Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A dual-branch Swin Transformer fusion method for fine-grained classification of ancient Chinese jewelry

Lei Hao
Shaanxi University of Science and Technology
Yue Zhu
Shaanxi University of Science and Technology
Zihao Wang
Shaanxi University of Science and Technology
Zongming Liu
liuzm@sust.edu.cn

Shaanxi University of Science and Technology

Article

Keywords: Ancient Chinese jewelry, Fine-grained image classification, Swin Transformer, Coordinate attention, Dual-branch fusion

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11063812/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Intelligent recognition of the types and materials of ancient Chinese jewelry is essential for digital archiving and cross-collection cataloging. However, high inter-class similarity, natural aging and display-case glare, and long-tailed sample distributions hinder existing methods from jointly capturing shallow texture and deep semantics. We propose a dual-branch fusion classification method based on Swin Transformer. Coordinate attention is embedded in the first two backbone stages to enhance slender jewelry responses and suppress background interference. A dual-branch fusion head exploits shallow color-texture and deep shape semantics. Class-frequency-weighted Focal Loss and a masked multi-task strategy leverage fully and weakly annotated samples. On a self-built dataset, the model achieved 85.00% Top-1 accuracy and 82.45% macro-F1 for jewelry type, and 82.22% Top-1 accuracy and 73.51% macro-F1 for material. It outperformed ResNet-50 and ConvNeXt-T. Ablation experiments validated coordinate attention, dual-branch fusion, and Focal Loss. t-SNE and Grad-CAM showed improved subject focus and material separability. This offers a lightweight path for fine-grained multi-attribute cultural relic recognition and digital heritage preservation.

Introduction

As the epitome of traditional Chinese arts and crafts, ancient Chinese jewelry has transcended mere decoration and adornment to become an important material carrier of social hierarchy, ritual norms, aesthetic tastes, and multicultural exchange across successive dynasties. From bone hairpins of the Shang and Zhou dynasties and gold–silver inlay of the Warring States, Qin, and Han periods to repoussé and filigree of the Tang Dynasty, inlay and kingfisher-feather inlay of the Song and Yuan dynasties, and filigree inlay of the Ming and Qing dynasties, ancient Chinese jewelry has accumulated, over thousands of years of evolution, an extremely rich system of types, material craftsmanship, and decorative motifs. The exquisite combination of diverse materials such as gold, silver, jade, and gemstones, and the morphological evolution of numerous types, including Single-pronged Hairpin and Double-pronged Hairpin, crowns, bracelets, and rings, not only demonstrate the outstanding craftsmanship and aesthetic pursuits of artisans across dynasties but also embody profound ritual and ethical connotations, folk beliefs, and auspicious symbolism[1]. As ancient Chinese jewelry is a non-renewable cultural heritage resource, information on its type, materials, and decorative patterns urgently requires systematic digital recording and intelligent curation. Traditionally, however, the classification and identification of such artifacts have relied heavily on visual assessment and documentary comparison by archaeology and museology experts; for the systematic organization of large-scale museum collection data, manual analysis is inefficient and highly subjective. Therefore, deep learning-based image recognition technology has become a key pathway for automated classification and digital preservation of cultural relics[2]. Through this technical approach, structured classification and semantic annotation of ancient Chinese jewelry images can provide strong technical support for digital archiving of cultural relics, cross-collection resource integration, and the systematic transmission of intangible cultural heritage. In recent years, deep learning has provided a feasible approach for batch classification of cultural relics, and multimodal semantic modeling and value assessment for cultural heritage have gradually attracted attention: a multimodal model for cultural relics based on multi-layer attention visual encoding and a distributed language computing framework[3], and ViT-HVE, a Vision Transformer-based framework for cultural heritage value recognition and weighted assessment[4], provide references for moving cultural relic images from single classification toward semantic understanding.

Although significant progress has been made in image classification, the recognition of type and materials of ancient Chinese jewelry still faces three challenges. First, existing CNN, ViT, and hybrid models are mostly trained on natural images or modern jewelry data, showing insufficient adaptability to material aging, cluttered backgrounds, and slender typological structures in cultural relic images, and their transferability has not been fully validated. Second, type classification relies on deep morphological semantics, whereas material discrimination relies on shallow color and texture. Existing methods are mostly designed for a single task and fail to fully exploit the complementary information between shallow texture and deep semantics, making it difficult to simultaneously satisfy the parallel recognition needs of

type and material. Third, publicly available high-quality samples of ancient jewelry are limited, and the categories of type and material exhibit an extremely long-tailed distribution. In addition, some samples have only type annotations but lack material labels. General supervised strategies are prone to bias toward head classes, resulting in insufficient robustness for recognizing rare materials and typologically similar forms. Meanwhile, model decisions lack interpretability, making it difficult to provide credible evidence for cultural relic authentication.

To address the above issues, this paper proposes a dual-branch fusion classification method based on Swin Transformer. The method uses Swin-T as the backbone, leveraging its hierarchical window self-attention to simultaneously capture local texture and global semantics. Coordinate attention modules are embedded after Stage 1 and Stage 2 to enhance the spatial response of slender jewelry subjects along the horizontal and vertical directions and suppress interference from display-case frames, labels, and watermarks. A dual-branch material fusion head is designed to concatenate and fuse shallow color-texture features with deep morphological semantic features, compensating for fine-grained information loss caused by multiple downsampling operations. Class-frequency-weighted Focal Loss is introduced into the dual tasks of type and material, and a masked multi-task strategy is adopted so that weakly annotated samples without material labels participate only in type supervision, enabling unified utilization of fully annotated and weakly annotated samples. In addition, this paper constructs an image dataset of ancient Chinese jewelry, covering 9 type categories and 4 material categories, and splits it into training, validation, and test sets grouped by artifact to avoid data leakage caused by multi-view images of the same artifact.

Experimental results show that the proposed method outperforms baselines such as ResNet-50 and ConvNeXt-T in both type and material classification. Ablation experiments validate the effectiveness of coordinate attention, the dual-branch fusion head, and Focal Loss. t-SNE and Grad-CAM visualizations further demonstrate that the model can focus on the jewelry subject and improve the separability of material features. The main contributions of this paper include: constructing a reusable fine-grained dataset of ancient jewelry; proposing a coordinate-attention-enhanced Swin-T backbone; designing a shallow–deep dual-branch material fusion head; and introducing a masked multi-task training strategy for weakly annotated samples. The above work provides a lightweight and effective technical pathway for multi-attribute fine-grained recognition of cultural relics and empirical support for intelligent classification and digital preservation of cultural heritage.

1 Literature review

1.1 Analysis of Ancient Chinese Jewelry

Analysis of ancient Chinese jewelry mostly employs non-invasive or minimally invasive methods such as optical microscopy, optical coherence tomography[5], micro X-ray fluorescence imaging[6], and Raman spectroscopy[7], but the equipment is expensive and difficult to apply on a large scale; deep learning provides a feasible approach for batch preliminary screening of cultural relics. Although deep learning has made significant progress in natural image classification, directly transferring it to fine-grained recognition of ancient Chinese jewelry still faces challenges, mainly due to the visual complexity of cultural relic images and the long-tailed distribution of data[8]. Jewelry has many categories and similar types: the contours of head ornaments such as single-pronged Hairpin, double-pronged Hairpin, and combs are highly similar, and body ornaments such as crowns, necklaces, bracelets, and armlets also exhibit obvious inter-class overlap; the same type also shows considerable intra-class variation due to differences in era, region, and craft school, constituting a typical fine-grained recognition problem[9]. Meanwhile, aging phenomena such as metal oxidation, color changes, jade erosion, and texture wear are common, and photographic interference such as display-case glare, cluttered backgrounds, and watermark occlusion further weakens discriminative texture and color features, making it difficult for traditional classification models to extract stable and discriminative representations. In addition, ancient jewelry is a rare cultural relic, and publicly available high-quality images are limited. The extant quantities of different types and materials

vary greatly. This long-tailed distribution places higher demands on the generalization and robustness of models under limited samples[10]. To address the long-tail and imbalance problems in cultural heritage images, existing studies have adopted resampling, cost-sensitive loss, or deep imbalanced learning to improve minority-class recall[11].

