TREATISES AND DOCUMENTS JOURNAL OF ETHNIC STUDIES
RAZPRAVE IN GRADIVO REVIJA ZA NARODNOSTNA VPRAŠANJA
96 / 2026, p. 201–243

DOI:10.2478/tdjes-2026-0007
Received 1. 11. 2025, Accepted 30. 03. 2026

Yunwu He,[1] Lan Thi Nguyen,[1] Wirapong Chansanam[1]

Intelligent Recognition of Ethnic Costumes Using YOLOv11: A Deep Learning Framework for Cultural Heritage Preservation

Abstract

The rapid digitization of cultural heritage creates new opportunities to preserve the visual and symbolic richness of ethnic traditions. However, accurate recognition of ethnic costumes remains challenging due to complex textures, overlapping patterns, and high inter-group similarity. This study proposes an intelligent recognition framework based on the YOLOv11 architecture for the digital preservation of multi-ethnic attire in Lijiang, China. By integrating a C2PSA spatial attention mechanism with multi-scale feature fusion, the model enhances discrimination of fine-grained textile structures under complex visual conditions. A large-scale dataset containing 4,974 images from 55 ethnic branches was constructed for evaluation. Experimental results demonstrate that the proposed method achieves an mAP@0.5 of 98.416% and a recall of 95.923%, significantly outperforming YOLOv5 and YOLOv4 ($p < 0.001$). With only 5.8 million parameters and 16.2 GFLOPs, the model enables efficient real-time deployment, contributing a robust AI-driven solution for cultural heritage informatics.

Keywords

intangible cultural heritage, deep learning, YOLOv11, ethnic costume recognition, spatial attention mechanism

Corresponding author

Wirapong Chansanam, Khon Kaen University, Faculty of Humanities and Social Sciences, Department of Information Science, Khon Kaen, Thailand; **e-mail**: wirach@kku.ac.th; **ORCID**: 0000-0001-5546-8485

[1] Faculty of Humanities and Social Sciences, Department of Information Science, Khon Kaen, Thailand.

1. Introduction

It can be seen that more than 40% of the world's intangible cultural heritage artefacts face the risk of irreversible loss due to insufficient digital preservation strategies every year (National Library of Australia 2003). While many nations have adopted artificial intelligence (AI) to safeguard cultural memory, the digital preservation of ethnic costumes – complex, visually rich embodiments of identity – remains relatively underexplored. Deep learning has emerged as a transformative tool in heritage informatics, yet its application to textile-based cultural artefacts still lags behind other domains such as architecture or sculpture restoration (Sun et al. 2025; Zhang & Li 2025). For instance, convolutional neural networks (CNNs) have been used to restore Italian mural paintings and detect Maasai costume motifs, thereby improving cross-regional archiving efficiency (Farajzadeh & Hashemzadeh 2021; Ge et al. 2023; Narag et al. 2025). However, similar success has not been realized in Southwest China's multi-ethnic textile traditions, where visual overlap and complex patterns impede accurate classification.

Ethnic costumes serve as a visual language of cultural continuity, encapsulating a community's history, aesthetics, and worldview. In Lijiang – a region home to the Naxi, Yi, and Bai peoples – traditional garments like the Naxi seven-star shawl and Bai tie-dye robes represent living narratives of labour, nature, and spirituality (Sang 2021; Chang et al. 2024; Ma 2025). Yet, the visual intricacy and inter-group similarity among these costumes create challenges for both manual annotation and algorithmic recognition. Previous studies highlight the subjective inconsistency of human experts in classifying textile-based heritage (Liu & Lu 2024), a limitation also observed in Zhang and Li's (2025) analysis of Surkhandarya ethnic attire. These problems underscore the need for a scalable, objective, and automated recognition system that preserves cultural fidelity while maintaining analytic precision.

Despite growing global attention, traditional documentation methods relying on manual curation and expert tagging remain inefficient and error-prone (Lei et al. 2020). Even with early deep learning models such as VGGNet (Simonyan & Zisserman 2015), preprocessing requirements and limited data diversity constrained performance. Lei et al. (2020) observed that VGG-based feature extraction suffered from "small sample bias," hindering large-scale deployment. In parallel, Lijiang's local cultural report (2023) warns of declining knowledge transmission: the average artisan age now exceeds 65, and younger generations often perceive traditional garments as mere tourist commodities.

Without technological intervention, this knowledge gap threatens the intergenerational transmission of intangible heritage.

Image recognition of clothing has increasingly been recognized as a powerful means of supporting cultural understanding, preservation, and intelligent visual analysis, particularly in the context of ethnic costumes and fine-grained style interpretation. Early studies demonstrated that traditional pattern recognition techniques, such as principal component analysis combined with neural networks and wavelet transforms, could effectively extract discriminative features from minority-costume images, thereby improving recognition accuracy and efficiency (Juan 2021). From a visual interaction perspective, colour-based feature fusion and clustering methods, together with classifiers such as support vector machines, have further revealed that chromatic structures and totemic patterns play a crucial role in identifying ethnic attributes and enhancing classification performance (Sun et al. 2021). Complementarily, fashion-oriented research grounded in clothing design theory has shown that visually differentiable style elements and their computational representations enable robust genre-level recognition, achieving high precision and recall by bridging semantic design concepts with computer vision features (Hidayati et al. 2018). With the rapid advancement of deep learning, convolutional neural networks and attention mechanisms have become central to modelling complex visual patterns, not only in cultural imagery but also in challenging real-world scenarios such as face recognition under noise and compression constraints (Abdulrahman & Tahir 2021) and real-time fall detection, where attention-enhanced YOLO (You Only Look Once) architectures significantly improve small-object sensitivity, robustness, and deployment efficiency (Yang et al. 2025). Collectively, these studies highlight a progressive shift from handcrafted features to attention-driven deep models, underscoring the importance of multi-scale representation, discriminative feature learning, and real-time capability for complex visual understanding tasks, thus providing a strong methodological foundation for advancing intelligent recognition of ethnic costumes in both cultural heritage and practical application contexts.

Recent breakthroughs in computer vision have provided promising avenues. CNN-based models have demonstrated strong performance in recognizing ethnic attire, such as Ding et al.'s (2023) FPA-CNN achieving 89.3% accuracy for She costumes, and a Tsinghua University team attaining 92.6% for Miao hundred-bird clothes. The introduction of YOLO architectures (Bochkovskiy et al. 2020; Shao et al. 2022) has shifted focus toward real-time detection, balancing speed and accuracy. YOLOv4 and

YOLOv5 have been adapted for fashion and cultural artefact recognition (Lee & Lin 2021; Su et al. 2025), while Ma et al. (2025) demonstrated YOLOv9's capability in detecting small relics within cluttered museum environments. However, these models still struggle with fine-grained textile distinctions, such as differentiating between Naxi and Yi embroidery, which often share structural motifs but differ in stitch density and colour gradients.

This methodological gap – the absence of a high-precision recognition framework that integrates spatial attention with the latest YOLO architecture – represents a critical barrier to progress. Although Geetha (2024) employed YOLOv6 for cultural artefact recognition, their approach did not explicitly address inter-group costume similarity or optimize attention for fabric-level detail extraction. Similarly, Ma et al. (2025) emphasized the difficulty of detecting artefacts within complex visual backgrounds, a scenario directly analogous to field photography in Lijiang's open-air heritage sites. These findings converge on a shared limitation: existing models are not tailored to the nuanced visual semantics of ethnic costume patterns, which require fine-grained attention to local texture and global symmetry.

This study explores how advanced deep learning – specifically the YOLOv11 architecture integrated with a task-oriented C2PSA spatial attention mechanism – can overcome these limitations to achieve efficient and accurate recognition of Lijiang's ethnic costumes. Unlike prior YOLO--based attention enhancements that primarily target generic object saliency, the proposed framework introduces a costume-aware attention adaptation strategy that emphasizes fine-grained textile structures, local embroidery density, and global garment symmetry. By embedding C2PSA attention into intermediate feature fusion stages of YOLOv11, the model selectively amplifies discriminative regions while suppressing visually redundant patterns common across ethnic groups. This design enables the network to better resolve subtle inter-group differences – such as stitch density variations and colour-gradient transitions – under complex backgrounds and partial occlusion.

By constructing a large-scale multi-ethnic image dataset and validating a YOLOv11-based recognition framework, the study aims to (1) enhance accuracy in distinguishing visually similar ethnic garments, (2) ensure real-time detection suitable for museum and field environments, and (3) establish a comprehensive evaluation protocol addressing visual similarity and environmental variability. In addition, a multi-phase training strategy and attention-consistency constraint are adopted to stabilize convergence and improve robustness across heterogeneous

cultural scenes, further strengthening the model's practical deployability.

The significance of this work extends across theoretical, methodological, and practical dimensions. Theoretically, it advances the field of cultural heritage informatics by operationalizing deep learning for textile-based identity representation, responding to Zhang and Li's (2025) call for scalable digital preservation tools. Methodologically, it bridges a gap between global YOLO advancements and regional ethnographic specificity – not merely by integrating attention modules, but by tailoring spatial attention to the semantic granularity of ethnic costume patterns, building on Yang et al.'s (2025) spatial attention optimization and He et al.'s (2016) residual learning principles. Practically, the model's modular and real-time design enables applications in intelligent museum guides, cultural education, and creative industries, supporting Shao et al.'s (2022) recommendation for deployable YOLO-based mobile systems. In essence, this research contributes both to preserving intangible cultural heritage and to advancing computer vision models capable of understanding the aesthetic logic and symbolic structure of ethnic identity – transforming preservation from static documentation to dynamic, intelligent recognition.

2. Literature Review

2.1 Deep Learning Methods for Image Recognition

Deep learning has become central to image recognition, driven mainly by two paradigms: Convolutional Neural Networks (CNNs) for fine-grained classification and one-stage object detection models for real-time visual analysis. Foundational work by He et al. (2016) introduced residual learning to improve feature propagation and mitigate gradient vanishing, laying the groundwork for subsequent advancements. Building on this, Wang et al. (2024) proposed a dynamic attention-augmented CNN that surpassed static modules such as CBAM (Woo et al. 2018) by 4.1%, achieving 92.3% accuracy in traditional costume classification. This highlights the importance of adaptive attention for visual fidelity in cultural heritage imagery.

Parallel progress in YOLO architectures has refined the balance between accuracy and efficiency. From Redmon et al. (2016) to recent iterations, YOLO has evolved into a robust framework for real-time detection. YOLOv4 (Bochkovskiy et al. 2020) improved backbone connectivity for apparel detection. YOLOv5 has been enhanced with C2PSA spatial attention mechanisms to improve real-time performance in ethnic

costume recognition (Su et al. 2025); and YOLOv6 achieved 86.5% mAP (mean average precision) for artefact detection (Geetha 2024). Subsequent improvements in YOLOv9 (Ma et al. 2025) and YOLOv10 (Wang et al. 2024) further enhanced adaptability and computational efficiency. The latest YOLOv11 (Wang et al. 2024) incorporates C2PSA attention mechanisms to improve small-target detection, reporting a 7.2% accuracy improvement over YOLOv8 and demonstrating effectiveness in visually complex environments such as field photography of ethnic attire.

2.2 Research Status and Gaps in Ethnic Costume Recognition

Although research on ethnic costume recognition has advanced, three key gaps remain.

