățare aprofundată YOLOv5s-SED, recunoașterea imaginilor*

INTRODUCTION

The Miao ethnic group, one of China's most ancient and culturally significant minorities, is predominantly distributed across Guizhou, Yunnan, Hunan, and neighbouring regions. Renowned for its intricate silver ornaments and vibrant costumes, it serves as a vital component of China's intangible cultural heritage and a symbolic representation of global cultural diversity. Miao ethnic costumes not only encapsulate the essence of the Miao people's "non-written civilisation", but also function as living testimonies to human cultural diversity [1]. Through distinctive material forms, these costumes reflect the ethnic group's historical memory, cosmological beliefs, and spiritual traditions, where specific patterns such as butterfly motifs symbolise ancestral spirits, and dragon designs represent water deities. This fusion of history, philosophy, and artisanal craftsmanship positions Miao costumes as essential carriers of the Miao cultural DNA [2]. Therefore, their preservation value extends

beyond the material realm, encapsulating the spiritual continuity of civilisation.

Indeed, Miao ethnic costumes, renowned for their distinctive materials, intricate techniques, and symbolic patterns, offer rich artistic expression and play an essential role in Chinese cultural and artistic traditions [3]. However, the Miao population's sub-tribal divisions and dispersed geographical distribution have resulted in a highly heterogeneous costume system [4]. Although this cultural diversity enhances their value, it also presents significant challenges to systematic protection. Traditional manual classification and visual recognition methods demonstrate inefficiency and poor generalisation [5].

Consequently, developing efficient and accurate recognition methods is crucial for the digital preservation and transmission of Miao costumes. More importantly, these methods fail to decode the semantic relationships between visual patterns and their cultural meanings, a gap our model aims to bridge by linking convolutional feature maps to symbolic annotations in the dataset [6].

With the advancement of intelligent recognition technologies, deep learning has emerged as a promising tool for automatically extracting visual features (e.g., texture, shape, colour) and integrating them with classification or recognition algorithms [4, 7, 8]. Traditional image processing and feature-based methods, such as Scale-Invariant Feature Transform (SIFT), Local Binary Patterns (LBP), and Gabor filter-based texture analysis, have been widely used in textile pattern recognition. However, these approaches rely on manually designed low-level features and fail to capture the deep cultural semantics embedded in symbolic motifs such as butterflies and dragons. Consequently, they offer shallow feature representation and limited interpretability. Moreover, conventional methods show poor adaptability to complex visual scenarios involving garment folds, silver ornament occlusion, and intricate embroidery, while variations in illumination, fabric deformation, and colour-layered structures further reduce their robustness. Despite advances in computer vision for textile inspection and fashion analytics, existing recognition systems still struggle to handle the visual and semantic complexity of traditional ethnic costumes. Most industrial frameworks remain optimised for standardised modern textiles and are thus incapable of interpreting the rich textures and symbolic ornamentation characteristic of heritage garments, creating a persistent gap in the digital documentation and intelligent management of cultural textiles within museums and heritage institutions.

YOLOv5s, as an advanced deep learning-based costume recognition framework, provides an effective solution for Miao costume recognition [9]. Compared to traditional methods, this architecture automatically extracts key garment features through end-to-end learning, with its convolutional neural network capturing high-level semantic characteristics including embroidery patterns, symmetrical designs, and silver ornament arrangements our implementation, visual features are aligned with ethnographic studies (e.g., tribal-specific pattern dictionaries) via a post-processing knowledge graph, establishing a bidirectional pipeline connecting visual recognition with cultural decoding. The framework also demonstrates high efficiency, robust spatial perception, and adaptability to small datasets, enabling better performance in complex scenes [10]. Additionally, the model supports techniques such as transfer learning and data augmentation, helping alleviate data scarcity in the domain of ethnic costume recognition.

However, despite its demonstrated advantages, YOLOv5s exhibits notable limitations in both feature channel selection and local detail representation when processing Miao costume imagery [11]. Specifically, the architecture's lack of an explicit attention mechanism results in suboptimal extraction of fine-grained details – particularly embroidery stitches and localised ornamental elements which serve as critical visual identifiers for accurate cultural recognition. A representative case involves the model's tendency to misclassify spiritually significant motifs such as the "fish-scale pattern" (a traditional symbol representing fertility in Miao culture) as generic texture artefacts when lacking specialised enhancement [12]. While the lightweight design confers computational efficiency benefits, this advantage comes at the cost of reduced capability to discern subtle local features under challenging conditions, including partial occlusion and fabric deformation [13]. Furthermore, constraints imposed by limited training dataset availability and inconsistent image quality frequently lead to either underfitting or overfitting scenarios, thereby substantially compromising the model's generalisation capacity. Overall, these identified limitations collectively impede the practical deployment of YOLOv5s for intangible cultural heritage costumes recognition applications, highlighting the urgent need for both architectural refinements and training protocol optimisations [14].