1.2 Application of Computer Vision in Cultural Relic Image Classification

Current deep learning methods for cultural relic classification mainly fall into two categories. Convolutional neural network (CNN)-based frameworks, such as ResNet[12], EfficientNet[13], and MobileNet[14], have advantages in texture extraction owing to their local receptive fields and translation equivariance, but their global context modeling is limited, making it difficult to capture cross-region semantic associations. Transformer-based architectures, such as the Vision Transformer (ViT), obtain a global receptive field through multi-head self-attention and can model long-range dependencies, but their computational complexity grows quadratically with image size, and they output only single-resolution features without a multi-scale hierarchical structure[15], making it difficult to meet the need for joint multi-scale representation of cultural relic images. Recently, hybrid CNN–Transformer architectures have been used for fine-grained recognition of art and cultural relics, such as convolutional Transformers for art image recognition[16], multi-channel graph convolution combined with multi-directional Gaussians for microscopic image classification of porcelain[17], spatial–spectral graph convolutional networks for pigment mapping of historical artifacts[18], and CNN-based SEM image classification for pigment identification[19]. These works indicate that the recognition of material and craft attributes of cultural relics can be enhanced through multimodal, multi-scale, and graph-structure modeling.

Swin Transformer introduces hierarchical feature construction and shifted-window self-attention, reducing complexity to linear while maintaining global modeling, and achieves leading performance on classification and detection benchmarks[20]. It has been applied in fields such as industry[21], agriculture[22], medicine[23], transportation[24], remote sensing[25], and cultural research[26], providing a more promising backbone network for fine-grained recognition of cultural relics. However, existing methods are mostly designed for a single task and fail to fully exploit the complementary information between shallow texture and deep semantics, making it difficult to meet the need for parallel recognition of both type and material of ancient Chinese jewelry.

Multi-level feature fusion has already made progress in cultural heritage image analysis. Zhao et al.[8] proposed a shallow texture enhancement and cross-layer attention aggregation model for ancient mural classification, validating the value of shallow texture for material and craft recognition. Wang et al.[9] designed an Inception Transformer with dual-layer routing attention and multi-scale feature collaborative enhancement for ancient Chinese architectural images, improving the efficiency of multi-scale feature utilization. Multi-scale fusion strategies have also advanced in NSST-domain enhancement of bronze X-ray images[27] and multi-scale local geometric Transformer-Mamba classification of pottery figurine point clouds[28], indicating that cross-scale geometric and texture cues have general value for fine-grained recognition of cultural relics. Shallow color, texture, and surface texture are crucial for material discrimination, while deep morphological semantics are indispensable for type classification; only through their synergy can accurate recognition be achieved. Coordinate attention decouples 2D positional information into horizontal and vertical 1D encodings, enhancing spatial position awareness without significantly increasing computational cost[29], which helps suppress background interference such as display-case frames and labels and focus on the artifact subject.

1.3 Jewelry and Gemstone Image Classification

Jewelry and gemstone image classification has also been explored. Khalilian et al. combined CNN with long short-term memory networks for jewelry and rock classification using laser-induced breakdown spectroscopy data, achieving interpretable discrimination[30]. Singh and Kaewprapha early applied CNN to five-class jewelry image classification,

validating the feasibility of CNN in jewelry visual recognition[31]; Chow and Reyes-Aldasoro constructed an automated visual pipeline for gemstone classification[32]; although the generative adversarial network proposed by Alex et al. targets medical images, its idea of using data augmentation to alleviate small-sample problems offers useful insights for cultural relic recognition[33]. The above works mostly target modern jewelry or a single gemstone category[34] and have not addressed the need for fine-grained recognition of both type and material of ancient Chinese jewelry. In traditional Chinese painting sentiment classification, an attention-guided CNN–Transformer hybrid model also validated the effectiveness of hybrid architectures for semantic modeling of artistic images[35].

2 Research methodology

2.1 Dataset preparation

The images in this paper were obtained from public collection catalogs of various museums and archaeological illustration albums, as shown in Fig. 1, covering 9 type categories—Single-pronged Hairpin, Double-pronged Hairpin, Comb, Crown, Ear ornament, Neck ornament, Bracelet, Armlet, and Ring—and 4 material categories: gold, silver, jade, and gemstones. Typical examples are shown in Fig. 1. A total of 900 images were originally collected. After category filtering and removal of duplicate samples, 835 valid images were obtained, consolidated into 711 independent artifacts. All 835 images have type annotations; among them, 757 have material annotations for the four categories of gold, silver, jade, and gemstones, while the remaining 78 are marked as unlabeled due to missing material information, such as rare material data including bone, pearl, coral, and kingfisher-feather inlay, and are not included in the material classification task. To avoid data leakage caused by multi-view images of the same artifact being split across training and test sets, the split was performed with artifacts as groups. At an 8:1:1 ratio, the training set was assigned 569 artifacts totaling 675 images, and the validation and test sets were each assigned 71 artifacts, corresponding to 80 images each.

To mitigate the risk of overfitting caused by small samples and long-tailed class distributions, a mild data augmentation strategy was applied only to the training set: random horizontal flipping, small-angle rotation of ± 15°, random resized cropping, slight color jitter, and brightness–contrast perturbation. To enrich sample diversity without compromising the realistic semantics of jewelry artifacts, the original 675 training images were expanded to 1800 augmented samples. During augmentation, each type category was balanced to 200 images in the type dimension, and minority classes were oversampled in the material dimension to alleviate the long-tailed distribution. The validation and test sets retained the original images throughout, with no data augmentation applied. Complete dataset statistics are presented in Table 1.

Table 1
Statistics of the ancient Chinese jewelry datasets

Artifact category	**Single-pronged Hairpin**	**Double-pronged Hairpin**	**Comb**	**Crown**	**Ear Ornament**	**Neck Ornament**	**Bracelet**	**Armlet**	**Ring**	**Total**
Training set	189	123	66	36	96	34	49	22	62	675
Validation set	24	11	7	5	9	5	8	2	9	80
Test set	19	17	7	4	13	4	6	3	7	80
Material category	**Gold**	**Silver**		**Jade**		**Gemstone**	**Unlabeled**		**Total**	
Training set	460	64		64		20	67		675	
Validation set	40	14		13		7	6		80	
Test set	38	12		15		10	5		80	

Note: The training set contains 569 artifacts with a total of 675 original images, which were expanded to 1800 through offline data augmentation; the validation and test sets each contain 71 artifacts and 80 images. All 835 images have type annotations; the material classification task uses only the four valid annotated categories of gold, silver, jade, and gemstone, and unlabeled samples do not participate in material classification training or evaluation.

As shown in Fig. 2, the dataset contains several typical cases that are difficult to classify. In jewelry images, natural aging and the acquisition environment jointly cause metal oxidation, color change, jade erosion, and decorative pattern abrasion on artifact surfaces, while at the image level, display-case glare, cluttered backgrounds, and watermark occlusion are common. These factors compress effective inter-class visual differences and weaken fine-grained cues related to material and type, thereby imposing higher demands on the robustness and discriminative power of model representations.

2.2 Swin Transformer Backbone

The Vision Transformer (ViT) can effectively model long-range dependencies through global self-attention, but its computational complexity grows quadratically with image size, and it outputs a single-resolution feature map without a multi-scale hierarchical structure, making it difficult to meet the needs of fine-grained texture recognition in cultural relic images. Swin Transformer effectively overcomes these limitations by introducing hierarchical feature construction and a shifted window self-attention mechanism [20]. On the one hand, it progressively downsamples through Patch Merging to form a hierarchical representation comprising four stages, with feature map resolutions of 1/4, 1/8, 1/16, and 1/32 of the input size, respectively, and the number of channels doubling at each stage. This constructs a hierarchical feature pyramid similar to that of convolutional neural networks, aligning with the cultural relic cognition logic of “shallow layers extracting texture and deep layers representing morphology.” On the other hand, it constrains self-attention computation within local windows and enables cross-window information interaction through shifting operations between adjacent windows, reducing complexity to linear while maintaining global modeling capability.

The computational complexities of standard multi-head self-attention (MSA) and window multi-head self-attention (W-MSA) are respectively:

$$\Omega\ (MSA) = 4hwC^2 + 2(hw)^2C$$

1

$$\Omega\left(W-MSA\right)=4hwC^2+2M^2hwC$$

2

Where *h* and *w* are the height and width of the feature map, respectively, *C* is the number of channels, and *M* is the window size. When the input image resolution is high, $2M^2hwC \ll 2(hw)^2C$, the computational cost is significantly reduced.