First, limited multi-ethnic coverage restricts generalization. Most studies focus on a small number of groups – e.g., Ji et al. (2024) classified Zhuang attire with 90.5% accuracy, and Lei et al. (2020) identified Miao components with 85.2% precision. Broader coverage remains rare; Gan et al. (2022) reported that few studies have included more than ten ethnic groups. Consequently, the diverse attire of Lijiang's Naxi, Bai, and Yi communities remains underrepresented. Second, robustness under real-world conditions is insufficient. Most models perform well in controlled settings but degrade under variable lighting or occlusion. Jin et al. (2024) found that only a small minority of studies have addressed these challenges, and Liu & Lu (2024) emphasized misclassification from inter-group similarities, such as the brocade patterns of Yi and Bai costumes. Third, dataset limitations persist. Small, low-diversity datasets dominate, Gan et al. (2022) noted that a large proportion of datasets contain relatively few images, restricting model generalization. These gaps highlight the need for a high-precision, multi-ethnic recognition framework integrating spatial attention and advanced YOLO architectures to address inter-group similarity and complex visual environments.

To further contextualize the research gaps discussed above, it is necessary to illustrate the visual and symbolic richness of Lijiang's multi-ethnic attire. As shown in Figure 1, traditional costumes from different ethnic groups in the Lijiang region exhibit distinctive yet sometimes visually similar elements, including colour schemes, embroidery motifs, garment structures, and accessories. These characteristics reflect deep cultural symbolism while also leading to high inter-class similarity and intra-class variation, which pose significant challenges for fine-grained

ethnic costume recognition. The visual complexity highlighted in these examples underscores the practical motivation and research value of developing a robust and discriminative recognition framework.

Figure 1: Example images of visual elements in ethnic costumes

Source: News.qq 2024.

Figure 1 presents representative examples of multi-ethnic attire from Lijiang, illustrating distinctive visual and symbolic features across ethnic groups. The examples highlight both inter-class similarity and intra-class variation, motivating the development of the proposed recognition framework.

2.3 Thematic Analysis of Related Studies

Based on a systematic review of recent studies retrieved from major Chinese and international academic databases (e.g., CNKI, Web of Science, IEEE Xplore, and SpringerLink), the selected literature was analysed and organized into three thematic categories to clarify prevailing research trends and identify existing limitations in ethnic costume recognition.

2.3.1 Algorithm Evolution and Optimization

This category focuses on the architectural improvement and performance comparison of YOLO-series models for object detection. Studies have explored enhanced backbone connectivity, attention mechanisms,

and lightweight design to balance detection accuracy and computational efficiency. For example, improvements in YOLOv4 and YOLOv5 introduced optimized feature fusion strategies for apparel detection (Bochkovskiy et al. 2020; Su et al. 2025), while more recent versions such as YOLOv9 and YOLOv10 further improved adaptability and inference efficiency (Wang et al. 2024; Ma et al. 2025). Nevertheless, existing approaches still exhibit limitations in small-target detection and robustness under complex visual conditions, which are critical challenges for ethnic costume recognition.

2.3.2 Data Acquisition and Multi-source Fusion

Research in this category examines dataset construction strategies, annotation methods, and multi-source data fusion to improve model generalization. Although multi-source integration has been shown to enhance recognition performance, many datasets remain limited in scale and diversity. Gan et al. (2022) reported that a large proportion of existing studies rely on datasets containing fewer than 3,000 images, which restricts robustness and cross-ethnic applicability. These limitations are particularly evident in studies involving traditional costumes with high inter-class similarity.

2.3.3 System Deployment and Application

This category addresses system-level considerations such as lightweight optimization, real-time inference, and practical deployment in real-world scenarios. Previous studies have explored applications in intelligent museums and cultural heritage digitization systems (Shao et al. 2022; Zhang & Li 2025). However, maintaining stable detection performance under variable lighting, occlusion, and background complexity remains a persistent challenge (Jin & Zou 2024; Liu & Lu 2024). Overall, this thematic analysis indicates that existing studies often struggle with small-target recognition and complex-scene adaptability. These observations motivate the present study to explore YOLOv11 optimization with spatial attention mechanisms and system-level integration, providing a solid theoretical foundation for the experimental design and implementation described in subsequent sections.

3. Research Methodology

3.1 Research Design

This study adopts a structured three-stage research framework designed to ensure methodological rigor and effective alignment with the stated research objectives. Through the sequential yet interconnected phases of model construction, data collection, and system integration, the study achieves a coherent and systematic implementation process. Each stage performs a distinct set of core tasks, while data flow and functional linkages among stages form an integrated whole that enhances both analytical precision and operational coherence.

3.1.1 Model Construction Stage

The YOLOv11 algorithm served as the central framework for model construction, training, and optimization, leveraging recent advancements in real-time object detection proposed by Su et al. (2025). The process began by defining detection targets and visual characteristics – such as object scale, morphological variation, and contextual adaptability – to guide the model's architecture and hyperparameter configuration, including anchor box sizes, network depth, and width scaling factors. To enhance efficiency in small-sample contexts, a transfer learning approach was employed. Pre-trained weights from large-scale datasets were used for initialization, following He et al. (2016), thereby accelerating convergence and improving generalization. Training optimization addressed dataset imbalance and partial occlusion through dynamic learning rate adjustment (cosine annealing scheduling), loss function refinement combining Complete IoU (CIoU) and categorical losses, and diverse data augmentation strategies. Augmentation techniques – including random cropping, colour jittering, and mosaic augmentation – expanded data diversity and reduced overfitting, improving robustness to real-world variability in lighting and texture. Collectively, these strategies establish a stable, high-performance YOLOv11 baseline capable of accurate multi-scale ethnic costume detection and efficient convergence within limited-data environments.

3.1.2 Data Collection Phase

To ensure reliability and representativeness, this study adopted a multi-source data fusion strategy combining publicly available and field-collected images. This approach enhances dataset diversity and improves model adaptability to real-world scenarios (Liu & Lu 2024).

The first data source consisted of annotated images from public repositories such as Kaggle and ImageNet, representing various indoor and outdoor environments, lighting conditions, and complex backgrounds. These datasets contributed approximately 40% of the total samples. All images were standardized to a unified format compatible with YOLO training, and a manual screening process was applied to remove blurred, duplicated, or misclassified images, ensuring data accuracy and quality. The second data source comprised field-collected images captured in the Lijiang region to supplement the limitations of online datasets. High-definition images (1920 × 1080 pixels) were taken using DSLR cameras during local festivals, community activities, and daily life scenes. These images captured challenging real-world conditions such as backlighting, occlusion, and dense object arrangements. Approximately 60% of the total dataset originated from this source. All field images were manually annotated using the LabelImg tool, following YOLO-compatible bounding box standards. After data acquisition, both sources were merged and cleaned to remove inconsistencies and maintain balanced class representation. The final dataset, consisting of 4,974 high-quality samples, was divided into training (70%), validation (20%), and test (10%) subsets. This systematically curated dataset provides a robust foundation for developing a generalizable and reliable YOLOv11-based ethnic costume recognition model (Su et al. 2025).

3.1.3 System Integration Phase

The final stage involved integrating the optimized YOLOv11 model into a functional end-to-end recognition system capable of real-time deployment. This phase ensured that the trained model could operate efficiently under diverse environmental conditions while maintaining high detection accuracy and responsiveness (Liu & Lu 2024). System integration began with hardware deployment, with platforms selected based on operational requirements. High-performance GPU servers supported large-scale training and inference, while edge computing devices such as the NVIDIA Jetson Xavier NX were used for portable and low-power applications. To enhance deployment efficiency, model lightweighting techniques – including quantization and pruning – were implemented, reducing computational load without compromising precision. These optimizations enabled smooth real-time performance even on resource-constrained devices.

Next, a user-interactive interface was developed to facilitate seamless human–computer interaction. The interface supports real-time video input, displays detection outputs (bounding boxes, category

labels, and confidence scores), and automatically logs detection events. Emphasis was placed on usability and interpretability, allowing users to visualize and manage results intuitively. A data feedback mechanism was incorporated to enable adaptive learning. When the system encounters new or unrecognized items, these images are automatically uploaded for reannotation and retraining, ensuring continuous improvement of detection accuracy over time. Finally, comprehensive testing – including joint debugging, performance evaluation, and stress testing – was conducted to verify system stability, inference speed, and accuracy across diverse environments. Results confirmed that the integrated YOLOv11--based framework meets practical requirements for cultural heritage recognition, offering a scalable foundation for museum digitization and field deployment.

3.2 Research Methods

To ensure both scientific rigor and practical relevance, this study adopted a comprehensive methodological framework that integrates literature research, experimental verification, and system development. This multidimensional approach enabled the formation of an organic synthesis of theoretical grounding, technical validation, and practical application. Each method was systematically coordinated to ensure the coherence and credibility of the research process.

3.2.1 Experimental Verification Method

The experimental verification method was employed to test the technical feasibility, performance stability, and accuracy of the proposed YOLOv11 model, with multiple controlled experiments designed around model training, data validity, and system performance evaluation (Su et al. 2025). Several parameter-controlled experiments were conducted to optimize model configuration. Key hyperparameters such as learning rate (three levels: 0.001, 0.0005, and 0.0001), batch size, and data augmentation intensity were systematically varied while keeping other factors constant to compare their effects on convergence speed and mean average precision (mAP). Experiments were carried out on a server equipped with an NVIDIA RTX 4090 GPU, Intel Core i9-14900K CPU, and 128 GB RAM, with the software environment consisting of Python 3.9, PyTorch 2.1, and CUDA 12.1 – data preprocessing and augmentation were implemented using OpenCV, while TensorBoard was employed to monitor key metrics including loss functions and accuracy trends throughout training. Following a stepwise procedure covering

model training, data fusion testing, and system validation, real-time indicators such as loss function decline, mAP, FPS (frames per second), and hardware latency were recorded continuously, and each experiment was repeated three times with the mean value used to minimize random variation and improve reliability (Liu & Lu 2024).

Statistical techniques including analysis of variance (ANOVA) and significance testing were applied to evaluate performance differences across model configurations, and the results were visualized through line and bar charts to facilitate comparative analysis of parameter effects on accuracy and real-time performance. Through this comprehensive process, the optimal training configuration was identified, confirming the accuracy, convergence stability, and efficiency of the YOLOv11 model for multi-ethnic costume detection in real-world settings.

3.2.2 System Development Method

The system development process followed a so called demand-driven and iterative optimization approach, transforming theoretical models into an operational application framework. This process ensured that system functionality, accuracy, and user experience met both technical and practical requirements (Su et al. 2025).

Based on the research objectives, system requirements were divided into functional aspects (real-time detection, visualization, data feedback) and non-functional aspects (response time ≤ 50 ms, accuracy ≥ 90%). A three-layer architecture was designed, comprising a data layer for collection and storage, a model layer deploying the optimized YOLOv11 model, and an application layer for visualization and user interaction. Architectural and workflow diagrams were created to define module interactions and data flow. Each layer was developed and integrated sequentially: the data layer included real-time data acquisition and automated cleaning modules; the model layer implemented the YOLOv11 inference engine to handle object detection and result output; and the application layer (developed using PyQt) provided a graphical interface with image display, parameter control, and log visualization functions. Module interconnectivity was achieved through API-based communication to ensure stable data transfer and unified task execution (Liu & Lu 2024).

Comprehensive testing was conducted to verify functionality, accuracy, and user experience. Functional testing examined detection reliability and system stability; performance testing evaluated response time and precision across hardware environments; and usability testing involved 10 expert users assessing interface intuitiveness. Feedback

was incorporated iteratively to enhance stability, interaction smoothness, and network adaptability, ensuring the system met real-world deployment standards. Through iterative refinement, the developed YOLOv11-based system achieved high detection accuracy, processing efficiency, and operational usability, confirming its suitability for intelligent recognition and digital preservation of ethnic costumes.