In response to these limitations, this study presents an enhanced deep learning model for high-precision Miao ethnic costume recognition. Specifically, our approach innovatively integrates the Squeeze-and-Excitation (SE) attention mechanism and Deformable Convolution v2 (DCNv2) into the YOLOv5s backbone network, achieving superior recognition accuracy and robustness. On one hand, the SE module dynamically models inter-channel dependencies, allowing the model to focus on culturally significant features and improve channel-wise feature representation. The SE-generated channel weights highlight convolutional filters activated by culturally diagnostic patterns (e.g., zigzag stitches encoding migration routes), effectively transforming low-level pixels into interpretable cultural symbols [14, 15]. On the other hand, DCNv2 substantially improves spatial adaptability by learning dynamic sampling locations that adjust according to object deformation and geometric variations, demonstrating exceptional performance in challenging scenarios involving fabric folds and ornament occlusion [16]. This capability proves crucial to maintaining semantic integrity, ensuring that culturally important but distorted motifs (e.g., partially obscured

"sunburst" patterns representing divinity) are correctly interpreted as complete symbolic units. The synergistic combination of these modules enhances both channel-wise feature representation and spatial feature extraction, delivering significant performance gains without compromising the model's real-time processing capability, thereby advancing digital heritage preservation methodologies.

To summarise, this study addresses key limitations in YOLOv5s —- insufficient feature extraction, weak local expression, and data scarcity – through an optimised framework integrating Squeeze-and-Excitation (SE) attention and Deformable Convolution v2 (DCNv2) modules. The enhanced model improves feature learning across both channel and spatial dimensions, significantly boosting recognition accuracy, robustness, and generalisation in complex environments. In addition to improving recognition accuracy, the proposed framework addresses the pressing need for efficient digital archiving of traditional textiles. Accurate costume recognition facilitates the categorisation, storage, and retrieval of cultural garments in digital databases, supporting long-term textile conservation and cultural heritage digitisation. By linking intelligent recognition technology with textile documentation, the study provides a technical foundation for sustainable cultural preservation. Crucially, it also establishes a technical bridge between computer vision and ethnography by enabling direct mapping between the model's attention heatmaps and cultural symbolism, such as high-attention regions corresponding to totemic patterns. This provides researchers with an effective tool for rapid annotation and interpretation of costume heritage while reducing reliance on large-scale, high-quality datasets. The framework offers a robust technical foundation for intelligently preserving and revitalising traditional ethnic cultures, successfully balancing accuracy with efficiency. Its promising potential extends to broader applications in multi-ethnic costume recognition and digital heritage systems. Innovations of this study include:

1. To address limited datasets and poor image quality, images of Miao ethnic costumes were collected via web crawlers and search engines, then annotated using LabelImg. A dataset of 4,468 images and annotations was constructed.
2. To improve recognition robustness and generalisation in complex scenarios, the YOLOv5s model was enhanced by incorporating the SE attention module and DCNv2 operator, improving the model's ability to learn inter-channel correlations and enhancing recognition accuracy.
3. Comparative experiments were conducted between the original and improved models under identical training and testing conditions. Model hyperparameters and optimisation strategies were adjusted, further improving recognition performance.

In conclusion, this study makes an important contribution to the intelligent recognition and cultural interpretation of Miao ethnic costumes. By integrating SE attention and DCNv2 modules into the YOLOv5s framework, it effectively enhances feature extraction, recognition accuracy, and cultural semantic understanding. The research not only advances technical innovation in ethnic costume recognition but also provides a valuable foundation for the digital preservation and revitalisation of intangible cultural heritage.

METHODOLOGY

Clothing recognition model and its working principle

The YOLOv5s-SED model comprises five core components: the Input module, Backbone, Neck, Head, and Output module, as illustrated in figure 1. Each component plays a specific role in the recognition of

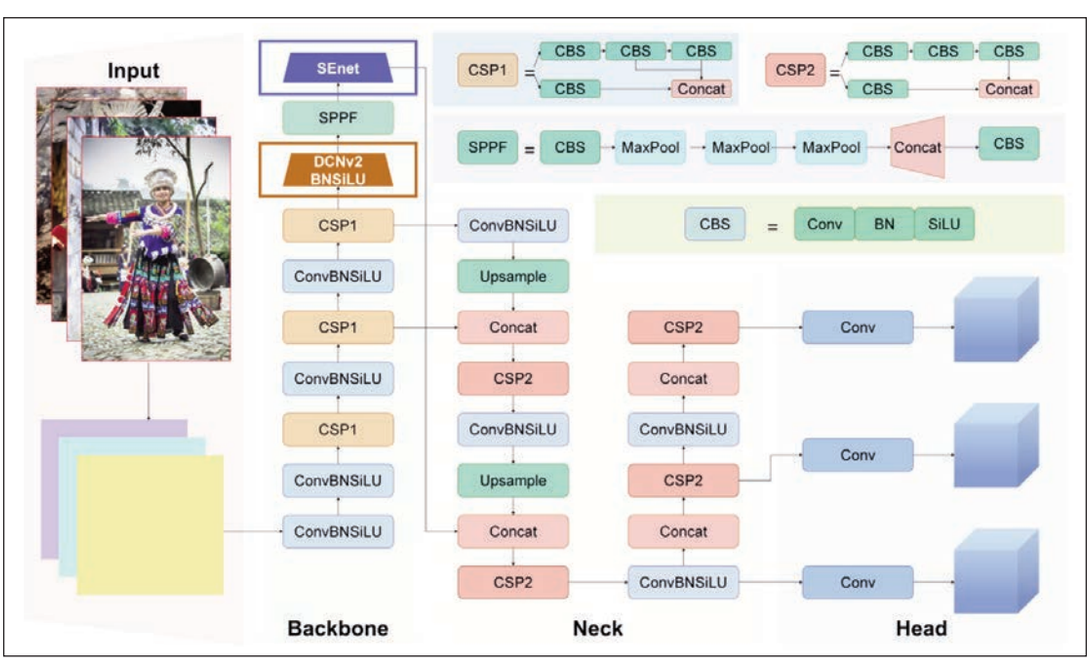

Fig. 1. Structure diagram of the YOLOv5s-SED model

Miao ethnic costumes, from data preprocessing to final output interpretation.