For the fine-grained recognition task of ancient Chinese jewelry, this paper adopts Swin-Tiny as the feature extraction backbone and loads ImageNet-1K pretrained weights for transfer learning, leveraging general visual priors to enhance model generalization. Compared with traditional CNN models, Swin-T, with its hierarchical window self-attention mechanism, can simultaneously capture local texture details and global semantic information at a controllable computational cost, making it suitable for multi-scale visual recognition tasks.

2.3 Overall Framework

This paper proposes a dual-task multi-scale feature fusion method based on the Swin-T backbone network, and the overall architecture is shown in Fig. 3. The network takes a 224×224 color image as input, which sequentially passes through Patch Partition and Linear Embedding into the Swin-T backbone. Swin-T contains four stages and performs progressive downsampling through Patch Merging to output multi-scale features. In this paper, coordinate attention (CA) modules are introduced after Stage 1 and Stage 2, respectively, to spatially weight the feature maps, suppress irrelevant background interference such as display-case frames, labels, and watermarks, and guide the network to focus on jewelry target regions.

At the output of the backbone, the network splits into two task branches. Stage 4 features are processed by global average pooling and layer normalization to obtain a 768-dimensional vector. One branch feeds into the type classification head, performing 9-class jewelry type recognition through Dropout and a fully connected layer; the other branch is mapped to 256 dimensions as the semantic branch input to the material fusion head. Meanwhile, Stage 4 deep features and CA-enhanced Stage 2 shallow features are fed into the dual-branch fusion head to perform material recognition.

2.4 Coordinate Attention Enhancement

Existing attention mechanisms have inherent limitations in feature enhancement. Channel attention, such as SENet, only encodes inter-channel importance while ignoring spatial position information; spatial attention, such as CBAM, can supplement position awareness but relies on large-receptive-field convolutions and incurs considerable computational redundancy. Coordinate attention decouples 2D positional information into horizontal and vertical 1D encodings and embeds them into channel attention, striking a balance between positional awareness accuracy and computational lightness [29].

The specific procedure of CA is as follows: given an input feature $X \in \mathbb{R}^{C\times H\times W}$, global average pooling is first performed along the horizontal and vertical directions separately to generate direction-aware feature vectors:

$$z_c^h\left(h\right)=\frac{1}{W}\sum_{0\leq i<W}x_c(h,i)$$

3

$$z_c^w\left(w\right)=\frac{1}{H}\sum_{0\leq j<H}x_c(j,w)$$

4

Subsequently, the features from the two directions are concatenated and passed through a 1×1 convolution F1, batch normalization, and SiLU activation for dimensionality reduction (reduction ratio r = 16). They are then split back into two directional vectors f^h and f^w, which separately learn weights for the two directions and are each passed through a 1×1 convolution and a Sigmoid function to generate directional attention weights. Finally, the weights are multiplied element-wise with the original feature:

$$f = \delta\ \left(F_1\left(\left[z^h, z^w\right]\right)\right)$$

5

$$g^h = \delta\ \left(F_h\left(f^h\right)\right) \quad , g^w = \delta\ \left(F_w\left(f^w\right)\right)$$

6

$$y_c(i,j) = x_c(i,j) \times\ g_c^h\left(i\right) \times\ g_c^w\left(j\right)$$

7

For ancient Chinese jewelry images, the CA module has good task adaptability: jewelry artifacts are mostly slender, left–right symmetric, and centrally arranged in composition, with discriminative subjects extending along the vertical or horizontal direction; meanwhile, interfering elements in the image background, such as display-case frames, text captions, and watermarks, also exhibit orthogonal strip-like distributions. CA encodes positional information along two orthogonal directions, thereby effectively enhancing artifact subject features and suppressing background interference. The structure of the coordinate attention module is shown in Fig. 4.

2.5 Material Dual-Branch Fine-Grained Fusion Head

Material discrimination requires integrating feature information from different levels of the network: color and surface texture information, such as the luster of metal, the grayish-white surface formed by oxidation of silver ornaments, and the warm, translucent texture of jade, is mainly contained in shallow features. However, gold ornaments and amber-like gemstones can exhibit similar warm yellow appearances; relying solely on shallow color and texture easily leads to confusion, and deep morphological semantic information extracted by the deep network is needed for discrimination. If only deep global features are used for material classification, after multiple downsampling operations, the color and texture cues carried by shallow features will be largely lost, which is detrimental to material recognition.

To this end, this paper designs a dual-branch material fusion head, whose structure is shown in Fig. 5. The texture branch performs global average pooling on the CA-enhanced Stage 2 shallow features and maps them to 256 dimensions through a fully connected layer; the semantic branch performs global average pooling and layer normalization on Stage 4 deep features and similarly maps them to 256-dimensional features. The features from the two branches are concatenated to obtain a 512-dimensional joint feature vector, which is then fed into a two-layer fully connected network with Dropout (0.3) to output prediction logits for the four material categories: gold, silver, jade, and gemstone. This fusion head introduces only about 0.37 M additional parameters, resulting in low model overhead.

2.6 Multi-Task Joint Training Loss

Both the type classification and material classification tasks in this paper suffer from class imbalance and small-sample problems. In the type task, the number of Single-pronged Hairpin and Double-pronged Hairpin samples exceeds one hundred, whereas classes such as crowns have only a few dozen images; in the material task, the number of gold

samples is significantly larger than that of silver samples. To address the above long-tailed distribution problem, this paper introduces class-frequency-weighted Focal Loss in both task classification heads for balanced optimization[36]:

$$FL\left(p_t\right) = -\alpha_t \cdot \left(1 - p_t\right)^{\gamma} \cdot \log(p_t)$$

8

where p_t is the predicted probability of the true class, $x\ \gamma$ is the focusing parameter, and a_t is the class-frequency weight, used to suppress the contribution of head classes and boost long-tailed classes.

In addition, the dataset contains some samples with weak material annotations. These samples mainly include organic jewelry such as bone, pearl, coral, and kingfisher-feather inlay, which cannot be assigned to the four material categories of gold, silver, jade, and gemstone, and therefore lack valid material labels. To make full use of all data, this paper proposes a masked multi-task loss strategy: the type loss supervises all samples in a batch, while the material loss is computed only on the subset of samples with valid material labels. The overall joint loss function is defined as:

$$L = L_{type} + \lambda \cdot \frac{1}{|S|}\sum_{i \in S} L_{material}^{(i)}$$

9

where S is the set of indices of samples with material labels in the batch, and λ is the material task weight coefficient, defaulting to 0.5. When $|S| = 0$, this loss term is set to 0. This design allows samples without material labels to still contribute to type supervision and backbone gradient updates, enabling unified utilization of fully annotated and weakly annotated samples.

3 Results

3.1 Experimental set up

3.1.1 Hardware and Software Environment

All networks were initialized with ImageNet-1K pretrained weights. During training, the input size was fixed at 224×224, the optimizer was AdamW (learning rate 1e-4, weight decay 1e-4), a cosine annealing schedule was used, the batch size was 16, and the maximum number of training epochs was 60. To ensure reliable results, experiments were independently repeated with 3 different random seeds, and the results in Table 2 are reported as mean ± standard deviation. Experimental environment: Windows 11, PyTorch 2.1.2, torchvision 0.16.2, GPU: NVIDIA GeForce GTX 1660 SUPER.

3.1.2 Evaluation Metrics

For the type and material tasks, Top-1 accuracy (Acc) and macro-average F1 (macro-F1) on the test set are reported separately. The test set contains 80 images in total. type classification is evaluated on all 80 images; material classification excludes unlabeled samples and computes Acc and F1 on 75 images with valid material annotations.

Top-1 accuracy measures the proportion of correctly predicted samples among all test samples:

$$\mathrm{Acc}=\frac{\mathrm{N}_{\mathrm{correct}}}{\mathrm{N}_{\mathrm{total}}}$$

10

Where $N_{correct}$ is the number of correctly predicted samples, and N_{total} is the total number of samples in the test set.