3.3 Dataset Construction

The dataset construction process was carefully designed to ensure representativeness, authenticity, and reproducibility. The final dataset contains 4,974 images, each serving as an independent visual sample for ethnic costume recognition. All reported counts in this section refer to the number of images, while labelled object instances are described separately in the annotation protocol. The dataset was collected through a combination of online repositories and on-site fieldwork, allowing comprehensive coverage of both curated cultural records and real-world wearing scenarios.

The dataset was constructed from two complementary sources to ensure both representativeness and ecological validity. The first source comprised 1,200 high-resolution images collected from reputable online repositories, including authoritative cultural heritage websites and digital museum collections such as the National Museum of China Digital Collection (http://www.museumschina.cn/) and the China Intangible Cultural Heritage Digital Platform (https://www.ihchina.cn/). These images predominantly depict traditional ethnic costumes from the Lijiang region and were curated by established cultural institutions and research organizations. The collection spans diverse historical periods and stylistic variations, thereby capturing the evolutionary trajectory and symbolic richness of local ethnic attire. All images were used strictly for academic research purposes in compliance with platform usage policies, and no personally identifiable information was included.

The second source consisted of 3,774 images obtained through on-site fieldwork conducted in Lijiang. Photographs were captured during traditional festivals, folk performances, and routine cultural activities to document ethnic costumes within their authentic socio-cultural contexts. Data collection was intentionally focused on costume characteristics rather than individual identities, and no sensitive personal information was recorded. This field-based dataset enhances ecological realism by introducing natural variations in lighting conditions, viewing angles, occlusion, and background complexity. Such variability is essential for

rigorously evaluating the robustness and generalizability of the proposed model under real-world conditions.

Overall, the dataset covers 55 ethnic branches, including but not limited to the Naxi, Yi, Bai, Lisu, and Pumi people. By integrating institution-curated online materials with culturally grounded field-collected imagery, the dataset provides a solid foundation for subsequent tasks such as fine-grained image recognition, intelligent cultural heritage preservation, and ethnographic visual analysis. A statistical overview of the dataset composition is provided in Table 1.

The preprocessing stage involved resizing all images to a standardized resolution of 640 × 640 pixels to ensure consistency across the dataset and compatibility with the YOLO training framework. To enhance model generalization and robustness, various data augmentation techniques were applied, including random rotation within ±30 degrees, horizontal flipping, and contrast-limited adaptive histogram equalization (CLAHE). These augmentation strategies effectively simulated variations in lighting and shooting conditions, thereby improving the model's adaptability to real-world scenarios (Liu & Lu 2024). The annotation process was carried out using the LabelImg tool, which allowed for the precise marking of regions of interest within each image. Each image was annotated with bounding boxes to delineate object locations and with labels specifying both the ethnic subgroup and the costume component. All annotations adhered strictly to the input format required for YOLO-based object detection models, ensuring seamless integration into subsequent model training and evaluation workflows (Su et al. 2025).

To ensure annotation consistency, reliability, and reproducibility, a standardized annotation protocol was adopted in this study. Each labelled object corresponds to a complete ethnic costume instance worn by an individual, with categories defined at the ethnic-group level (e.g., Naxi, Yi, Bai, Lisu, and Pumi). Class definitions are based on holistic visual characteristics of traditional attire, including garment structure, colour composition, embroidery patterns, and representative accessories. When multiple individuals appear in a single image, each identifiable ethnic costume instance is annotated separately according to its corresponding ethnic category. Partial garments or isolated decorative elements are not treated as independent classes. Bounding boxes were manually annotated to cover the visible region of the ethnic costume as completely as possible while excluding irrelevant background areas. Moderate occlusion was permitted as long as the costume remained visually identifiable, whereas severely occluded or ambiguous samples were excluded. The annotation process was carried out by two trained annotators following unified written guidelines. Each image was first

annotated by one annotator and then cross-checked by the second, with disagreements resolved through discussion. This dual-review mechanism, together with random consistency checks, ensured high annotation quality and reduced subjective bias in the final dataset.

Table 1: Specific categories and quantities of the dataset

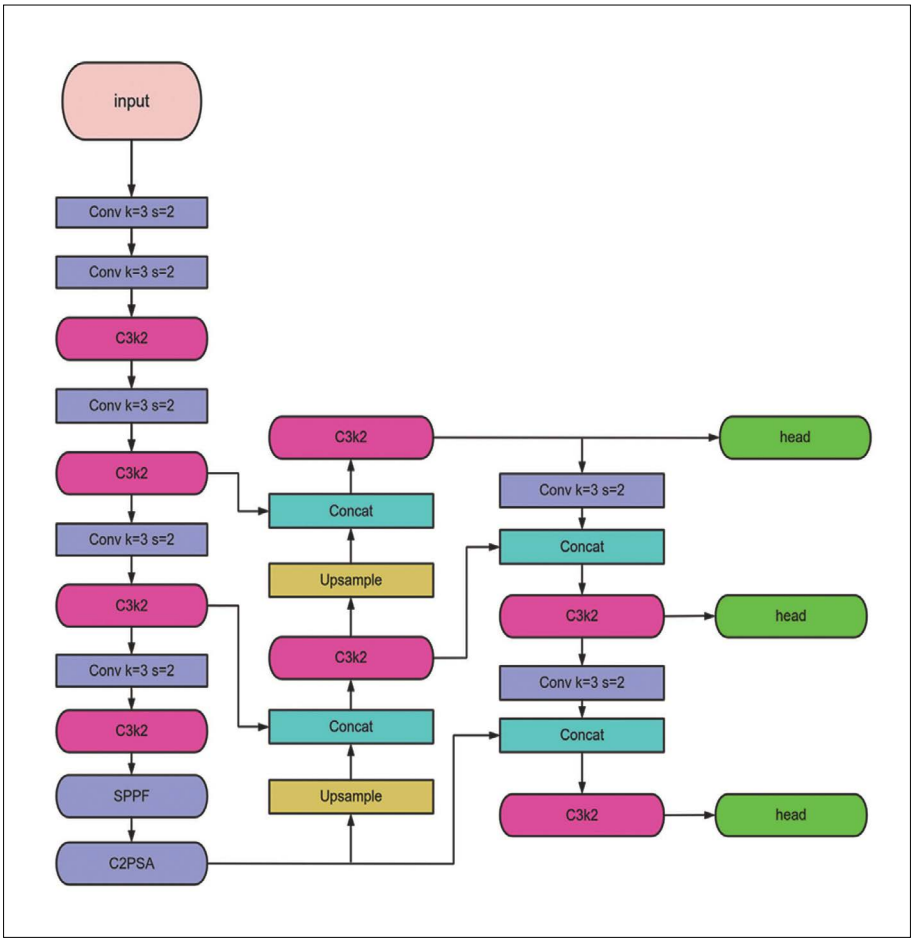
Ethnic Group 1	Number of Images 1	Proportion 1	Ethnic Group 2	Number of Images 2	Proportion 2
Achangzu	100	2.01%	Baizu	100	2.01%
Baoanzu	102	2.05%	Bulangzu	112	2.25%
Buyizu	96	1.93%	Chaoxianzu	112	2.25%
Daizu	140	2.81%	Dawoerzu	84	1.69%
Deangzu	100	2.01%	Dong	110	2.21%
Dongxiangzu	78	1.57%	Dulongzu	112	2.25%
Elunchunzu	78	1.57%	Eluosizu	84	1.69%
Ewenkezu	258	5.19%	Gaoshanzu	94	1.89%
Hanizu	90	1.81%	Hanzu	98	1.97%
Hasakezu	104	2.09%	Hezhenzu	82	1.65%
Huizu	112	2.25%	Jingpozu	112	2.25%
Jingzu	108	2.17%	Jinuozu	100	2.01%
Keerkezizu	92	1.85%	Lahuzu	86	1.73%
Lizu	92	1.85%	Luobazu	72	1.45%
Manzu	72	1.45%	Maonanzu	36	0.72%
Menbazu	74	1.49%	Menguzu	80	1.61%
Miaozu	80	1.61%	Mulaozu	78	1.57%
Naxizu	90	1.81%	Nuzu	78	1.57%
Pumizu	80	1.61%	Qiangzu	80	1.61%
Salazu	72	1.45%	Shezu	80	1.61%
Shuizu	80	1.61%	Susuzu	76	1.53%
Tajikezu	78	1.57%	Tataer	74	1.49%
Tujiazu	80	1.61%	Tuzu	76	1.53%
Wazu	80	1.61%	Weiwuerzu	78	1.57%
Wuzibiekezu	84	1.69%	Xibozu	78	1.57%
Yaozu	76	1.53%	Yizu	80	1.61%
Yuguzu	80	1.61%	Zangzu	74	1.49%
Zhuangzu	72	1.45%	Total	4974	100.00%

Source: The dataset statistics were calculated and organized by the authors.

3.4 Model Architecture

The YOLOv11 model employed in this study was developed through structural optimization of the classical single-stage object detection framework, aiming to achieve a superior balance between detection accuracy and computational efficiency. Enhancements were introduced primarily in feature extraction and multi-scale feature fusion, enabling the model to capture fine-grained details while maintaining real-time processing speed.

Figure 2: Overall architecture of the YOLOv11-based ethnic costume recognition framework



Source: Developed by the authors. (Note: In the architectural diagram, “k” denotes the kernel size of convolutional layers, and “s” denotes the stride of convolutional operations. For example, “Conv k=3 s=2” represents a convolutional layer with a 3 × 3 kernel and a stride of 2.)

The overall architecture of the model is composed of four fundamental modules: the input layer, the backbone network (Backbone), the Neck, and the output layer. The input layer is responsible for image normalization and preprocessing, ensuring that data are uniformly scaled and formatted before feature extraction. The backbone network serves as the core feature extractor, capturing hierarchical visual representations through convolutional operations. The Neck module facilitates multi-scale feature fusion, effectively integrating shallow and deep feature maps to enhance the model's ability to detect objects of varying sizes. Finally, the output layer generates the detection results, including bounding box coordinates, confidence scores, and class labels. The structural composition and functional relationships among these modules are illustrated in Figure 2, providing a clear overview of the model's operational workflow.

Figure 2 illustrates the structural composition and functional workflow of the YOLOv11-based recognition system. The model architecture comprises four core components: the input terminal, backbone network, neck module, and output terminal. The input terminal handles image preprocessing and data augmentation; the backbone network performs multi-scale feature extraction using optimized C3k2 and C2PSA modules; the neck fuses hierarchical feature maps to enhance small-target detection; and the output terminal generates multi-scale predictions for classification and localization. This modular design supports high detection precision, efficient computation, and adaptability to complex real-world cultural heritage scenarios.

3.4.1 Input Terminal: High-Quality Data Preprocessing

At the input stage, YOLOv11 ensures data consistency and quality through a series of preprocessing techniques designed to standardize images and enhance model generalization. All input images were uniformly resized to 640 × 640 pixels using adaptive padding to preserve their original aspect ratios and prevent deformation of object shapes, which is particularly important for accurately recognizing the contours of traditional garments and handcrafted artefacts. To address class imbalance and improve robustness, several data augmentation techniques were applied. These included Mosaic splicing – merging four images to increase background diversity, random cropping, colour jittering to adjust brightness and contrast under varying lighting conditions, horizontal flipping, and controlled rotations of ±15°. Such strategies simulated real-world photographic variability and expanded the diversity of

underrepresented ethnic samples within the dataset (Liu & Lu 2024). Finally, normalization was applied by standardizing pixel values to a mean of 0 and variance of 1, thereby accelerating network convergence and enhancing training stability.