Input module: data augmentation and anchor optimisation

To enhance training data diversity and generalisation, the input stage employs Mosaic data augmentation, which combines four randomly selected images into a single composite input. This approach increases the model's ability to learn under varied lighting conditions, spatial arrangements, and object scales. Additionally, an adaptive anchor box generation strategy is implemented. It uses K-means clustering to align anchor sizes with the actual distribution of costume dimensions in the dataset, improving localisation precision across diverse garment shapes.

Backbone: feature extraction with attention and deformable convolution

The Backbone module extracts structural and semantic features from input images. To improve feature differentiation, a Squeeze-and-Excitation (SE) attention mechanism is embedded in the Spatial Pyramid Pooling-Fast (SPPF) layer. This module models inter-channel relationships, enabling the model to emphasise culturally relevant visual cues, such as embroidery patterns, while minimising background noise.

To handle garment deformation and occlusion, Deformable Convolution v2 (DCNv2) is selectively integrated into the deeper layers of the Backbone. While early layers retain standard convolution to preserve fine details, DCNv2 in later layers enables the model to adapt to non-rigid transformations such as folds or twists in clothing. This targeted design strikes a balance between detailed local feature capture and computational efficiency.

Neck: multi-scale feature fusion

The Neck module performs feature aggregation from multiple levels using a Feature Pyramid Network (FPN). This structure allows the model to maintain semantic consistency across different resolutions and enhances its ability to detect both fine and coarse features typical of Miao ethnic costumes. Through hierarchical fusion, the model adapts better to the scale and complexity of visual patterns.

Head: multi-scale detection

The Head module comprises three parallel detection branches tailored for different object sizes:

1. Small-Scale detection Head identifies distant or intricately detailed costume elements, leveraging high spatial resolution.
2. Medium-Scale detection Head balances resolution and receptive field to detect garments at moderate distances.
3. Large-Scale detection Head targets prominent costume features in close-up views, using broader receptive fields and rich semantic information.

This multi-scale architecture improves detection across varied viewing angles and spatial conditions common in real-world scenarios.

Output module: post-processing and result refinement

In the final stage, Non-Maximum Suppression (NMS) eliminates redundant bounding boxes by selecting only the highest-confidence detections. This results in clean, non-overlapping outputs with precise localisation and labelling of Miao costume components. The final predictions are optimised for accuracy and ready for application in heritage preservation.

SE attention mechanism

The SE attention mechanism refines feature maps by selectively emphasising informative regions. It includes two stages (figure 2).

In the squeeze stage, global average pooling is applied to each feature channel, compressing the spatial information into a single descriptor that reflects the overall importance of that channel. This

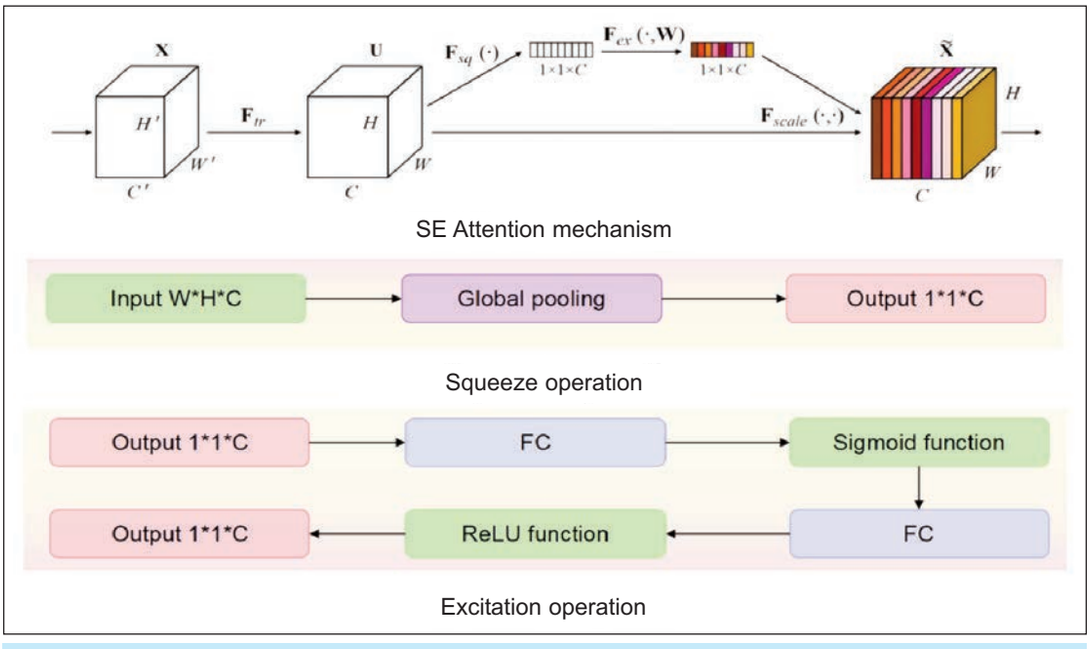

Fig. 2. SE attention mechanism and mechanism analysis

process enables the network to capture the relative significance of each feature channel within the global context. In the excitation stage, these aggregated descriptors are passed through fully connected layers, where non-linear transformations are performed to generate adaptive weights. These weights are then used to re-scale the original feature maps, allowing the model to emphasise informative features while suppressing less important ones, thus enhancing its representational capability [17].