Macro-average F1 takes the arithmetic mean of the per-class F1 scores, which can more faithfully reflect the recognition ability for minority classes under a long-tailed distribution. Therefore, this paper conducts model comparison and analysis based on macro-average F1.

$$Macro - F1 = \frac{1}{C}\sum_{i=1}^{C} F1_i$$

11

$$F1_i = \frac{2P_iR_i}{P_i + R_i}$$

12

Where C is the total number of classes; P_i and R_i are the precision and recall of the i-th class, respectively.

3.2 Model Performance Evaluation Metrics

Figure 6 shows the accuracy curves of different models on the validation set. In Fig. 6(a) for type classification, the proposed method begins to lead from around the 5th epoch, and after convergence its validation accuracy stabilizes between 0.81 and 0.84; ResNet-50 ranks second, at approximately 0.78–0.81; the remaining models are approximately 0.72–0.75. The proposed method maintains a stable advantage of 3 to 6 percentage points over mainstream CNN baselines, indicating that Swin Transformer combined with coordinate attention facilitates the learning of type semantic features. Material classification has lower overall accuracy than type classification, indicating greater task difficulty. The proposed method still ranks first, at approximately 0.71–0.74 after convergence; ResNet-50 and ConvNeXt-T are approximately 0.68–0.72; EfficientNet-B0 and MobileNetV2 are approximately 0.65–0.68.

In Fig. 6(b), the standard deviation shading of the material curves is wider, indicating that this task is more affected by sample imbalance and random initialization. The proposed method still maintains a leading position on the more difficult material task, with relatively controllable fluctuations and good generalization stability.

The proposed method achieved the best performance on both validation tasks. Its advantage is particularly pronounced in type classification, the primary task, validating the effectiveness of Swin Transformer's global modeling capability combined with CA attention for extracting type semantic features. It also leads in material classification, the auxiliary task; however, owing to class imbalance in material categories, its overall accuracy is lower than that of type classification, and the gaps among models are smaller. Overall, the proposed method exhibits superior and more stable classification performance on the validation set.

To verify the effectiveness of the proposed method, it was compared with four mainstream backbones under the same multi-task protocol: ResNet-50, MobileNetV2, EfficientNet-B0, and ConvNeXt-T, all loaded with ImageNet pretrained weights. The comparison models uniformly adopted a type head plus a linear material head to compare backbone differences; the contribution of the dual-branch fusion head was verified by ablation experiments. The results are shown in Table 2.

Table 2
Comparison of different methods on the test set

Method	type Acc	type macro-F1	Material Acc	Material macro-F1	Params
ResNet-50	83.33%±1.83%	78.87%±3.92%	79.56%±2.12%	67.27%±1.23%	23.53 M
MobileNetV2	81.67%±3.77%	76.62%±1.30%	73.78%±3.68%	58.16%±4.89%	2.24 M
EfficientNet-B0	81.25%±3.31%	78.72%±3.40%	76.44%±2.89%	67.72%±1.24%	4.02 M
ConvNeXt-T	82.50%±2.52%	79.99%±2.22%	80.89%±4.34%	69.49%±6.32%	27.83 M
Ours	85.00%±1.44%	82.45%±2.82%	82.22%±1.70%	73.51%±4.77%	27.91 M

In the type classification task, the proposed method achieved a Top-1 accuracy of 85.00%±1.44% and a macro-average F1 of 82.45%±2.82%, ranking best among the five comparison methods on both metrics. Among the baseline models, ResNet-50 achieved a Top-1 accuracy of 83.33%±1.83%, and ConvNeXt-T achieved a macro-average F1 of 79.99%±2.22%. Compared with these, the proposed method improves accuracy by 1.67 percentage points and macro-average F1 by 2.46 percentage points. Macro-average F1 assigns equal weight to all classes, which can mitigate the influence of classes with abundant samples such as Single-pronged Hairpin and Double-pronged Hairpin. The significant improvement in this metric indicates that the model still has good recognition ability for long-tailed types such as armlets and neck ornaments, and that the overall performance gain does not rely solely on head classes with sufficient samples. In terms of training stability, the standard deviation of accuracy for the proposed method is 1.44%, with a smaller standard deviation than all baseline models (1.83%–3.77%). Meanwhile, the performance improvement does not come with a significant increase in parameter count; the proposed model has 27.91 M parameters, basically at the same order of magnitude as ConvNeXt-T (27.83 M).

In the material classification task, the performance gain of the proposed method is more pronounced. The model achieved a Top-1 accuracy of 82.22%±1.70% and a macro-average F1 of 73.51%±4.77%. For comparison, ConvNeXt-T achieved Top-1 accuracy and macro-average F1 of 80.89%±4.34% and 69.49%±6.32%, respectively; compared with this baseline, the proposed method improves accuracy and macro-average F1 by 1.33 and 4.02 percentage points, respectively. Among samples with valid material annotations, gold accounts for nearly 50%, and the test set contains 38/75 gold samples. Overall accuracy is easily dominated by the high-proportion gold class, whereas macro-average F1 can better reflect the recognition performance for rare materials such as silver and gemstones. The improvement in macro-average F1 is notably larger than that in accuracy, consistent with the design intention of the dual-branch fusion head: the shallow color and texture features preserved by this module play a more critical role in discriminating rare materials. In terms of stability across repeated experiments, the standard deviation of material macro-average F1 for ConvNeXt-T is as high as 6.32%, indicating significant performance fluctuations across training runs, while the corresponding standard deviation of the proposed method is 4.77%; moreover, the standard deviation of material accuracy for the proposed method is only 1.70%, the lowest among all comparison methods. In addition, it can be observed that the impact of backbone network scale on the material task is greater than that on the type task. MobileNetV2 and EfficientNet-B0 have similar accuracy levels on the type task, at 81.67% and 81.25%, respectively; however, their gap in material macro-average F1 reaches 9.56 percentage points, with corresponding results of 58.16% and 67.72%, among which MobileNetV2 is the lowest among all methods. This phenomenon indicates that fine-grained material recognition places higher demands on the feature representation capability of the backbone network.

3.3 Ablation Experiments

To verify the contribution of each improved component in this paper, multiple model variants were set up with Swin-T as the backbone for ablation experiments: (1) the full model; (2) removing the coordinate attention module; (3) removing the dual-branch fusion head, with the material branch using only deep features of the network for classification; (4) replacing

Focal Loss with standard cross-entropy loss; (5) the baseline model, retaining only the Swin-T backbone and a simple linear dual head, without introducing coordinate attention, the dual-branch fusion head, or Focal Loss. The ablation experiment results are shown in Table 3.

Table 3
Ablation experiment results

Model variant	type Acc	type macro-F1	Material Acc	Material macro-F1	Params
Full model	85%±1.44%	82.45%±2.82%	82.22%±1.70%	73.51%±4.77%	27.91 M
w/o coordinate attention	77.50%±2.36%	75.88%±3.15%	77.33%±2.48%	68.15%±6.12%	27.90 M
w/o dual-branch fusion head	81.25%±1.98%	80.42%±2.53%	80.00%±2.15%	69.83%±4.38%	27.80 M
w/oFocal Loss	80.00%±2.67%	77.36%±3.44%	81.33%±1.93%	70.56%±5.23%	27.91 M
Swin-T	78.75%±2.85%	78.00%±3.27%	78.67%±2.61%	69.27%±5.91%	27.53 M

The experimental results are shown in Table 3. After removing the coordinate attention module, Acc and macro-F1 for type classification decreased by 7.50 and 6.57 percentage points, respectively, and Acc and macro-F1 for material classification decreased by 4.89 and 5.36 percentage points, making it the component with the largest contribution to overall performance, while the two coordinate attention modules add only about 0.01 M parameters. The type Acc of this ablation group is 77.50%, slightly lower than the Swin-T baseline of 78.75%. This phenomenon can be explained by random training fluctuations and module coupling effects. The performance fluctuation under multi-seed averages is usually 1–2 percentage points, but the performance degradation caused by removing coordinate attention is much larger than this range, so the conclusion is not affected. When the dual-branch fusion head is removed, material Acc and macro-F1 decrease by 2.22 and 3.68 percentage points, respectively, indicating that shallow texture features are crucial for material recognition; type Acc simultaneously decreases by 3.75 percentage points, which is caused by the coupling effect of material branch gradients on the shared backbone under multi-task joint training. Replacing Focal Loss with ordinary cross-entropy loss reduces type Acc and macro-F1 by 5.00 and 5.09 percentage points, respectively, and material macro-F1 by 2.95 percentage points, while material Acc decreases only slightly by 0.89 percentage points. The reason for this difference is that material classification accuracy is mainly dominated by gold samples, which account for 50.7% of the test set, and is insensitive to changes in minority classes; macro-F1 weights all classes equally and better reflects the performance degradation of long-tailed classes such as silver, jade, and gemstone. This also explains why the material macro-F1 of the full model (73.51%) is noticeably lower than its Acc (82.22%): class-weighted loss can alleviate the bias caused by the long-tailed distribution but cannot completely solve the inherent problem of insufficient samples for rare materials. Overall, compared with the Swin-T baseline, the proposed model improves Acc and macro-F1 by 6.25 and 4.45 percentage points on the type task, and by 3.55 and 4.24 percentage points on the material task; the number of parameters increases by only about 0.38 M, of which the dual-branch fusion head accounts for 0.37 M and coordinate attention for less than 0.01 M, indicating low computational overhead for the performance improvement. The current model still has considerable room for optimization in rare material recognition.