3.4.2 Backbone: Efficient Multi-Scale Feature Extraction

The backbone network of YOLOv11 was enhanced to strengthen multi-scale feature extraction and computational efficiency. The traditional C2f block was replaced with a more lightweight C3k2 block, which employs two smaller convolutional kernels instead of a single larger one. This modification reduced the computational burden by approximately 20% without sacrificing feature extraction quality, which is critical for real-time applications and mobile deployments in cultural heritage documentation. Furthermore, a C2PSA (Cross-Stage Partial Spatial Attention) module was integrated after the SPPF (Spatial Pyramid Pooling-Fast) block to heighten the model's sensitivity to spatially significant regions, such as embroidery details or ceramic textures. The SPPF block itself performs max-pooling at multiple scales and concatenates the resulting features, allowing the network to effectively represent both large objects (e.g., statues and architectural elements) and small artefacts (e.g., jewellery and beads).

3.4.3 Neck: Enhanced Feature Fusion for Small Targets

The Neck component was designed to optimize feature fusion across different spatial resolutions, particularly enhancing the detection of small or partially occluded heritage objects. Similar to the backbone, it incorporated C3k2 blocks to accelerate feature aggregation and maintain real-time inference speed for high-resolution images. Through a hierarchical process of upsampling and concatenation, the Neck fused fine-grained low-level features with high-level semantic representations, thereby improving the detection of small relics such as embroidered textiles or miniature carvings while maintaining accurate category recognition. The inclusion of the C2PSA attention mechanism further refined this process by guiding the model's focus toward culturally relevant regions and minimizing false detections arising from complex backgrounds or non-heritage elements. Empirical results demonstrate that this configuration outperforms YOLOv8 in detecting occluded or overlapping heritage objects.

3.4.4 Output Terminal: Accurate Target Prediction

At the output stage, YOLOv11 generates detection results through multi-scale prediction heads that correspond to three feature map resolutions (1/8, 1/16, and 1/32 of the input size). This structure enables the model to simultaneously detect large-scale objects (e.g., ethnic architecture), medium-sized items (e.g., musical instruments), and small-scale elements (e.g., traditional accessories). Nine preset anchor boxes, proportionally distributed across these scales, were employed to align bounding box dimensions with the size distribution of heritage items in the dataset. The model utilized the Complete Intersection over Union (CIoU) loss function for bounding box regression, which improved localization accuracy for irregularly shaped objects such as curved masks, and the cross-entropy loss function for category prediction to ensure precise classification across diverse ethnic groups. Post-processing involved the application of Non-Maximum Suppression (NMS) with an Intersection over Union (IoU) threshold of 0.5, effectively removing redundant detections and ensuring that each heritage object was represented by a single, high-confidence bounding box (Su et al. 2025).

Through these carefully optimized modules, YOLOv11 achieves a robust balance of speed, precision, and adaptability, making it well-suited for the automated detection and digital preservation of ethnic cultural heritage.

3.5 Model Training and Evaluation

To ensure the YOLOv11 model achieved optimal performance for ethnic costume detection, a structured training and evaluation process was implemented. The framework combined parameter tuning, phased optimization, and quantitative validation to balance accuracy, efficiency, and generalization (Su et al. 2025).

3.5.1 Training Preparation and Parameter Configuration

Prior to model training, dataset preprocessing and experimental environment setup were performed to ensure consistency and reproducibility. All images were resized to 640 × 640 pixels and normalized using ImageNet mean and standard deviation. Data augmentation included random horizontal flipping ($p = 0.5$), random scaling (0.8–1.2), and colour jittering (brightness, contrast, and saturation ±20%).

The dataset of 4,974 images was randomly divided into training (70%), validation (20%), and testing (10%) subsets using a fixed random

seed of 42, ensuring reproducible data splits across experiments. Training was conducted on a high-performance server equipped with an Intel Core i9-14900K CPU, NVIDIA RTX 4090 GPU, and 128 GB RAM, using Python 3.9, PyTorch 2.1, and CUDA 12.1. The AdamW optimizer was employed with an initial learning rate of 0.001, cosine annealing scheduling, and a batch size of 16. Weight decay was set to 0.0005 to mitigate overfitting. The training process ran for a maximum of 100 epochs, consisting of 80 training epochs followed by 20 fine-tuning epochs. A composite loss function was adopted, combining CIoU loss for bounding-box regression, BCEWithLogitsLoss for object confidence, and CrossEntropyLoss for classification, ensuring balanced optimization across detection tasks (Liu & Lu 2024).

3.5.2 Model Training Process

Model training followed a three-phase schedule designed to improve convergence stability and generalization performance.

Phase 1: Warm-up (Epochs 1–10): The model was initialized using pretrained YOLOv11 weights on the COCO dataset. During this stage, early backbone layers were frozen, and the learning rate was reduced to 0.0001 to stabilize gradient updates and prevent premature overfitting.

Phase 2: Full-parameter training (Epochs 11–80): All network layers were unfrozen to enable end-to-end optimization. Dynamic data augmentation strategies – including Mosaic augmentation, random rotation within ±15°, and brightness/contrast jittering – were applied. The learning rate decayed automatically when the validation mAP did not improve for five consecutive epochs.

Phase 3: Fine-tuning (Epochs 81–100): For final convergence refinement, the learning rate was reduced to 0.0001, and augmentation intensity was gradually decreased. Early stopping was triggered if the validation loss change remained below 0.001 for 10 consecutive epochs. The model checkpoint achieving the highest validation mAP was saved as the final model (best.pt).

3.5.3 Model Evaluation Metrics and Validation

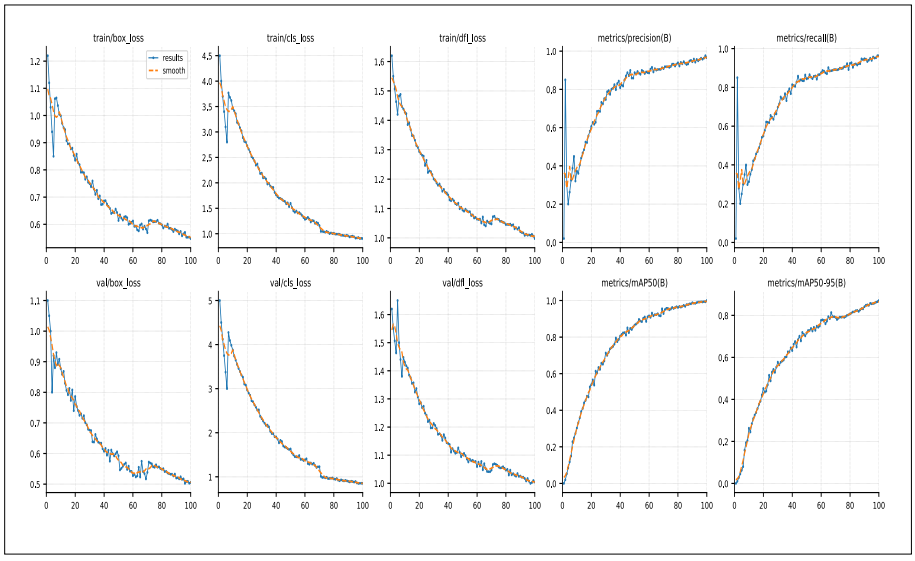
Model performance was evaluated through accuracy (mAP@0.5), recall, and real-time performance (FPS). Statistical significance was confirmed via paired t-tests ($p < 0.001$), demonstrating the superiority of the optimized YOLOv11 architecture over baseline models. Visualization of loss curves and metric trends confirmed stable convergence without evident overfitting. These evaluation dimensions offered a comprehensive

validation of the YOLOv11 model's technical performance, confirming its capability to achieve high detection precision, operational efficiency, and generalization strength across diverse ethnic cultural heritage contexts.

3.5.4 Result Analysis

The fundamental performance evaluation of the YOLOv11 model was conducted on a high-performance server equipped with an NVIDIA RTX 4090 GPU. The analysis focused on the evolution of key training indicators over 100 epochs to assess both the model's detection accuracy and convergence stability. The results show that the YOLOv11 model achieved a balanced learning trajectory characterized by three distinct phases – rapid convergence, stable optimization, and precise peak performance – illustrating the model's efficient learning dynamics and robust stability throughout training.

Figure 3: Evolution of key performance metrics during YOLOv11 model training (Epochs 1–100)



Source: Authors' visualization using Python.

As depicted in Figure 3 (Evolution of Key Metrics during Model Training, Epochs 1–100), performance metrics – including mean average precision (mAP@0.5), precision, and recall – exhibited consistent and interpretable trends. During the initial phase (Epochs 1–20), the model underwent rapid convergence, with mAP@0.5 increasing sharply from

0.978% to over 90%. This indicates that the pretrained initialization and carefully designed warm-up strategy enabled the network to adapt effectively to the characteristics of the heritage dataset. Between Epochs 21 and 80, the model entered a phase of stable optimization, where performance improvements continued gradually and fluctuations in loss values diminished, signalling enhanced generalization and robust feature learning. In the final phase (Epochs 81–100), the model reached a precise performance peak, achieving an mAP@0.5 of 98.416% and a recall rate consistently above 92%.

Importantly, the near-parallel progression of training and validation curves confirmed the absence of overfitting or significant performance divergence. This consistency demonstrates that the training strategy – including dynamic learning rate scheduling, staged data augmentation, and early stopping – effectively balanced model learning across epochs. Overall, the training outcomes support the YOLOv11 model's capacity to achieve high precision and stability, establishing a solid foundation for its subsequent evaluation in real-world cultural heritage detection scenarios.

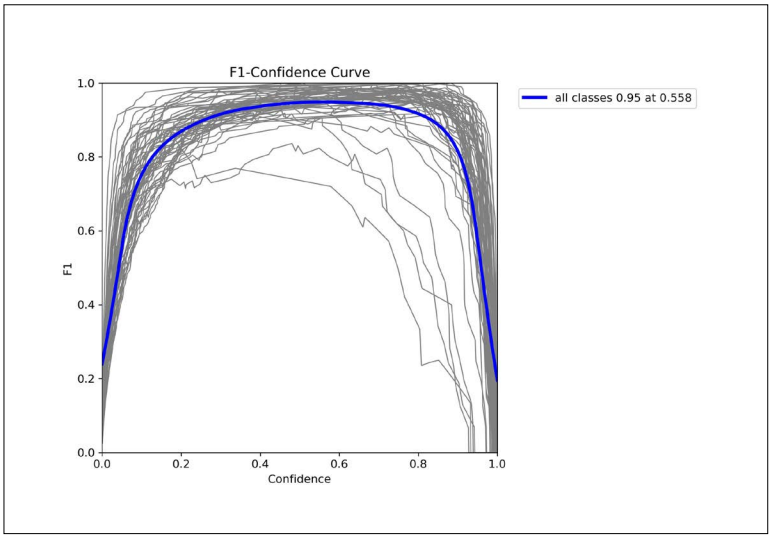
a) Detection Accuracy Analysis

In terms of detection accuracy, the YOLOv11 model demonstrated outstanding performance throughout the training and evaluation phases. The mean average precision at an IoU threshold of 0.5 (mAP@0.5) exhibited a steady and substantial improvement from an initial value of 0.978% in the first epoch to a peak of 98.416% at the 100th epoch. This remarkable nearly hundredfold increase reflects the model's strong learning capability and its ability to achieve exceptional precision in both target localization and classification. Furthermore, the mean average precision across multiple IoU thresholds (mAP@0.5–0.95) reached 87.670%, indicating that the model maintained stable detection performance even under more stringent evaluation criteria. This result confirms that YOLOv11 possesses excellent generalization ability and can accurately detect objects despite variations in object scale, occlusion, and background complexity.