This mechanism helps the model focus on culturally significant costume features, such as embroidery, symbolic motifs, and silver ornaments, while reducing noise from irrelevant background elements.

The process of the SE channel attention mechanism is as follows. Suppose the feature map of the input image X is U, where the dimension of U is $C\times H\times W$, where C is the number of channels, and H and W represent the height and width, respectively [18]. The mapping from the input image to the feature map is according to the following formula:

$$u_c = v_c \times X = \sum_{s=1}^{c'} v_c^s \times X_s \quad (1)$$

Squeeze phase

In the Squeeze operation, to alleviate the problem of channel dependence, the SE module performs global average pooling on the input feature map, uses the output of the previous layer as the input for the current layer, and compresses it into a $1\times1\times C$ feature vector. The pooled feature vector is denoted as $Z \in R^C$, where Z_c represents the compressed feature of channel c.

$$Z_c = F_{sq}(u_c) = \frac{1}{H\times X}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i=j) \quad (2)$$

Excitation stage

To fully leverage the compressed information obtained during the Squeeze operation, the Excitation stage is designed to model the dependencies between feature channels. Its primary goal is to selectively enhance informative features while suppressing irrelevant or redundant ones [19]. This process is implemented using two fully connected (FC) layers that serve as a lightweight gating mechanism, effectively capturing inter-channel relationships.

The output of the Excitation stage is a set of channel-wise attention weights, which are applied to the original feature map via element-wise multiplication [20]. This recalibrates the feature map, amplifying channels that carry semantically meaningful information and attenuating those that contribute less to the recognition task. As a result, the model's attention is directed toward discriminative visual cues, such as intricate embroidery or symbolic motifs, crucial for accurately identifying Miao ethnic costumes in complex backgrounds.

DCNv2 deformable convolution

Deformable convolution differs from standard convolution by introducing learnable offsets at each sampling point within the convolutional kernel [16]. This modification allows the network to adjust its sampling locations dynamically, resulting in a flexible and adaptive receptive field that better captures geometric variations during training. Consequently, the model is able to extract more informative and spatially diverse features from input images, particularly in scenarios involving irregular shapes or deformations common in traditional garments [21].

The following equation represents the output of a standard convolution operation:

$$Y = \sum_{P_n\in R} w(p_n)\cdot x(p_0 + p_n) \quad (3)$$

$$Y = \sum_{P_n\in R} w(p_n)\cdot x(p_0 + p_n + \Delta p_n) \quad (4)$$

As illustrated in figure 3, the incorporation of learnable offsets enables the sampling grid to deviate from the fixed, regular structure of standard convolution. This flexibility allows the model to better align with the contours and structures of irregular costume elements, thereby enhancing its ability to detect fine-grained and contextually significant features [16].

However, this flexibility introduces new challenges. In Deformable Convolution v1 (DCNv1), sampling positions are no longer fixed, which can result in the capture of irrelevant background information [22, 23]. This may dilute the model's focus and interfere with the learning of task-relevant features, particularly when foreground and background elements are closely intertwined.

To mitigate this issue, Deformable Convolution v2 (DCNv2) introduces a critical enhancement: the modulation scalar Δmk for each sampling point. In addition to the learnable offset Δpn, this scalar functions as an attention weight, enabling the network to

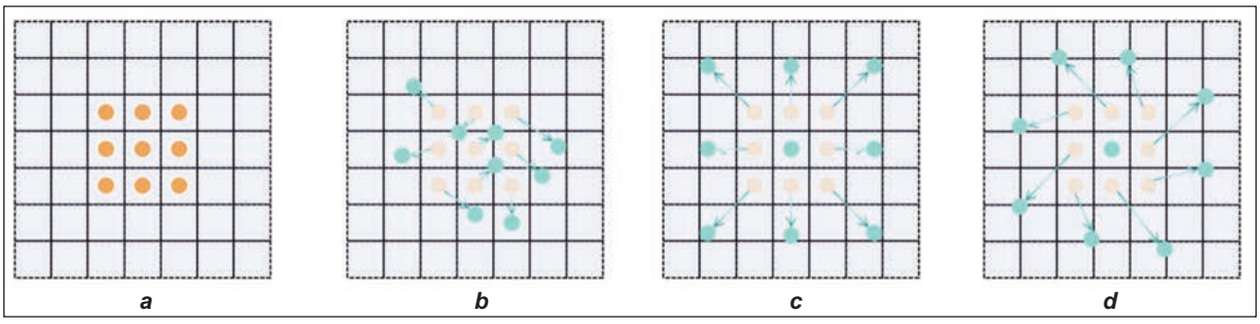

Fig. 3. Ordinary convolution sampling rules and variable convolution sampling rules: *a* – ordinary convolution sampling; *b* – the deformable sampling rule 1; *c* – the deformable sampling rule 2; *d* – the deformable sampling rule 3

emphasise important regions while down-weighting or ignoring irrelevant ones. Specifically, higher weights are assigned to sampling points within regions of interest, while near-zero weights are allocated to areas unrelated to the task, such as background noise. This attention-based refinement significantly enhances the model's ability to distinguish meaningful features from distracting elements in visually complex scenes, leading to more accurate and robust feature representation.