3.4 Visualization Analysis

Figure 7 shows the t-SNE feature distributions of the full model proposed in this paper and the Swin-T baseline on a total of 160 images from the validation and test sets, colored by ground-truth class. From the material perspective, the four classes of Gold, Silver, Jade, and Gemstone samples from the Swin-T baseline overlap severely in the two-dimensional projection space, lacking clear decision boundaries between classes, indicating that the features extracted by the baseline model cannot effectively distinguish different materials. In contrast, for the proposed method, Gold and Jade form two relatively clear distribution bands along the upper right and lower right, respectively, Silver is concentrated in the upper left region, and Gemstone tends to be distributed toward the lower part, with the four material classes overall showing a clear

trend of inter-class separation. This improvement benefits from the explicit collaborative utilization of shallow color-texture features and deep semantic features by the dual-branch material fusion head, enabling the model to capture fine-grained discriminative cues such as metallic luster, oxidation hue, and jade texture. From the type perspective, the nine type classes of both methods do not form compact cluster structures in the two-dimensional projection, and considerable overlap remains among classes. This indicates that inter-class differences in type are not sufficiently significant in the low-dimensional embedding space, and their discrimination mainly relies on high-dimensional semantic features. Overall, the proposed method achieves significant improvement over the Swin-T baseline in material feature separability, while for type recognition, CA and the fusion structure mainly preserve the original feature distribution without negatively affecting type features.

Figure 8(a) shows the type classification confusion matrix of the full model on the test set, row-normalized (each row sums to 1; rows represent predicted classes and columns represent true classes), so the values on the diagonal indicate the precision of each class. As can be seen, the precision of Double-pronged Hairpin, Crown, Neck Ornament, and Armlet is 1.00; the precision of Single-pronged Hairpin, Comb, Ear Ornament, and Ring is 0.79, 0.88, 0.91, and 0.88, respectively; Bracelet has the lowest precision, at only 0.56. Misclassifications are mainly concentrated between classes with similar morphology: among samples predicted as Single-pronged Hairpin, 17% are truly Double-pronged Hairpin and 4% are truly Ring; among samples predicted as Comb, 12% are truly Double-pronged Hairpin; among samples predicted as Ear Ornament, 9% are truly Neck Ornament; among samples predicted as Bracelet, 22% are truly Ear Ornament, 22% are truly Armlet, and 56% are truly Bracelet; among samples predicted as Ring, 12% are truly Bracelet. These misclassifications reflect the subtle contour differences between single- and double-pronged hairpins and annular ornaments, as well as the difficulty of distinguishing armlets, bracelets, and rings in terms of thickness and diameter.

Figure 8(b) shows the material classification confusion matrix, also row-normalized, computed over 75 valid annotated images. The precision values on the diagonal are Gold 0.85, Silver 0.82, Jade 0.92, and Gemstone 0.67. The misclassification pattern is as follows: Gold and Silver are confused with each other; among samples predicted as Gold, 5% are truly Silver, and among samples predicted as Silver, 18% are truly Gold, which is related to the similar appearance caused by gilding processes and the gray oxidation of silver ornaments. In addition, among samples predicted as Gold, 7% are truly Jade and 3% are truly Gemstone; among samples predicted as Gemstone, 17% are truly Gold, 8% are truly Silver, and 8% are truly Jade. Overall, the bottleneck of material recognition lies, on the one hand, in the fine-grained distinction of materials with similar appearances and, on the other hand, in the scarcity of samples for rare classes. Gemstone has only 10 test images, so a single misclassification can affect precision by as much as 10 percentage points.

As shown in Fig. 9(a), Grad-CAM distributions indicate that the proposed method focuses more accurately on discriminative jewelry regions in both type and material classification tasks, with more concentrated and reasonable attention than baseline models. In type classification, the heatmap regions of the proposed method concentrate on main body parts such as single-pronged Hairpin, double-pronged Hairpin, comb contours, ear-ornament pendants, and crown tops, with clear boundaries, effectively suppressing display-case glare, background texture, and watermarks. This benefits from coordinate attention encoding positional information along horizontal and vertical directions, strengthening spatial responses on slender jewelry subjects. Compared with baseline models, the heatmap regions of Swin-T and CNN models are more dispersed: ResNet-50 and ConvNeXt-T tend to diffuse responses to background or non-discriminative regions, MobileNetV2 and EfficientNet-B0 exhibit localization shifts, and some samples focus on multiple irrelevant regions, failing to stably lock onto discriminative parts.

As shown in Fig. 9(b), in the material classification task, the attention of the Swin-T baseline is concentrated on local texture details, with relatively sharp heatmap regions; the proposed method has broader heatmap coverage, accounting for the overall morphology and structural information of artifacts in addition to local texture. This difference arises from the cross-task feature interaction introduced by the dual-branch fusion mechanism, enabling the model to rely not only on local material cues but also to comprehensively utilize type structure to assist material discrimination. Quantitatively, the

material Acc of the proposed method is 82.22%, higher than the 78.67% of Swin-T, indicating that a more global attention distribution is instead beneficial to material classification, validating the effectiveness of multi-task feature complementarity. The above visualization results are consistent with the ablation experiments, indicating that coordinate attention and the dual-branch fusion head can enhance the model's ability to focus on jewelry subjects and material cues, thereby achieving better performance in type and material classification tasks.

4 Conclusion

This paper proposes an image classification method for ancient Chinese jewelry based on Swin Transformer dual-branch fusion, aiming to address challenges such as high inter-class similarity among types, interference from natural aging and display-case glare, and long-tailed sample distributions. The method enhances shallow spatial responses through coordinate attention, uses a dual-branch fusion head to jointly model shallow color-texture and deep morphological semantics, and adopts class-frequency-weighted Focal Loss and a masked multi-task strategy to unify the use of fully annotated and weakly annotated samples. The model achieves type Top-1 accuracy of 85.00% and macro-average F1 of 82.45%, and material Top-1 accuracy of 82.22% and macro-average F1 of 73.51%, outperforming baselines such as ResNet-50 and ConvNeXt-T overall. Its core innovations include: a coordinate attention enhancement module for slender jewelry subjects, a shallow–deep dual-branch material fusion head, a multi-task training strategy for long-tailed and weakly annotated data, and a reusable fine-grained dataset of ancient jewelry.

The significance of this work lies not only in improving classification performance but also in promoting interdisciplinary integration between artificial intelligence and cultural heritage. Experiments show that architecture design respecting the visual characteristics of cultural relic images can effectively improve fine-grained recognition and provide a methodological reference for computational art analysis that integrates domain-specific knowledge. The contributions of this study to cultural computing also include the constructed dataset and evaluation protocol, which can support digital archiving of cultural relics, cross-collection resource integration, and subsequent research. However, its broader applicability still depends on validation across more materials, types, and artistic traditions.

Future work can proceed in both short-term and long-term directions. In the short term, high-quality samples for rare classes can be expanded, multimodal priors such as craft schools, historical periods, and documentary information can be introduced, and discriminability can be enhanced through local key-region cropping. In the long term, a fine-grained classification system oriented toward cultural relic attributes needs to be constructed, combining weakly supervised localization and cross-domain multi-scale fusion to improve model sensitivity to local geometric differences, and explainable artificial intelligence techniques can be used to uncover decision rationales. The ultimate goal is to build an intelligent system that can serve as a bridge between traditional cultural relic knowledge and modern computational methods, providing sustainable technical support for the digital preservation and knowledge dissemination of ancient jewelry while maintaining cultural authenticity and academic rigor.