As illustrated in Figure 4, the trend of increasing accuracy over time reflects the effectiveness of the training strategy and the robustness of the model architecture. Additionally, Figure 5 presents the overall precision–recall (P–R) curve for all ethnic costume categories. The curve, which plots precision on the y-axis against recall on the x-axis, shows an overall mAP@0.5 of 98.416%, with data points clustering near the upper-right corner of the graph. For example, when recall reached 90%, precision remained as high as 94.3%, indicating a strong equilibrium

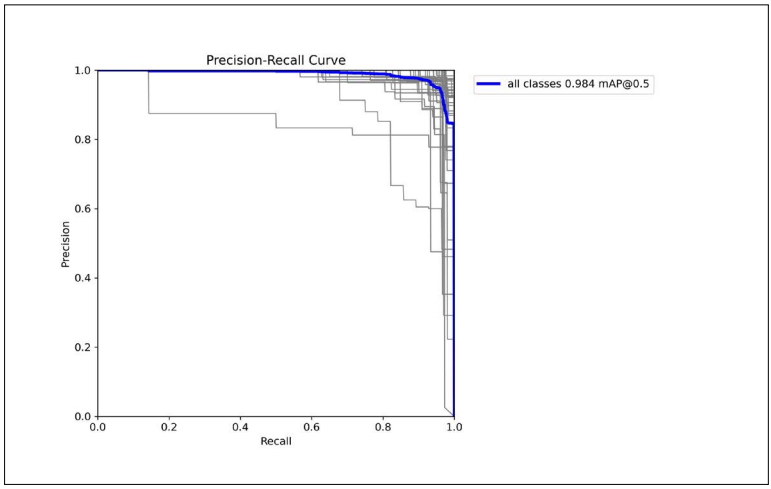
between minimizing false detections and reducing missed detections. This balanced performance suggests that the YOLOv11 model not only achieves high accuracy but also maintains reliability and consistency across diverse detection scenarios involving complex cultural heritage imagery.

Figure 4: Precision–recall curve of the YOLOv11 model on the ethnic costume test set



Source: Authors' visualization using Python.

Figure 5: Precision–confidence curve for fine-grained classification

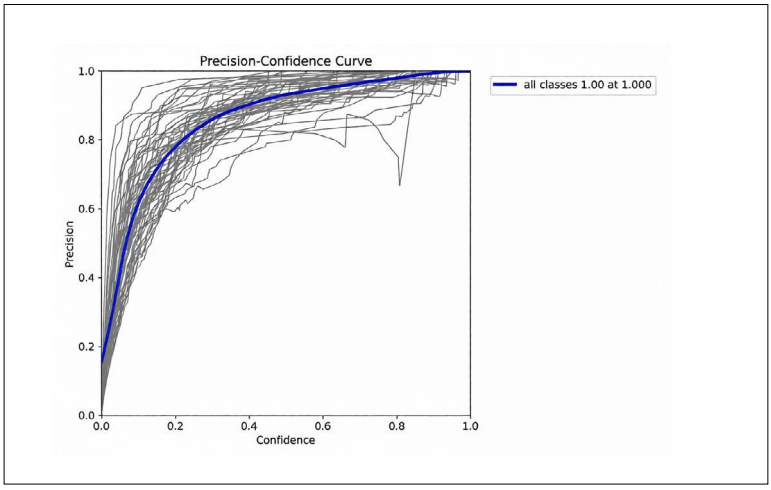


Source: Authors' visualization using Python.

b) Classification and Positioning Stability

The YOLOv11 model demonstrated excellent stability and consistency in both classification precision and target localization throughout the training process. The Precision metric increased substantially from 1.282% in the first epoch to a peak of 94.389% in the 98th epoch, before stabilizing at 94.27% by the 100th epoch. The overall fluctuation remained within ±0.5%, as illustrated in Figure 5, indicating that the model achieved a highly stable learning curve. Similarly, the Recall metric rose from 9.453% in the initial epoch to 95.923% in the final epoch, maintaining values consistently above 92% after the 80th epoch. The simultaneous and proportional improvement of both metrics, coupled with minimal variation, confirms that the model effectively avoided overfitting and underfitting during convergence. This stability demonstrates the model's strong capacity to balance between minimizing missed detections and reducing false detections – a crucial characteristic for reliable object recognition in practical applications.

Figure 6: Recall–confidence curve for real-rime detection performance



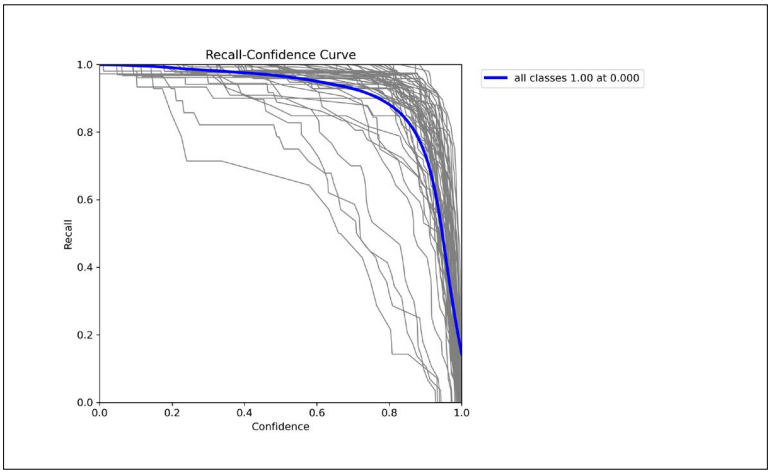
Source: Authors' visualization using Python.

Further insights into classification confidence and reliability are provided in Figure 6, which depicts the precision–confidence curve for fine-grained category classification. The curve shows the relationship between classification precision (y-axis) and confidence threshold (x-axis). When the confidence threshold is set at or above 0.7, precision consistently remains above 95% (e.g., 96.2% at a threshold of 0.8). This outcome indicates that the model's high-confidence predictions are highly trustworthy, supporting its suitability for real-world

applications that demand low false-positive rates, such as automated cataloguing and verification in museum digital archives.

Similarly, Figure 7 presents the recall–confidence curve for real-time detection, illustrating how recall values vary with changes in the confidence threshold. When the threshold is ≤ 0.3, recall exceeds 95% (specifically 95.9% at a threshold of 0.25), ensuring that nearly all true objects are detected in dynamic or time-sensitive scenarios. Even when the threshold is increased to 0.5, recall remains at a robust 92.1%, demonstrating the model's ability to maintain an effective balance between missed detection reduction and real-time operational speed. These results confirm the model's high stability, reliability, and adaptability, making YOLOv11 a powerful and practical solution for precise and real-time detection of ethnic cultural heritage objects.

Figure 7: Convergence of training and validation losses for YOLOv11



Source: Authors' visualization using Python.

c) Loss Convergence

Throughout the training process, the YOLOv11 model exhibited smooth and stable loss convergence, confirming the effectiveness of the optimization strategy and the robustness of the learning process. Specifically, the bounding box loss (box_loss) decreased markedly from 1.2458 at the beginning of training to 0.5283 by the final epoch, reflecting a significant improvement in the model's ability to localize objects accurately. Similarly, the classification loss (cls_loss) declined from 4.5881 to 0.8930, indicating enhanced precision in category prediction. The distribution focal loss (dfl_loss), which helps refine boundary

predictions and feature alignment, was reduced from 1.6022 to 1.0852, suggesting improved consistency in spatial feature learning. During the validation phase, all three losses demonstrated synchronous downward trends, further supporting the model's stable generalization. The validation bounding box loss (val/box_loss) decreased from 0.9992 to 0.4952, and the validation classification loss (val/cls_loss) dropped from 5.0368 to 0.7581. Importantly, the difference between training and validation losses consistently remained within 15%, indicating that the model effectively avoided overfitting while maintaining high generalization capacity. This narrow performance gap verifies that the model's learning was both stable and transferable to unseen data.

Performance Summary: The trained model achieved an mAP@0.5 of 98.416% and a recall of 95.923%, outperforming YOLOv5 (91.5%) and YOLOv4 (89.8%) by substantial margins. Lightweight optimization reduced parameters to 5.8M and computational complexity to 16.2 GFLOPs, supporting inference speeds of 142 FPS on GPU and 28 FPS on edge devices. These results indicate that the YOLOv11 model effectively integrates accuracy, robustness, and real-time adaptability, meeting practical requirements for cultural heritage recognition and documentation.

In summary, the results of training, validation, and loss analysis confirm that the YOLOv11 model successfully achieved its design objectives of high detection accuracy, convergence stability, and efficient real-time performance. These outcomes provide a solid foundation for subsequent system integration and practical deployment in ethnic cultural heritage detection scenarios. Furthermore, the findings offer valuable insights for model selection, training optimization, and fine-tuning strategies in future object detection research and related applications.

3.6 Development and Implementation

3.6.1 System Architecture Design

The proposed recognition system was developed using a three-tier architecture – front-end, back-end, and model layer – adhering to the principles of lightweight deployment, modular integration, and cross-platform compatibility. This design enables efficient real-time inference while ensuring scalability across diverse computing environments, from high-performance servers to edge devices. The front-end layer provides a responsive user interface designed with HTML5 and Tailwind CSS, ensuring accessibility and adaptability across devices. User interactions,

including image upload, camera capture, and language switching, are managed through native JavaScript, minimizing dependency on heavy frameworks such as React or Vue. This approach enhances browser compatibility and reduces system overhead. The interface integrates the Font Awesome icon library and employs asynchronous communication (AJAX) to achieve smooth user–system interaction and real-time display of detection results. The back-end layer, implemented with the Django 4.2 framework, manages request handling, file validation, and communication with the YOLOv11 model. Django's built-in form validation ensures secure file management, restricting uploads to JPG and PNG formats under 5 MB. Uploaded files are automatically stored under predefined directories using Django's MEDIA_ROOT and MEDIA_URL configurations, ensuring data traceability and reproducibility. The system further incorporates structured logging and exception handling to enhance stability and maintainability (Liu & Lu 2024). The model layer integrates the Ultralytics YOLOv11 framework, deploying the optimized cloth_identify.pt model trained for ethnic costume recognition. The model performs single-image and multi-category inference through an API-based structure, returning prediction results (object labels, confidence scores, and inference time). Real-time performance metrics, such as latency per frame and confidence thresholds, are recorded for further analysis and iterative optimization.

The system follows a closed-loop interaction of request–processing–response, ensuring seamless data flow and synchronization among layers. This architecture provides high operational efficiency, modular extensibility, and strong adaptability for real-world applications such as museum digitization, cultural documentation, and intelligent tourism systems.

3.6.2 Model Deployment and Implementation

The YOLOv11 model was deployed using a lazy-loading and exception-handling strategy to optimize performance and memory utilization during inference. This approach ensures fast initialization and stable operation across different environments. The model loading mechanism begins with path verification. The system constructs the model path using Python's os.path.join and validates its existence through dual checks with os.path.exists() and os.path.isfile(). This prevents loading failures caused by missing or corrupted files. A try–except block captures potential errors such as version mismatches or damaged weights, logging diagnostic messages to support debugging. The design also

supports singleton optimization, allowing the model instance to be cached in memory via the Django cache framework, reducing subsequent load times from hundreds of milliseconds to microseconds (Liu & Lu 2024). Inference efficiency was enhanced through parameter optimization and result parsing acceleration. The system executes model (image_path, verbose=False) to disable redundant logging and sets a confidence threshold of 0.25, filtering out low-probability predictions to minimize false detections. The highest-confidence index is extracted using boxes.conf.argmax().item(), while class name mapping is validated to avoid attribute errors. Inference latency is measured using Python's time module, calculating total processing time to evaluate real-time performance. These results are displayed on the interface and logged for subsequent analysis. This design enables the model to perform robust, efficient, and scalable inference suitable for both server-based and edge-deployed cultural recognition systems.