The mathematical formulation of DCNv2 can be expressed as:

$$Y = \sum_{P_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \cdot \Delta m_k \qquad (5)$$

The mathematical formulation of DCNv2 combines both spatial flexibility (via learned offsets) and selective weighting (via modulation scalars), enabling the model to dynamically adapt its focus based on the visual and contextual relevance of each region.

In this study, DCNv2 is incorporated to replace standard convolution in selected layers of the model. This substitution allows for a more complete and precise extraction of costume features, particularly those affected by occlusion, deformation, or variable perspectives. Ultimately, the use of DCNv2 strengthens the model's capacity to recognise Miao ethnic costumes under diverse and challenging visual conditions, enhancing both accuracy and robustness in real-world applications.

Dataset construction and processing

To overcome challenges associated with limited data availability and inconsistent image quality in the recognition of Miao ethnic costumes, this study implements a comprehensive and structured data collection strategy. The process begins with targeted keyword searches, such as "Miao ethnic costumes during festivals" and "details of traditional Miao ethnic costumes"-conducted on major search engines including Baidu and Google. These queries helped form a preliminary image dataset, emphasising authenticity and cultural specificity.

Building upon this foundation, a Python-based web crawler was developed to automate the collection of additional images. This crawler systematically retrieved data from a range of platforms, including websites dedicated to Miao ethnic heritage, traditional clothing e-commerce stores, and social media platforms like Weibo and Douyin. By targeting relevant pages and themes, the crawler extracted image URLs from the HTML structure of each webpage and downloaded the files in batches. Meanwhile, to avoid IP blocking or service denial, request intervals were carefully managed. After image collection, the data was further filtered to ensure quality and compliance. Images with low resolution (below 500×500 pixels), duplicates, or those subject to copyright restrictions were removed to ensure both technical quality and ethical compliance.

To compensate for the limitations of web-acquired data, field investigations were carried out in collaboration with local folk culture institutions. These efforts provided high-resolution photographs of rare and region-specific Miao costumes, captured during on-site visits. This component of the dataset adds valuable depth and authenticity, especially for styles that are underrepresented online.

After completing all image acquisition, a rigorous filtering process was applied to ensure dataset quality and representativeness. Only images that displayed a clear view of Miao ethnic costume features, across diverse body poses, camera angles, and complex background environments, were retained. Images lacking distinctive ethnic characteristics or depicting casual, non-traditional attire were excluded.

During the annotation phase, manual annotation was performed using LabelImg software. Annotators labelled key visual elements, including garment contours, silver ornaments, and embroidery motifs, features essential for accurate model training. Figure 4, a presents an overview of the constructed dataset, while figure 4, *b* illustrates the manual annotation process.

Fig. 4. Dataset section and annotation process: *a* – the dataset section constructed for this paper; *b* – the manual annotation process using LabelImg

Table 1

DATASET SECTION			
Costume category	**Image count**	**Data source**	**Description**
Ceremonial	1682	Online and field data	Encompasses attire for formal and ritual occasions
Daily	1456	Online and field data	Covers everyday, occupational, and children's attire
Craft-featured	1330	Authorised cultural resources	Features costumes with distinctive embroidery, batik, and silverwork
Total	**4468**	-	-

As a result, a curated and annotated dataset of 4,468 high-quality images was established. This dataset captures rich visual diversity and scene complexity, laying a strong foundation for enhancing the generalisation ability and recognition accuracy of the proposed deep learning-based Miao costume recognition model (table 1).

Experimental design

Ablation study

To identify the optimal integration point for Deformable Convolution v2 (DCNv2), different convolutional layers in the YOLOv5s backbone, already enhanced with the Squeeze-and-Excitation (SE) mechanism, were selectively replaced with DCNv2 layers. Four modified model variants were created, trained on the constructed dataset, and tested on the t1 image. By comparing recognition results across these versions, this experiment determines the most effective layer for incorporating DCNv2. The outcomes also support the theoretical framework discussed earlier in the paper.

Controlled experiment

The model variant that achieved the best performance in the ablation study was designated as YOLOv5s-SED. To thoroughly evaluate its effectiveness, both the baseline YOLOv5s model and the enhanced YOLOv5s-SED model were trained on the same dataset and tested on the t1 image. This direct comparison clearly validates the improvements in recognition accuracy and generalisation brought by the proposed enhancements, specifically in the context of Miao ethnic costume recognition.

Component analysis

To assess the individual impact of each architectural enhancement, two additional model variants were developed by independently incorporating the SE attention mechanism and DCNv2 into the baseline YOLOv5s model. Together with the YOLOv5s and YOLOv5s-SED models, all four versions were trained on the same dataset and evaluated using a representative sample of images. Ultimately, this experiment quantifies the individual and combined contributions of SE and DCNv2 to overall model performance, confirming that the integration of both components yields a significant advantage in complex visual recognition tasks.

RESULTS AND DISCUSSION

Ablation study

In this phase, the deformable convolutional layer DCNv2 was embedded into different layers of the backbone network to evaluate its impact on model performance. Four configurations were tested, each incorporating DCNv2 at progressively deeper layers. These models were designated SE-1, SE-2, SE-3, and SE-4.