Declarations

Competing interests

The authors declare no competing interests.

Funding

General Project of Ministry of Education Foundation on Humanities and Social Sciences (Project No.23YJC760034).

Author Contribution

Lei Hao: Conceived and implemented the project; performed data curation, formal analysis, and methodology design; wrote and edited the full manuscript. Zongming Liu: Designed the overall framework, reviewed literature, and supervised the study throughout. Yue Zhu: Performed data analysis and project implementation; prepared literature and figures. Zihao Wang: Provided technical support.

Acknowledgement

This study was supported by the General Project of Humanities and Social Sciences Research of the Ministry of Education (Project No.: 23YJC760034). We hereby express our gratitude.

Data Availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Li, Y. et al. *History of ancient Chinese jewelry* (Jiangsu Phoenix Literature and Art Publishing House, 2020).
2. Wang, Y. et al. Construction of Cultural Heritage Knowledge Graph Based on Graph Attention Neural Network. *Appl. Sci.* **14**, 8231 (2024).
3. Xu, Z. Y. et al. ACM,. Multimodal model of cultural relics based on multi-layer attention visual encoding and distributed language computing framework. In Proceedings of the 2025 International Conference on Culture Heritage and Digital Humanities (2025).
4. ViT-HVE. a vision transformer-based framework for recognition and weighted evaluation of cultural heritage values. *npj Herit. Sci.* **13**, (2025).
5. Du, J. et al. A Sui-Tang dynasty woman's crown: analyzing form, glass and class. *Herit. Sci.* **10**, 45 (2022).
6. Shi, Y. et al. Transmission and innovation on gold granulation: the application of tin for soldering techniques in ancient China. *Herit. Sci.* **10**, 122 (2022).
7. Liu, Y. et al. Demystifying ancient filigree art: Microanalytical study of gold earrings from Dongheigou cemetery (4th-2nd century BCE) in north-west China. *J. Archaeol. Science: Rep.* **41**, 103321 (2022).
8. Zhao, P. et al. An ancient mural classification model with shallow texture enhancement and cross-layer attention aggregation. *npj Herit. Sci.* (2026).
9. Wang, X. et al. An improved Inception Transformer with Bi-Level Routing Attention and multi-scale feature collaborative enhancement for Chinese ancient architecture image classification. *npj Herit. Sci.* (2026).
10. A coordinate-to. -semantic attention network for multi-label ancient Chinese architecture image classification. *npj Herit. Sci.* (2025).
11. Imbalanced Classification of Cultural Heritage Images Using Deep Neural Networks. In 2024 First International Conference on Pioneering Developments in Computer Science. IEEE, (2024).
12. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 770–778 (2016).
13. Tan, M., Le, Q. V. & EfficientNet Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning 6105–6114 (2019).

14. Howard, A. G. et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv* :170404861 (2017).
15. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (2021).
16. Liu, Y., Bai, H. & Wang, J. Fine-art recognition using convolutional transformers. *PeerJ Comput. Sci.* **10**, e2409 (2024).
17. Multiple channel GCN with. multiple directional Gaussian for porcelain microscopic image classification. *npj Herit. Sci.* **13**, 254 (2025).
18. Spatial-spectral graph. convolutional network for automatic pigment mapping of historical artifacts. *npj Herit. Sci.* **13**, 106 (2025).
19. Automated classification of natural. and synthetic pigments from SEM images using CNN-based deep learning. *npj Herit. Sci.* **13**, 607 (2025).
20. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision 10012–10022 (2021).
21. Jiang, J. et al. Masked Swin Transformer Unet for Industrial Anomaly Detection. *IEEE Trans. Industr. Inf.* **19**, 2200–2209 (2023).
22. Han, R. et al. Lightweight real-time detection of sweet potato virus disease by a hybrid CNN–Transformer network. *Comput. Electron. Agric.* **255**, 112319 (2026).
23. Hatamizadeh, A. et al. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (eds (eds Crimi, A. & Bakas, S.) 272–284 (Springer, Cham, (2022).
24. Su, X., Liu, S., Li, R., Bing, Z. & Knoll, A. Efficient Stereo Matching Using Swin Transformer and Multilevel Feature Consistency in Autonomous Mobile Systems. *IEEE Trans. Industr. Inf.* **20**, 7957–7965 (2024).
25. Duan, Y., Song, C., Zhang, Y., Cheng, P. & Mei, S. STMSF: Swin Transformer with Multi-Scale Fusion for Remote Sensing Scene Classification. *Remote Sens.* **17**, 668 (2025).
26. Zhao, J. et al. Multi-scale progressive Swin Transformer for ancient Tai Lue palm leaf manuscript restoration. *npj Herit. Sci.* **14**, 376 (2026).
27. Wu, M., Yang, L. & Chai, R. Research on Multi-Scale Fusion Method for Ancient Bronze Ware X-ray Images in NSST Domain. *Appl. Sci.* **14**, 4166 (2024).
28. MLGTM. Multi-Scale Local Geometric Transformer-Mamba Application in Terracotta Warriors Point Cloud Classification. *Remote Sens.* **16**, 3086 (2024).
29. Hou, Q., Zhou, D. & Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 13713–13722 (2021).
30. Khalilian, P. et al. Jewelry rock discrimination as interpretable data using laser-induced breakdown spectroscopy and a convolutional LSTM deep learning algorithm. *Sci. Rep.* **14**, 5169 (2024).
31. Singh, V. & Kaewprapha, P. A comparative experiment in classifying jewelry images using convolutional neural networks. *Sci. Technol. Asia*. **23**, 7–17 (2018).
32. Yan Chow, B. H. Carlos Reyes-Aldasoro, C. Automatic gemstone classification using computer vision. *Minerals* **12**, 60 (2022).
33. Alex, V., KP, M. S., Chennamsetty, S. S. & Krishnamurthi, G. Generative adversarial networks for brain lesion detection. In Medical Imaging 2017: Image Processing, Vol. 10133, 101330G (2017).
34. Mohamad Saufi, N. et al. Element identification of different gemstones by using LIBS-PCA method. In Proceedings of the 7th International Conference on the Applications of Science and Mathematics 155–163 (2021).

35. Han, X. An improved attention guided convolutional neural network and transformer hybrid model for emotion classification in traditional Chinese paintings. *Sci. Rep.* **16**, 23240 (2026).
36. Lin, T. Y. et al. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision 2980–2988 (2017).