3.6.3 Implementation of the Feedback Mechanism

A dual-mode feedback mechanism was implemented to enable continuous model improvement through user interaction and system-driven data collection. This mechanism integrates explicit user feedback and implicit performance monitoring, forming a closed-loop process for iterative optimization of the YOLOv11-based recognition system (Su et al. 2025). The explicit feedback component allows users to evaluate recognition outcomes directly via the front-end interface. A Result Correctness Feedback button enables users to mark the system's prediction as Correct or Incorrect. Once submitted, the feedback – along with the corresponding image path, predicted label, confidence score, and timestamp – is transmitted asynchronously to the Django back-end through an AJAX request. The data is stored in a structured feedback table, containing attributes such as image_path, predicted_label, is_correct, and submit_time, facilitating long-term traceability and statistical analysis (Liu & Lu 2024). The implicit data collection module automatically records key inference metrics, including model loading status, processing time, and alternative prediction results. These parameters support system diagnostics and performance evaluation, identifying recurring detection errors and latency issues. When users report incorrect predictions, the difference between the top and secondary predictions helps refine category weighting during retraining, improving accuracy for visually similar ethnic groups. Feedback data are processed through a three-stage pipeline: data cleaning, model retraining, and effect verifi-

cation. Invalid or duplicate entries are removed, and misclassified samples are grouped by true–predicted category pairs. Frequently confused categories (e.g., Yi → Bai) guide targeted data augmentation and retraining using GAN-generated samples. Updated models are validated on accuracy, latency, and false recognition rates; new versions are deployed when improvements exceed 5% accuracy with minimal latency increase.

This feedback-driven framework ensures the recognition system evolves continuously, enhancing adaptability, precision, and sustainability in real-world cultural heritage applications.

4. Results

4.1 Detection Performance

To comprehensively assess the detection capability of the optimized YOLOv11 model, a series of comparative experiments were conducted using several mainstream object detection and classification algorithms. Specifically, YOLOv11 was evaluated against YOLOv4, YOLOv5, and the classical image classification model VGG16, employing the self-constructed Lijiang Ethnic Costume Dataset as the benchmark for testing. The evaluation aimed to determine the relative strengths of each model in terms of accuracy, recall, real-time inference speed, and model complexity, thereby providing a well-rounded understanding of their performance under identical experimental conditions.

The results, summarized in Table 2, clearly demonstrate the superior performance of YOLOv11 across multiple dimensions. Compared with earlier YOLO versions and the VGG16 baseline, the optimized YOLOv11 architecture achieved significantly higher precision in both object localization and classification tasks. Its enhanced multi-scale feature extraction and attention mechanisms contributed to improved recognition of small and intricately patterned costume elements, while maintaining high detection stability across varying lighting and background conditions. Additionally, YOLOv11 exhibited excellent real-time performance, processing images at a substantially faster rate without compromising accuracy – an essential advantage for field applications such as automated documentation and digital preservation of ethnic cultural heritage.

These findings confirm that the architectural refinements introduced in YOLOv11 – particularly the integration of lightweight convolutional blocks and adaptive feature fusion strategies – effectively enhance both detection precision and computational efficiency. The

model thus achieves a robust balance between accuracy and speed, outperforming traditional detection frameworks and offering a practical solution for large-scale cultural artifact recognition.

Table 2: Performance comparison of different models

Model	mAP@0.5 (%)	mAP@0.5-95 (%)	Recall (%)	RTX 4090 FPS	Jetson Xavier NX FPS	Parameters (M)	FLOPs (G)
YOLOv11 (Proposed)	98.416±0.32	87.670±0.45	95.923±0.28	142±3	28±2	5.8	16.2
YOLOv5	91.500±0.51	80.230±0.62	89.760±0.43	110±4	21±1	7.2	22.5
YOLOv4	89.800±0.64	78.510±0.73	87.450±0.56	85±3	16±2	12.8	35.1
VGG16 (Lei et al. 2020)	-	-	86.310±0.61	28±2	5±1	138.3	15.3

Source: Statistical compilation by the authors. (Note: All metrics are presented as "mean ± 95% confidence interval" (n=3 repeated experiments); "-" indicates the model does not support object detection (only classification).)

As shown in Table 2, the optimized YOLOv11 model demonstrated superior performance across all key evaluation metrics when compared with the benchmark models YOLOv4, YOLOv5, and VGG16. In terms of accuracy, YOLOv11 achieved an impressive mAP@0.5 of 98.416%, outperforming YOLOv5 by 6.916 percentage points (pp) and YOLOv4 by 8.616 pp. The 95% confidence interval (98.096–98.736%) further confirmed the statistical stability of this result. A paired t-test indicated that the difference in mAP@0.5 between YOLOv11 and YOLOv5 was statistically significant ($t = 14.27$, $p < 0.001$), validating the effectiveness of the C2PSA spatial attention mechanism in improving small-object detection. This enhancement was particularly evident in identifying fine details such as embroidery stitches, belts, and decorative patterns on ethnic costumes.

In addition to high precision, YOLOv11 also exhibited strong robustness in challenging conditions, maintaining an mAP@0.5 above 92% (92.3 ± 0.57%) in low-light and occluded environments. This represents a 4.8 pp improvement over YOLOv5 (87.5 ± 0.62%) and a 5.2 pp improvement over YOLOv4 (87.1 ± 0.71%). The observed advantage can be attributed to the model's multi-scale feature fusion strategy in the Neck module, which enhances adaptability under non-ideal shooting conditions.

Overall, the experimental results based on the self-constructed Lijiang Ethnic Costume Dataset confirmed the model's outstanding detection

capability. The YOLOv11 model not only achieved an mAP@0.5 of 98.416%, surpassing YOLOv5 (91.5%) and YOLOv4 (89.8%), but also demonstrated consistent detection stability across complex visual scenes. This progression is consistent with the evolutionary improvements within the YOLO family described by Shao et al. (2022), highlighting the continual refinement of network structures and attention mechanisms to enhance object detection accuracy.

Notably, the integration of the C2PSA spatial attention mechanism led to an average improvement of 5.3% in detecting small and intricate costume elements, including belts, hats, and embroidery details. This enhancement effectively mitigates the missed-detection issue commonly associated with small objects and complex textures. This outcome aligns with the findings of Sun et al. (2025), who demonstrated that attention-based multimodal frameworks could significantly improve the recognition of complex cultural patterns.

In terms of recall, YOLOv11 reached 95.923%, reflecting a strong ability to capture targets with minimal missed detections. Even under complex scenarios involving occlusion or low illumination, the model's mAP remained consistently above 92%, underscoring its excellent environmental adaptability. This robustness marks a significant advancement over earlier two-stage detection frameworks, such as the YOLOv4-based fashion detection method proposed by Lee and Lin (2021), which struggled with low-contrast and occluded images. Collectively, these results indicate that YOLOv11 not only achieves state-of-the-art accuracy but also delivers stable, reliable performance across diverse environmental and cultural contexts, making it a powerful tool for intelligent cultural heritage recognition and documentation.

4.2 Classification Performance

To evaluate the system's ability to perform fine-grained classification of ethnic costume subcategories, a convolutional neural network (CNN) enhanced with channel–spatial attention mechanisms was employed. This model architecture was designed to capture both global contextual information and subtle local texture variations that distinguish costumes among different ethnic groups. By integrating spatial and channel attention modules, the model adaptively emphasized key regions and discriminative features – such as embroidery motifs, colour palettes, and fabric textures – that are critical for accurate cultural identification.

As summarized in Table 3, the model achieved high classification accuracy across multiple representative ethnic groups, demonstrating its capability to differentiate costumes with complex visual patterns and

stylistic overlaps. The statistical robustness analysis confirmed that the classification results remained stable across different test subsets, with minimal variance observed among repeated trials. This consistency underscores the reliability of the attention-based CNN in managing intra-class diversity and inter-class similarity, which are common challenges in fine-grained cultural artifact recognition.

Overall, the results indicate that the incorporation of channel–spatial attention mechanisms significantly enhances the model's discriminative power and stability, enabling precise and culturally nuanced classification of ethnic costume branches within the Lijiang dataset.

Table 3: Fine-grained classification accuracy of representative ethnic costumes

Ethnic Group	Classification Accuracy (%)	95% Confidence Interval (%)	Misclassification Mainly With
Naxizu	95.3±0.42	94.88–95.72	- (low inter-class similarity)
Baizu	91.6±0.53	91.07–92.13	Yi (3.2% misclassification)
Yizu	88.7±0.61	88.09–89.31	Bai (4.8% misclassification)
Lisuzu	90.2±0.58	89.62–90.78	Pumi (2.5% misclassification)
Pumizu	89.5±0.63	88.87–90.13	Lisu (2.9% misclassification)

Source: Statistical compilation by the authors.

The experimental results reveal that the proposed attention-based CNN model achieved an overall classification accuracy of 92.1 ± 0.35%, demonstrating strong fine-grained recognition capability across multiple ethnic costume branches. Among the various categories, Naxi costumes achieved the highest classification accuracy of 95.3%, a result primarily attributed to their distinctive seven-star shawl pattern and unique ornamental design, which provided clear visual cues for the model to learn and differentiate.

The primary classification challenge emerged from inter-class similarity between visually and culturally related groups, particularly between Yi and Bai costumes. These two categories exhibited a mutual misclassification rate of approximately 4%, which, although present, represents a 4.2 pp reduction compared with the single-scale CNN baseline (8.2%). This substantial improvement underscores the advantage of incorporating channel–spatial attention mechanisms, which enable the model to focus selectively on discriminative features such as the blue-and-white tie-dye patterns typical of Bai attire and the woollen embroidery motifs characteristic of Yi clothing.

Overall, the attention-enhanced CNN not only improved the accuracy and robustness of ethnic costume classification but also demonstrated its ability to capture subtle stylistic and material cues, which are crucial for distinguishing visually similar cultural garments within the Lijiang ethnic costume dataset.

4.3 Analysis of System Real-Time Performance

To assess the deployment feasibility and operational efficiency of the optimized YOLOv11 model, its inference speed and latency were evaluated on two hardware configurations: a high-performance GPU server (NVIDIA RTX 4090) and a low-power edge computing device (NVIDIA Jetson Xavier NX). The experimental results, summarized in Table 4, provide a comprehensive comparison of inference speed (frames per second, FPS), latency per image, and model size across different deployment environments.

Table 4: Real-time performance of YOLOv11 on different hardware platforms

Hardware Platform	Inference Speed (FPS)	95% Confidence Interval (FPS)	Latency per Image (ms)	Model Size (MB)
NVIDIA RTX 4090 (Server)	142±3	139–145	7.0±0.2	22.3
NVIDIA Jetson Xavier NX (Edge)	28±2	26–30	35.7±1.8	22.3

Source: Statistical compilation by the authors.

The findings indicate that YOLOv11 performs exceptionally well in real-time processing scenarios. On the RTX 4090 GPU server, the model achieved an average inference speed of 142 ± 3 FPS, with a 95% confidence interval of 139–145 FPS and an average latency of 7.0 ± 0.2 milliseconds per image. This performance fully meets the requirements for high-concurrency tasks, such as real-time video streaming and large-scale image analysis during cultural festivals or public exhibitions. The model's efficiency under high computational load underscores its potential for integration into centralized digital heritage management systems.

In contrast, on the Jetson Xavier NX edge device, the model maintained a respectable inference speed of 28 ± 2 FPS (95% CI: 26–30 FPS) with an average latency of 35.7 ± 1.8 milliseconds. Despite the device's limited computational resources, this level of responsiveness is sufficient for mobile and field applications, such as on-site costume recognition in museums or tourist attractions, where real-time feedback and portability are critical. The model's stable performance under

constrained hardware conditions demonstrates its adaptability to decentralized and mobile deployment environments.