Table 2

ABLATION STUDY RESULTS			
Model	**Precision (%)**	**Recall (%)**	**mAP (Mean Average Precision) (%)**
SE-1	97.6	98.9	67
SE-2	96.1	99.6	68.9
SE-3	97.1	99.3	68.2
SE-4	97.1	99.3	70.7

As shown in table 2, the SE-1 model achieves the highest precision (97.6%) on the training set. However, when evaluated on the test image t1, it fails to detect any Miao ethnic costumes (figure 5, *a*), revealing poor generalisation to unseen data. This suggests that placing DCNv2 in early layers disrupts fine-grained feature extraction critical for initial representation learning.

In contrast, SE-2 successfully detects all costume targets in the test image (figure 5, *b*), despite exhibiting slightly lower training precision. SE-3 shows moderate performance; it correctly identifies three costumes but also misclassifies a non-Miao outfit (figure 5, *c*), indicating susceptibility to false positives. SE-4 delivers the best balance between recognition performance and generalisation, achieving the highest mAP (70.7%) and accurately identifying all relevant targets in the test image (figure 5, *d*).

These findings suggest that integrating DCNv2 into the deeper backbone layers optimally enhances the model's spatial adaptability without compromising the integrity of early-stage feature learning. Conversely, introducing it prematurely in the shallow layers undermines recognition robustness.

Controlled experiment

Based on the ablation study, SE-4 was selected as the optimal configuration and named YOLOv5s-SED.

Fig. 5. Recognition results on image t1 using different model variants: *a* – SE-1; *b* – SE-2; *c* – SE-3; *d* – SE-4

Table 3

CONTROLLED EXPERIMENT RESULTS			
Model	**Precision (%)**	**Recall (%)**	**mAP (Mean Average Precision) (%)**
YOLOv5s	97.1	99.3	70.7
YOLOv5s-SED	97.6	99.8	71.5

A controlled experiment was conducted to compare the recognition performance of YOLOv5s-SED against the baseline YOLOv5s model.

As shown in table 3, YOLOv5s-SED surpasses the baseline in all key metrics. While YOLOv5s performs well under most conditions, figure 6, *a* reveals a critical limitation: it misses one of four Miao costumes in a densely packed test image, likely due to occlusion effects. YOLOv5s-SED, by contrast, achieves perfect recognition in the same scenario (figure 6, *b*), confirming its superior spatial reasoning.

Performance over training iterations is shown in figure 7, *b*. YOLOv5s-SED demonstrates a smoother and steeper mAP growth trajectory, converging after approximately 250 epochs with a final mAP of 0.715, an improvement over the baseline's 0.707.

In figure 7, *a*, the shorter the test time for one frame, the better the real-time performance of the model. In