Figures

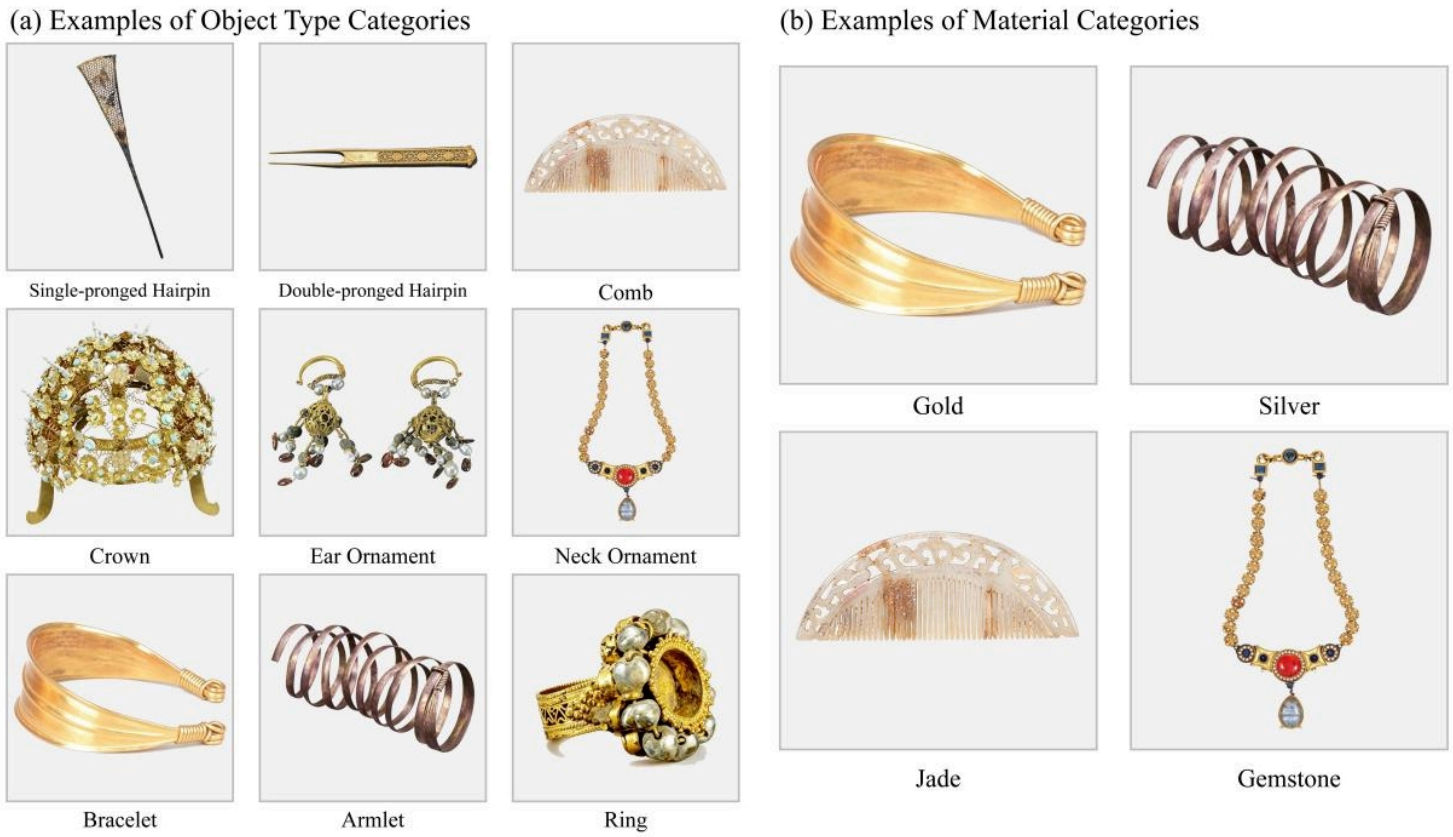


Figure 1

Overview of sample examples from the ancient Chinese jewelry dataset: (a) Examples of object type categories; (b) Examples of material categories.

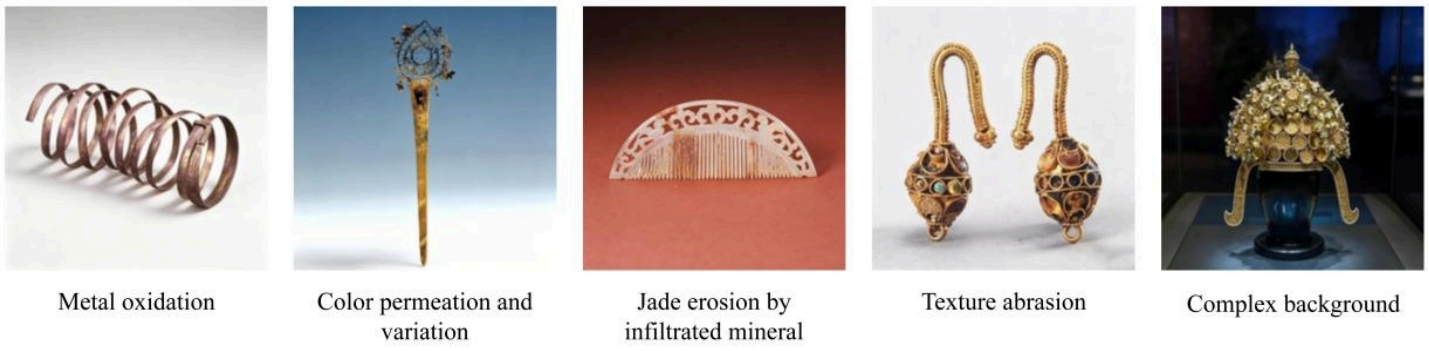


Figure 2

Examples of typical confusable samples.

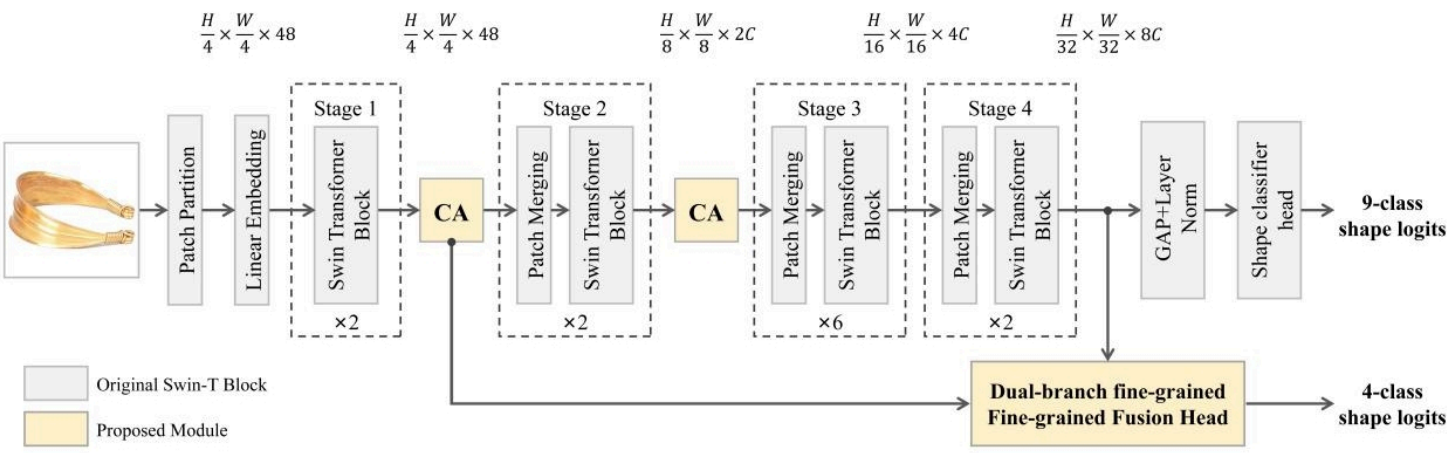


Figure 3

Overall architecture.