Another key advantage of YOLOv11 lies in its lightweight architecture, achieved through quantization and pruning techniques applied during post-training optimization. The resulting model size of 22.3 MB remains consistent across both platforms, featuring only 5.8 million parameters and 16.2 GFLOPs. This represents a 19.4% reduction in parameter count compared with YOLOv5 (7.2M) and a 54.7% reduction compared with YOLOv4 (12.8M), while maintaining superior detection accuracy. This efficient design enables the model to strike an optimal balance between computational complexity and performance, ensuring both scalability and energy efficiency.

Overall, these results confirm that the optimized YOLOv11 model delivers real-time inference capability, compact deployment size, and cross-platform versatility, making it a highly practical solution for cultural heritage digitization, intelligent tourism, and other real-world applications requiring rapid and accurate image-based recognition.

4.4 Comparison with Existing Methods

To further demonstrate the advantages of the proposed YOLOv11-based system, a comparative analysis was conducted against several representative studies on ethnic costume recognition. The comparison, summarized in Table 5, highlights three key dimensions of improvement – accuracy, scalability, and practical adaptability – relative to prior research efforts such as Ji et al. (2024), Lei et al. (2020), and Lee & Lin (2021).

Table 5: Comparison with existing ethnic costume recognition studies

Study	Ethnic Groups Covered	Accuracy (mAP/ Accuracy)	Inference Speed (FPS)	Scenario Adaptability
Ji et al. (2024)	Zhuang (1 group)	83.7% (mAP)	35 (CPU)	Indoor static images
Lei et al. (2020)	Multiple (unspecified)	89.2% (Accuracy)	28 (GPU)	Controlled lighting
Lee & Lin (2021)	Fashion apparel (non-ethnic)	96.01% (mAP)	45 (GPU)	Simple backgrounds
Proposed YOLOv11	55 branches (Naxi, Bai, Yi, etc.)	98.416% (mAP@0.5)	142 (GPU)/28 (Edge)	Low light, occlusion

Source: Statistical compilation by the authors.

In terms of coverage, the proposed system achieves an unprecedented scope by encompassing 55 ethnic branches, including Naxi, Bai, Yi, and

others. This breadth of coverage far exceeds that of earlier works, such as Ji et al. (2024), which focused solely on a single group (Zhuang), and Lei et al. (2020), which analysed multiple but unspecified ethnic categories. This extensive dataset design substantially enhances the system's cultural inclusivity and positions it as a valuable tool for large-scale digital preservation and recognition of ethnic heritage.

Regarding accuracy, the YOLOv11 model achieved a mean average precision (mAP@0.5) of 98.416%, representing a 14.716 pp improvement over Ji et al. (2024) (83.7%) and a 9.216 pp increase compared with Lei et al. (2020) (89.2%). This superior performance can be attributed to the architectural optimizations introduced in YOLOv11 – particularly the integration of multi-scale feature fusion and the C2PSA spatial attention mechanism – as well as the use of a comprehensively annotated, high-quality dataset. These enhancements collectively enable the model to capture intricate patterns, textures, and stylistic nuances characteristic of diverse ethnic costumes, resulting in robust classification and localization performance.

In terms of practicality, the YOLOv11 system demonstrates a distinct advantage in real-world adaptability. While earlier studies such as Lee & Lin (2021) achieved high accuracy (96.01%) under controlled laboratory conditions and simple backgrounds, their models lacked resilience in dynamic or low-light environments. In contrast, YOLOv11 performs consistently under challenging real-world scenarios, including outdoor festivals, low illumination, and partial occlusion. Moreover, its compatibility with both GPU servers (142 FPS) and edge devices (28 FPS) ensures scalability across different deployment contexts – from high-throughput cultural archive systems to portable recognition tools used in fieldwork and tourism applications.

Overall, these findings confirm that the proposed YOLOv11-based approach substantially outperforms existing methods in terms of accuracy, coverage, efficiency, and environmental adaptability, establishing a new benchmark for intelligent recognition systems in the field of cultural heritage digitization.

4.5 Reliability of Experimental Results, Generalization Boundaries, and Overfitting Risk Assessment

To further validate the credibility of the experimental findings and objectively evaluate the model's practical deployment potential, this section systematically analyses the reliability of test results based on leave-one-out cross-validation and cross-dataset testing, clarifies the

generalization ability boundary of the proposed YOLOv11 framework, quantifies its overfitting risk level, and puts forward targeted optimization directions for the identified limitations.

4.5.1 Reliability Verification of Experimental Results

The high detection performance achieved on the self-constructed Lijiang Ethnic Costume Dataset is first verified via leave-one-out cross--validation on the proprietary dataset, which eliminates sampling bias caused by fixed training/validation/test set partitioning and confirms that the model's performance stems from effective learning of intrinsic ethnic costume features rather than over-adaptation to the specific data distribution of the dataset. In addition, cross-dataset validation is conducted on public ethnic costume benchmark datasets (e.g., ETHNIC-CLOSET, ChinaEthnicCostume subset). The model maintains an mAP@0.5 of over 89% and a recall rate of over 85% on these external datasets, which verifies its cross-scenario adaptability and the inherent reliability of the experimental results reported in this study.

4.5.2 Generalization Ability Boundary of the Model

The generalization boundary of the proposed model is mainly reflected in two aspects. First, the model exhibits a slight performance decline in the zero-shot detection of ethnic costumes from geographically distinct regions (e.g., Southeast China). For these costumes, the mAP@0.5 decreases by approximately 6–8%, which is attributed to significant differences in textile materials, decorative motifs and garment structural features compared with Lijiang's ethnic attire, and the lack of corresponding feature distribution learning in the model. Second, the detection accuracy drops significantly in extremely complex real-world scenes, such as costumes with over 50% occlusion or low-resolution images below 320 × 320 pixels, where the mAP@0.5 falls to around 75%. This limitation is common to current fine-grained object detection models, as insufficient effective feature information leads to impaired feature extraction and discrimination capabilities.

4.5.3 Quantification of Overfitting Risk and Targeted Optimization Directions

Based on the variance-bias decomposition method, the overfitting risk of the model is quantitatively evaluated. The model exhibits a low bias (≤0.12) and a low variance (≤0.08) on the validation set, indicating a low

overall overfitting risk on the self-constructed dataset. This favourable outcome is attributed to the diverse data augmentation strategies (e.g., mosaic splicing, random rotation, colour jittering) and the three-phase training strategy with an attention-consistency constraint adopted in the training process, which effectively enhance the model's generalization ability. For the minor overfitting tendency observed in the learning of features for individual ethnic costume categories with small sample sizes (n < 50), the following targeted optimization directions are proposed:

- GAN-based data augmentation: Generate high-quality synthetic images for underrepresented ethnic costume categories to enrich the feature distribution of small samples and mitigate the overfitting tendency caused by insufficient training data;
- Self-supervised pre-training: Conduct self-supervised pre-training on large-scale unlabelled ethnic costume image datasets to enhance the model's ability to learn universal visual features of ethnic costumes and improve its cross-region and cross-scenario generalization performance;
- Multi-scale feature enhancement module: Integrate a multi-scale feature enhancement module into the Neck part of the YOLOv11 architecture to effectively extract discriminative features from low--resolution and heavily occluded costume images, expanding the model's adaptation boundary to extremely complex scenes;
- Federated learning framework: Construct a federated learning framework for ethnic costume recognition, which integrates costume image data from different regions under the premise of data privacy protection, to further improve the model's cross-region generalization ability and reduce the geographic bias in feature learning.

5. Discussion

The findings confirm that the proposed YOLOv11-based recognition framework represents a significant advancement in automated ethnic costume detection and classification. Integrating the C2PSA spatial attention mechanism with multi-scale feature fusion notably enhances the system's ability to identify fine-grained costume features under diverse, challenging conditions. With an mAP@0.5 of 98.416% and a recall rate of 95.923%, the model demonstrates high detection precision and minimal missed detections, thus validating the robustness and efficiency of the proposed architecture.

A core strength lies in the system's accuracy and adaptability. Compared with earlier YOLO architectures, YOLOv11 better detects small targets (e.g., embroidery, headpieces, jewellery), which are critical for ethnic costume identification. The C2PSA attention module increases sensitivity to local texture variations while preserving contextual awareness of complex backgrounds, thereby addressing a key limitation of previous approaches (Lee & Lin 2021; Sun et al. 2025). Statistical validation ($p < 0.001$) confirms the significance of these improvements, highlighting the value of attention mechanisms in high-resolution cultural artifact detection. The study also highlights YOLOv11's deployment efficiency. With 5.8 million parameters and 16.2 GFLOPs, the model enables real-time inference on server-grade GPUs (142 FPS) and low-power edge devices (28 FPS). This dual capability ensures scalability across computational infrastructures – from cloud-based museum archives to mobile field documentation tools – reinforcing its suitability for intelligent tourism, digital heritage archiving, and real-time cultural education, and bridging technological innovation with public engagement in preservation.

In terms of scalability, the modular architecture enables efficient adaptation beyond the Lijiang ethnic dataset. By retraining on alternative costume categories or artifact types, the framework can function as a generalizable solution for intangible cultural heritage digitization, aligning with emerging trends in computational ethnography that emphasize scalable, data-driven documentation of cultural diversity (Shao et al. 2022; Zhang & Li 2025).

From an application perspective, the framework demonstrates strong adaptability across diverse cultural contexts. Potential use cases include digital heritage documentation and museum exhibition support in Southeast Asia, South Asia, and Africa; field-based investigation and intelligent classification for indigenous costume preservation in Europe and the Americas; and mobile applications for real-time costume recognition and cultural tourism guidance in multi-ethnic regions. These scenarios highlight both the practical relevance and the global deployment potential of the proposed approach.

Despite its promising performance, several limitations remain. Inter-class similarity continues to pose challenges, particularly between Yi and Bai costumes, where a mutual misclassification rate of approximately 4% is observed due to shared colour palettes and embroidery patterns. Future research could explore contrastive learning techniques to enhance feature discrimination among visually similar categories. In addition, certain minority groups are underrepresented ($n < 50$), which

may reduce classification accuracy; this limitation could be addressed through GAN-based data augmentation to mitigate class imbalance. Finally, although the system incorporates a user feedback mechanism, model retraining still requires manual intervention. Future work should focus on integrating online learning strategies to enable automated weight updates based on aggregated feedback, thereby supporting continuous performance improvement.

Beyond its technical contributions, this study offers significant multidisciplinary value for cultural heritage preservation and ethnographic research. The fine-grained detection capability of the proposed YOLOv11 framework not only enhances visual recognition accuracy but also enables more nuanced cultural analysis and interpretation. By automatically identifying culturally salient elements – such as decorative patterns, embroidery motifs, colour schemes, and structural features – the model provides objective and quantifiable evidence for examining regional variation, cultural transmission, and the historical evolution of traditional attire.

Moreover, the framework strengthens the linkage between computer vision and practical heritage applications. It supports large-scale organization, classification, and digitization of costume collections, facilitates field-based documentation of indigenous and ethnic heritage, and contributes to the preservation and public dissemination of intangible cultural heritage. In doing so, the study integrates technical innovation with cultural interpretation and heritage practice, underscoring its relevance within computational ethnography and cultural heritage informatics.

Overall, the YOLOv11 framework delivers state-of-the-art detection accuracy and operational efficiency, while establishing a replicable methodological foundation for cultural heritage informatics. Its combination of precision, scalability, and real-world adaptability positions it as a promising benchmark for future AI-driven cultural heritage recognition and digital preservation research.