Fig. 6. Recognition results on image t1: *a* – YOLOv5s; *b* – YOLOv5s-SED

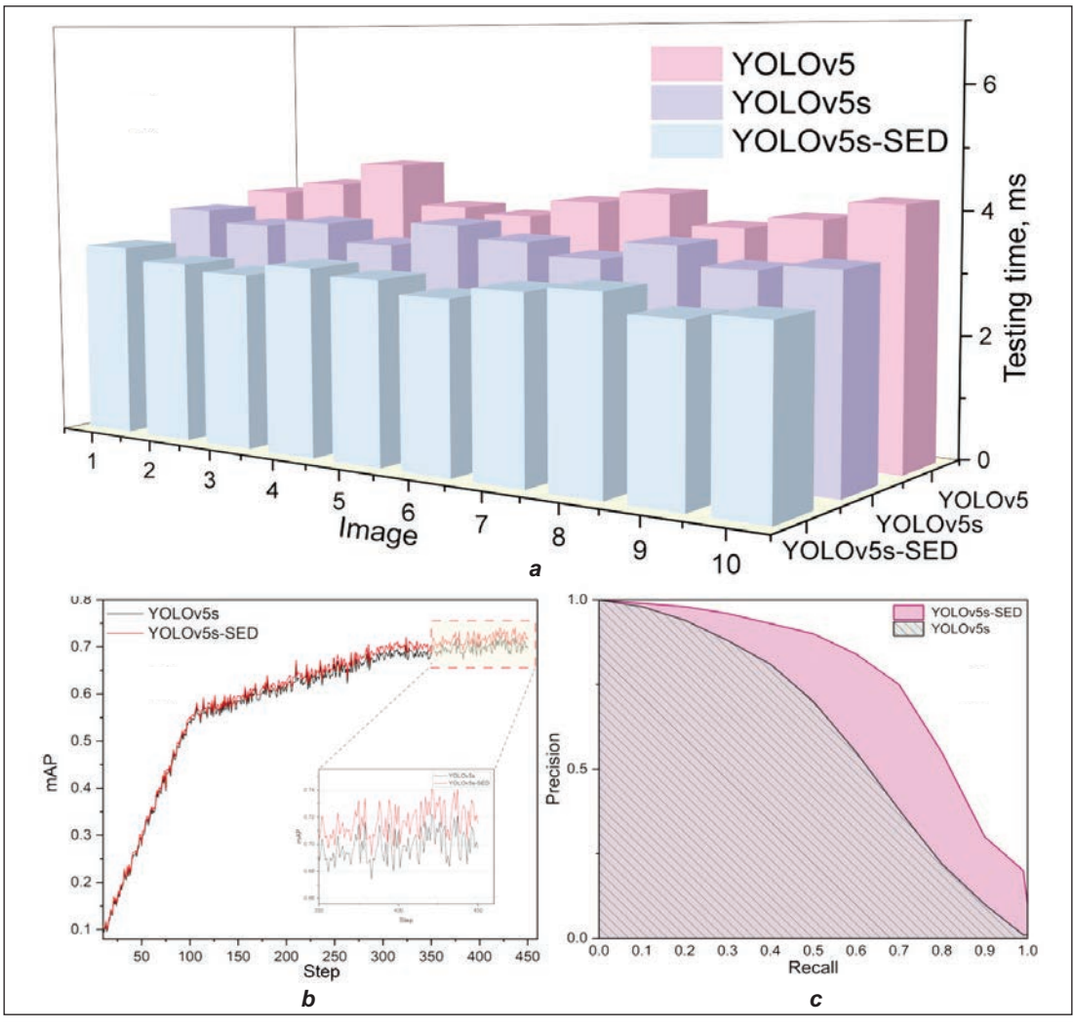

Fig. 7. Graphical representation of: *a* – comparison of experimental recognition speed; *b* – mAP@0.5 curve (the iteration count is scaled up by a factor of 350-450); *c* – comparison of R-P curves

real-time testing using 10 evaluation images, YOLOv5s-SED maintained lower inference latency and more consistent processing times (figure 7, *a*), while retaining a lightweight structure. This highlights its practicality for deployment in live cultural heritage recognition systems.

In figure 7, *c*, the larger the area under the curve, the better the average performance of the model at different thresholds, which also indicates a higher reliability of the model.

Furthermore, the Precision–Recall curve in figure 7, *c* illustrates the robustness of YOLOv5s-SED across classification thresholds. Its larger area under the curve (AUC) compared to YOLOv5s indicates a more reliable balance of precision and recall, with fewer false positives and missed recognitions.

Component analysis

To evaluate the individual contributions of SE attention and DCNv2, two additional variants of the baseline YOLOv5s were constructed, namely YOLOv5s + SE and YOLOv5s + DCNv2.

The baseline YOLOv5s performs reliably in low-complexity environments, showing high confidence and accurate localisation. However, under more challenging conditions, such as cluttered backgrounds or partial occlusion, its performance drops noticeably.

Table 4

COMPONENT ANALYSIS RESULTS			
Model	**Precision (%)**	**Recall (%)**	**mAP (Mean Average Precision) (%)**
YOLOv5s	97.1	99.3	70.7
+DCNv2	96.5	99.8	68.9
+SENet	96.9	99.3	69.2
YOLOv5s-SED	97.6	99.8	71.5

The YOLOv5s + DCNv2 variant demonstrates significantly enhanced recognition robustness, achieving higher IoU values, more stable confidence scores ($p<0.01$), and greater resilience to spatial deformations (figure 8, *b*). These results confirm DCNv2's superior adaptability to geometric variations in costume shapes and poses. In contrast, figure 8, *c* highlights common failure cases, such as missed costume regions and false recognitions on irrelevant backgrounds. This indicates that channel-only attention mechanisms, when lacking spatial modelling, may disrupt the crucial spatial–semantic alignment required for accurate costume recognition.

Fig. 8. Recognition results of component models on selected images

By integrating SE and DCNv2, the YOLOv5s-SED model achieves a synergistic enhancement in recognition performance. It effectively balances precision and recall, maintaining consistent accuracy and stability across the dataset. The integrated architecture strengthens both semantic sensitivity and spatial adaptability, making it exceptionally well-suited to the complex visual and structural features of Miao ethnic costumes.

CONCLUSIONS

This study focuses on the recognition of Miao ethnic costumes through deep learning, supported by the construction of a high-quality, annotated dataset comprising 4,468 images. To address the challenges of fine-grained costume recognition, we introduced an enhanced recognition model, YOLOv5s-SED, by incorporating the Squeeze-and-Excitation (SE) attention mechanism and Deformable Convolution v2 (DCNv2) into the YOLOv5s framework. As a result, this integration improves both semantic focus and spatial adaptability, leading to significant gains in recognition performance. The proposed model achieves a precision of 97.6%, a recall of 99.8%, and a mean Average Precision (mAP) of 71.5%, while substantially reducing missed recognitions and improving robustness across complex visual conditions.

However, despite these promising results, the current dataset is primarily composed of female ceremonial costumes, lacking representation of male garments, casual wear, and regional variations within the Miao community. Consequently, this limited diversity affects the model's generalisation capability. Additionally, the computational complexity introduced by the architectural enhancements poses challenges for real-time deployment on resource-constrained devices.

The proposed recognition framework holds significant potential for real-world applications in textile museums, archives, and cultural heritage institutions. By integrating the model into digital documentation workflows, curators and researchers can efficiently catalogue, retrieve, and interpret costume artefacts with high accuracy, thereby enhancing the accessibility and sustainability of textile heritage through intelligent digitisation. Looking ahead, future work will focus on expanding the dataset to encompass a wider variety of costume types, cultural subgroups, and environmental contexts, as well as exploring optimisation strategies such as depth-wise separable convolutions and knowledge distillation to reduce model size and inference latency. Furthermore, potential application directions include the use of augmented reality (AR) for immersive virtual try-on experiences and automated pattern generation for cultural design reproduction. Collectively, these efforts aim to advance the transition from static digital preservation toward dynamic cultural revitalisation, establishing a robust technological foundation for the sustainable safeguarding and innovation of Miao intangible cultural heritage.

REFERENCES

[1] Shi, T., Wu, X.H., Wang, D.B., Lei, Y., *The Miao in China: A review of developments and achievements over seventy years*, In: Hmong Studies Journal, 2019, 20, 1–23
[2] Quan, H., Li, Y., Liu, D., Zhou, Y., *Protection of Guizhou Miao batik culture based on knowledge graph and deep learning,* In: Heritage Science, 2024,12, 1, 202
[3] Mao, M., Pengli, W., *Cultural Inheritance of Miao Nationality in Western Hubei under the Background of Beautiful Countryside Construction,* In: Journal of Landscape Research, 2020, 12, 6, 83–94
[4] Hu, Y., Wu, S., Ma, Z., Cheng, S., I*ntegrating deep learning and machine learning for ceramic artifact classification and market value prediction,* In: Heritage Science, 2025, 13, 1, 1–17
[5] Wiley, V., Lucas, T., *Computer vision and image processing: a paper review,* In: International Journal of Artificial Intelligence Research, 2018, 2, 1, 29–36
[6] Zhang, Y., Zhao, H., Qi, L., Zhang, J., Zhang, T., *Research on the co-occurrence feature mining of the Qing Dynasty embroidery patterns based on temporal multilayer networks*, In: Heritage Science, 2025,13, 1, 1–19
[7] Nalbant, K.G., Bozkurt, B., *Application of machine learning methodology for textile defect detection,* In: Industria Textila, 2025, 76, 3, 372–386, https://doi.org/10.35530/IT.076.03.2024108
[8] Li, T., Lyu, Y.-X., Ma, L., Xie, Y., Zou, F.-Y., *Research on garment flat multi-component recognition based on Mask R-CNN,* In: Industria Textila, 2023, 74, 1, 49–56
[9] Luo, Q., Xu, X., Gao, H., *Research on the Inheritance of Miao Costume Culture Based on Digital Wireless Communication Technology*, In: Mobile Information Systems, 2022, 1, 4052341
[10] Hong, Y., Bruniaux, P., Zeng, X., Curteza, A., Liu, K., *Design and evaluation of personalized garment block for atypical morphology using the knowledge-supported virtual simulation method*, In: Textile Research Journal, 2018, 88, 15, 1721–1734
[11] Zhang, J., Zhang, Y., Liu, J., Lan, Y., Zhang, T., *Human figure detection in Han portrait stone images via enhanced YOLO-v5,* In: Heritage Science, 2024, 12, 1, 19
[12] Zhang, C., Wu, S., Chen, J., I*dentification of Miao embroidery in southeast Guizhou province of China based on convolution neural network*, In: Autex Research Journal, 2021, 21, 2, 198–206
[13] Jalandoni, A., Zhang, Y., Zaidi, N.A., *On the use of Machine Learning methods in rock art research with application to automatic painted rock art identification*, In: Journal of Archaeological Science, 2022, 144, 105629
[14] Zhan, J., Meng, Y., Zhang, L., Li, K., Yan, F., *Research on computer vision in intelligent damage monitoring of heritage conservation: the case of Yungang Cave Paintings,* In: Heritage Science, 2025, 13, 1, 50
[15] Liang, Y., Tohti, T., Hamdulla, A., *Multimodal false information detection method based on Text-CNN and SE module,* In: Plos one, 2022, 17, 11, e0277463
[16] Zhu, X., Hu, H., Lin, S., Dai, J., *Deformable convnets v2: More deformable, better results*, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019
[17] Wenyi, H., Hongkun, W., Yujia, D., *Identification Method of Tomato Diseases and Pests Based on SE Module and ResNet,* In: Agricultural Engineering, 2022, 12, 9, 33–40
[18] Munjal, B., Amin, S., Tombari, F., Galasso, F., *Query-guided end-to-end person search,* In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019
[19] Chao, X., Hu, X., Feng, J., Zhang, Z., Wang, M., He, D., *Construction of apple leaf diseases identification networks based on Xception fused by SE module*, In: Applied Sciences, 2021, 11, 10, 4614
[20] Gao, W., Shan, M., Song, N., Fan, B., Fang, Y., *Detection of microaneurysms in fundus images based on improved YOLOv4 with SENet embedded*, In: Journal of Biomedical Engineering, 2022, 39, 4, 713–720
[21] Chen, K., Wang, H., *MDCYOLO: Improved YOLOv5 Algorithm with Modified Deformable Convolution,* In: Proceedings of the 2023 4th International Conference on Machine Learning and Computer Application, 2023
[22] Liu, C., Hu, X., *Deep neural network with deformable convolution and side window convolution for image denoising,* In: Pattern Recognition Letters, 2023, 171, 92–98
[23] Mathew, D., Brintha, N.C., *Detection of garment manufacturing defects using CFPNet and deep belief network: an image-based approach*, In: Industria Textila, 2025, 76, 2, 160–170, https://doi.org/10.35530/IT.076.02.2024140

Authors:

RUI GUO[1,†], TING CHEN[2,†], YAN HONG[2], XIANYI ZENG[3]

[1]School of Fashion, Henan University of Engineering, Zhengzhou, Henan, 451191, China
e-mail: 58056753@qq.com

[2]College of Textile and Clothing Engineering, Soochow University, Suzhou 215021, China
e-mail: 20245215101@stu.suda.edu.cn

[3]Laboratoire de Génie et Matériaux Textiles (GEMTEX), University of Lille, ENSAIT, Roubaix 59056, France
e-mail: xianyi.zeng@ensait.fr

[†]RUI GUO and TING CHEN contributed equally to this work and should be regarded as co-first authors

Corresponding author:

YAN HONG
e-mail: hongyan@suda.edu.cn