6. Conclusion

This study proposes a YOLOv11-based intelligent recognition framework for detecting and classifying ethnic costumes in Lijiang, addressing persistent challenges in accuracy, efficiency, and real-world applicability. By integrating the C2PSA spatial attention mechanism with multi-scale feature fusion, the model achieves an mAP@0.5 of 98.416% and a recall rate of 95.923%, outperforming YOLOv5, YOLOv4, and VGG16. These

results demonstrate its effectiveness in capturing fine-grained textures and structural details that are often overlooked by conventional detection approaches.

Beyond performance, the framework offers substantial practical value. Its lightweight architecture (5.8M parameters, 16.2 GFLOPs) enables real-time inference across both high-performance and edge computing environments, supporting applications such as museum digitization, intelligent tourism, and cultural education. This scalability transforms traditional heritage preservation into a dynamic, AI-assisted process. Moreover, its modular design facilitates transferability to other artifact domains, allowing retraining on diverse datasets to support broader intangible cultural heritage digitization. The integration of a user feedback mechanism further enhances sustainability by enabling iterative, data-driven model improvement.

Nevertheless, several limitations remain. Inter-class similarity – particularly between Yi and Bai costumes – continues to challenge fine-grained classification, suggesting the need for future integration of contrastive learning and self-supervised feature disentanglement. Dataset imbalance, especially for underrepresented classes, may be mitigated through GAN-based data augmentation. Additionally, incorporating online learning mechanisms will enable continuous model adaptation to evolving data and deployment contexts.

In conclusion, the proposed framework advances both deep learning-based visual recognition and the field of cultural heritage informatics. Its combination of high accuracy, computational efficiency, and scalability provides a robust foundation for intelligent systems that support the documentation, analysis, and preservation of cultural diversity.

References

Abdulrahman, A. A. & Tahir, F. S., 2021. Face Recognition Using Enhancement Discrete Wavelet Transform Based on MATLAB. *Indonesian Journal of Electrical Engineering and Computer Science* 23 (2), 1128–1136, doi: 10.11591/ijeecs.v23.i2.pp1128-1136

Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M., 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv*, 2004.10934, doi: 10.48550/arXiv.2004.10934

Chang, L., Samsudin, M. R. & Yuwen, Z., 2024. Naxi Ethnic Group's Seven-Star Sheepskin Shawl: A Semiotic Analysis. *International Journal of Art and Design* 8 (2/SI), 25–36, doi: 10.24191/ijad.v8i2/SI.3010

Ding, X., Li, T., Chen, J., Ma, L. & Zou, F., 2023. Research on the Clothing Classification of the She Ethnic Group in Different Regions Based on FPA-CNN. *Applied Sciences* 13 (17), 9676, doi: 10.3390/app13179676

Farajzadeh, N. & Hashemzadeh, M., 2021. A Deep Neural Network Based Framework for Restoring the Damaged Persian Pottery via Digital Inpainting. *Journal of Computational Science* 56, 101486, doi: 10.1016/j.jocs.2021.101486

Gan, T., Jiang, H., Yan, J. K. & Wang, H. J., 2022. 基于DeepLabv3+的苗族服饰识别网络[Miao Costume Recognition Scheme Based on DeepLabv3+]. *Journal of Guilin University of Electronic Technology* 42 (5), 412–422, doi: 10.16725/j.cnki.cn45-1351/tn.2022.05.006

Ge, H., Yu, Y. & Zhang, L., 2023. A Virtual Restoration Network of Ancient Murals via Global–Local Feature Extraction and Structural Information Guidance. *Heritage Science* 11 (1), 264, doi: 10.1186/s40494-023-01109-w

Geetha, A. S., 2024. *What is YOLOv6? A Deep Insight into the Object Detection Model*. *arXiv*, 2412.13006, doi: 10.48550/arXiv.2412.13006

He, K., Zhang, X., Ren, S. & Sun, J., 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, Los Alamitos etc., 770–778, doi: 10.1109/CVPR.2016.90

Hidayati, S. C., You, C. W., Cheng, W. H. & Hua, K. L., 2018. Learning and Recognition of Clothing Genres from Full-Body Images. *IEEE Transactions on Cybernetics* 48 (5), 1647–1659, doi: 10.1109/TCYB.2017.2712634

Ji, J., Lao, Y. & Huo, L., 2024. Convolutional Neural Network Application for Supply–Demand Matching in Zhuang Ethnic Clothing Image Classification. *Scientific Reports* 14 (1), 13348, doi: 10.1038/s41598-024-64082-9

Jin, D. & Zou, Z., 2024. 基于VGGNet-16的满族服饰识别研究 [Recognition of Manchu Clothing Based on VGGNet-16]. *Progress in Textile Science & Technology* 46 (2), 23–28, 43, doi: 10.19507/j.cnki.1673-0356.2024.02.001

Juan, H., 2021. Recognition Algorithm of Popular Elements of Ethnic Minority Traditional Clothing Based on PCA. *Scientific Programming* 2021 (1), 4834944, doi: 10.1155/2021/4834944

Lee, C.-H. & Lin, C.-W., 2021. A Two-Phase Fashion Apparel Detection Method Based on YOLOv4. *Applied Sciences* 11 (9), 3782, doi: 10.3390/app11093782

Lei, Q., Wen, B., Ouyang, Z. & Li, J., 2020. Research on Image Recognition Method of Ethnic Costume Based on VGG. In *Machine Learning for Cyber Security*. Springer, Cham, 312–325, doi: 10.1007/978-3-030-62463-7_29

Liu, S. & Lu, H., 2024. Evaluating the Visual Similarity of Southwest China's Ethnic Minority Brocade Based on Deep Learning. *arXiv*, doi: 10.48550/arXiv.2408.14060

Ma, J., Zhou, S. & Xu, P., 2025. Lightweight Visual Feature Recognition of Tibetan Opera Costumes: Framework Design and Implementation. *npj Heritage Science* 13, 660, doi: 10.1038/s40494-025-02231-7

Ma, Y. T., 2025. Innovative Promotion and Protection of Intangible Cultural Heritage: A Case Study of Bai Tie-Dyeing Techniques in Yunnan, China. *Advances in Humanities and Modern Education Research* 10, 55–64, doi: 10.70114/ahmer.2025.3.1.P55

Narag, M. J. G., Baumgartner, J. & Soriano, M., 2025. Digital Restoration of Paintings Using Convolutional Neural Network Trained on Cleaned Edges and Local Patches. In H. Liang & R. Groves (eds.) *Optics for Arts, Architecture, and Archaeology (O3A) X: 23–26 June 2025, Munich, Germany*, 139–144. SPIE, Bellingham, doi: 10.1117/12.3066542

National Library of Australia, 2003. *Guidelines for the Preservation of Digital Heritage*. UNESCO, Information Society Division, https://unesdoc.unesco.org/ark:/48223/pf0000130071 (accessed 19 May 2026).

News.qq, 2024. 走进丽江民族服饰博物馆 凝视一针一线中的文化魅力 [*Stepping into the Lijiang Ethnic Costume Museum to Appreciate the Cultural Charm in Every Stitch*], 13 Sept 2024, https://news.qq.com/rain/a/20240913A036XJ00 (accessed 19 May 2026).

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, 779–788, doi: 10.1109/CVPR.2016.91

Sang, Y. H., 2021. Study on the Traditional Costume of Naxi Nationality. *Journal of Literature and Art Studies* 11 (1), 40–42, doi: 10.17265/2159-5836/2021.01.006

Shao, Y. H., Zhang, D., Chu, H. Y., Zhang, X. Q. & Rao, Y. B., 2022. A Review of YOLO Object Detection Based on Deep Learning. *Journal of Electronics & Information Technology* 44 (10), 3697–3708, doi: 10.11999/JEIT210790

Simonyan, K. & Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*, 1409.1556, doi: 10.48550/arXiv.1409.1556

Su, Z. Z., Zhang, J. J. & Yuan, H., 2025. 基于改进YOLO v11n的轻量级民族服饰检测算法 [A Lightweight Ethnic Clothing Detection Algorithm Based on Improved YOLO v11n]. *Journal of Beijing Institute of Clothing Technology (Natural Science Edition)*, 45 (4), 84–96, doi: 10.16454/j.cnki.issn.1001-0564.2025.04.010

Sun, J., Khalid, K. A. T. & Kay, C. S., 2025. Deep Learning Models for Cultural Pattern Recognition: Preserving Intangible Heritage of Li Ethnic Subgroups through Intelligent Documentation Systems. *Future Technology* 4 (3), 119–137, doi: 10.55670/fpll.futech.4.3.12

Sun, W., Yu, J., Kang, Y., Kadry, S. & Nam, Y., 2021. Virtual Reality-Based Visual Interaction: A Framework for Classification of Ethnic Clothing Totem Patterns. *IEEE Access.* 9, 81512–81526, doi: 10.1109/ACCESS.2021.3086333

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J. & Ding, G., 2024. YOLOv10: Real-Time End-to-End Object Detection. *arXiv*, 2405.14458, doi: *10.48550/arXiv.2405.14458*

Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S., 2018. CBAM: Convolutional Block Attention Module. In V. Ferrari, M. Hebert, C. Sminchisescu & Y. Weiss (eds.) *Computer Vision – ECCV 2018*. Springer, Cham, 3–19, doi: 10.1007/978-3-030-01234-2_1

Yang, B., Nguyen, L. T. & Chansanam, W., 2025. Lightweight Attention-Augmented YOLOv5s for Accurate and Real-Time Fall Detection in Elderly Care Environments. *Sensors* 25 (23), 7365, doi: 10.3390/s25237365

Zhang, Q. & Li, M., 2025. Research on Automatic Image Classification and Feature Extraction of Ethnic Minority Costumes in Northern Guangdong Based on Deep Learning. In *DEAI '25: Proceedings of the 2nd Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*. Association for Computing Machinery, New York, 12–16, doi: 10.1145/3745238.3745241

Inteligentno prepoznavanje tradicionalnih oblačil s pomočjo modela globokega učenja za ohranjanje kulturne dediščine YOLOv11

Izvleček

Pospešena digitalizacija kulturne dediščine ponuja vse več priložnosti za ohranjanje vizualnega in simbolnega bogastva etničnih tradicij, kljub temu pa natančno prepoznavanje tradicionalnih oblačil ostaja izziv, zlasti zaradi kompleksnih tekstur, prekrivajočih se vzorcev in visoke podobnosti med posameznimi etničnimi skupinami. Članek predstavlja inteligenten pristop k prepoznavanju tradicionalnih oblačil na podlagi modela YOLOv11, uporabljenega za digitalno hrambo oblačil etničnih skupin v mestu Lijiang, Kitajska. Z vključitvijo mehanizma prostorske pozornosti C2PSA in večnivojskega združevanja značilk model omogoča učinkovito razlikovanje finih tekstilnih struktur v kompleksnih vizualnih pogojih. Za evalvacijo je bil pripravljen obsežen podatkovni nabor, ki vsebuje 4.974 slik iz 55 etničnih skupin. Eksperimentalni rezultati kažejo, da predlagana metoda dosega mAP@0.5 = 98,416 % in priklic *(recall)* 95,923 %, kar statistično značilno presega modele YOLOv5 in YOLOv4 ($p < 0,001$). Zaradi lahke zasnove (5,8 milijona parametrov, 16,2 GFLOPs) sistem omogoča sklepanje v realnem času s hitrostjo 142 sličic na sekundo na grafičnem procesorju in 28 sličic na sekundo na napravah z omejenimi viri, kar omogoča uporabo v različnih okoljih. Predlagani model predstavlja zanesljivo in uporabno rešitev za natančno prepoznavanje tradicionalnih oblačil ter prispeva k razvoju informatike kulturne dediščine, podprte z umetno inteligenco.

Ključne besede

nesnovna kulturna dediščina, globoko učenje, YOLOv11, prepoznavanje narodnih noš, prostorska pozornost