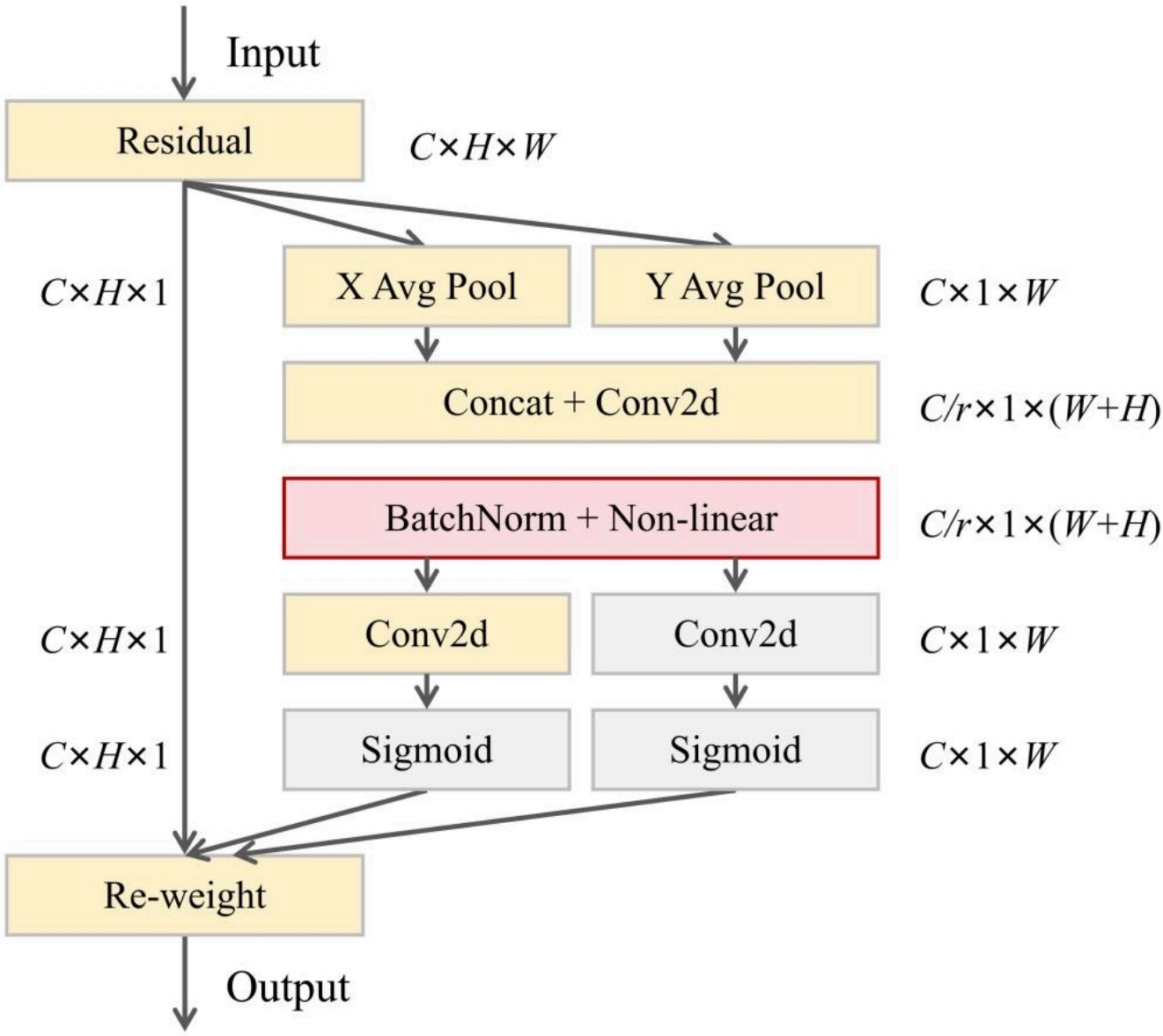


Figure 4

Structure of the coordinate attention module.

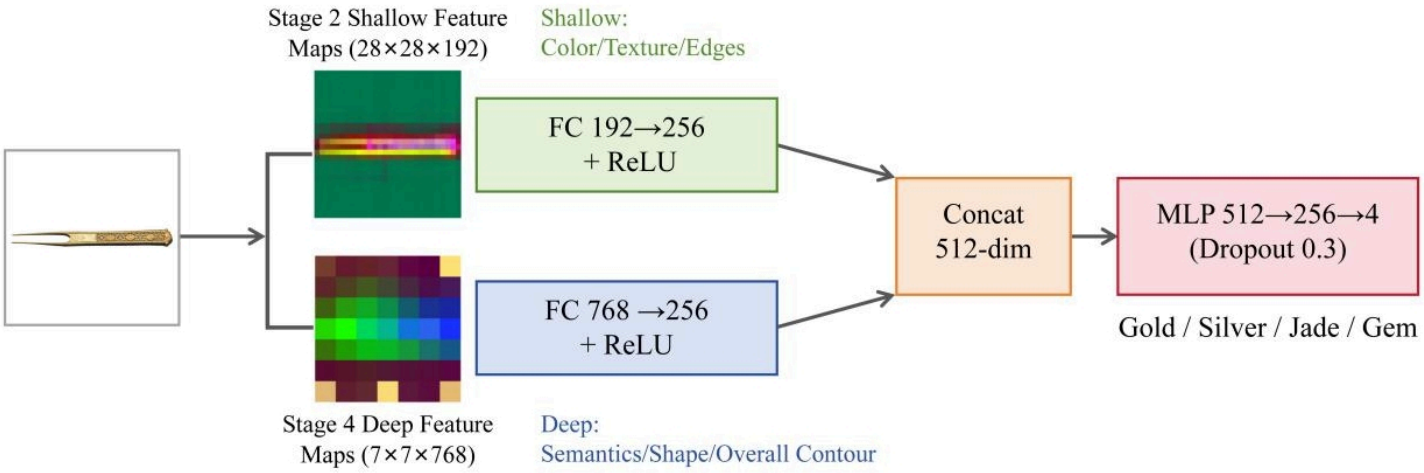


Figure 5

Material dual-branch fine-grained fusion head.

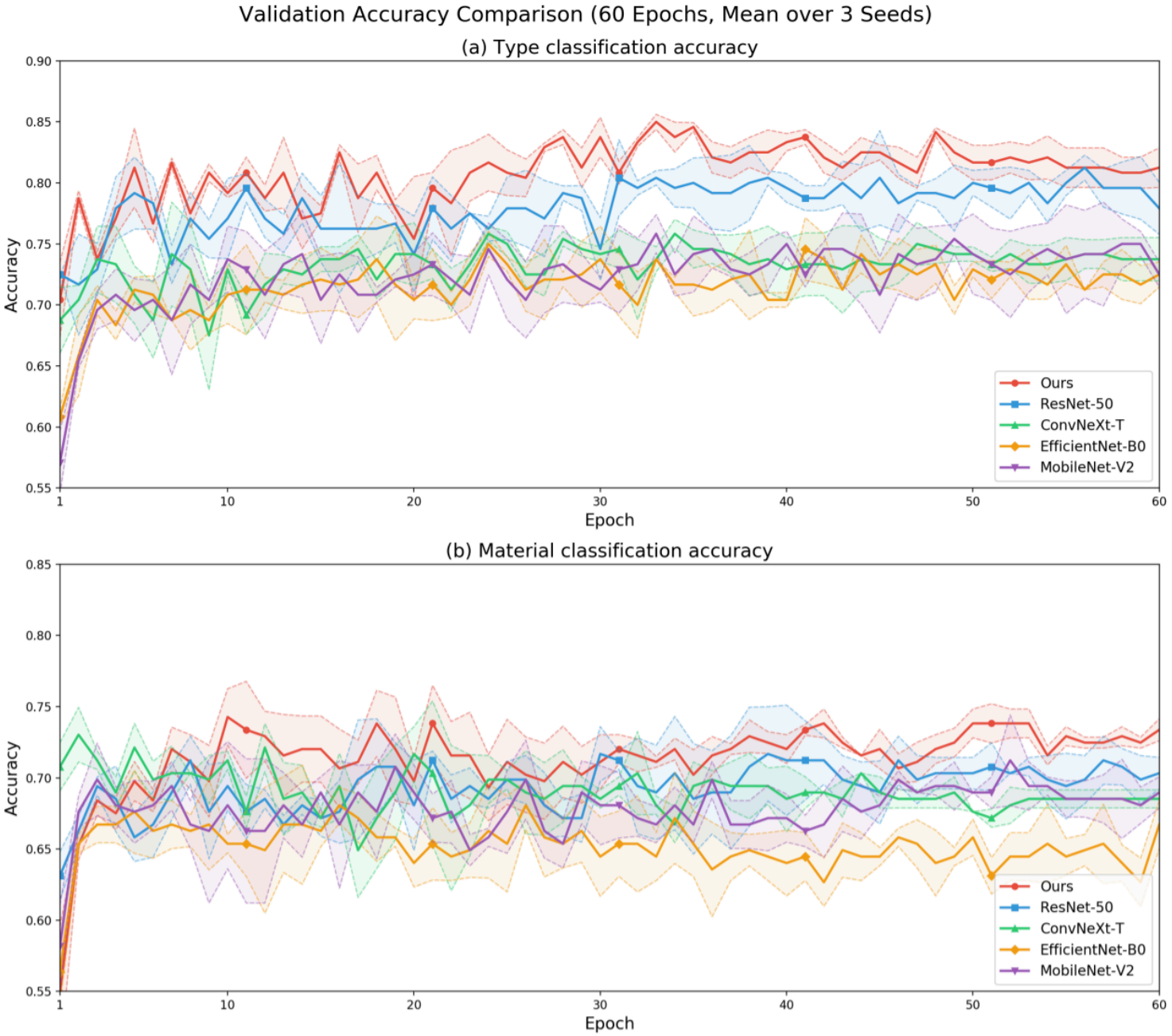


Figure 6

Validation accuracy curves of different models during training on the ancient Chinese jewelry dataset: (a) Type classification accuracy; (b) Material classification accuracy.

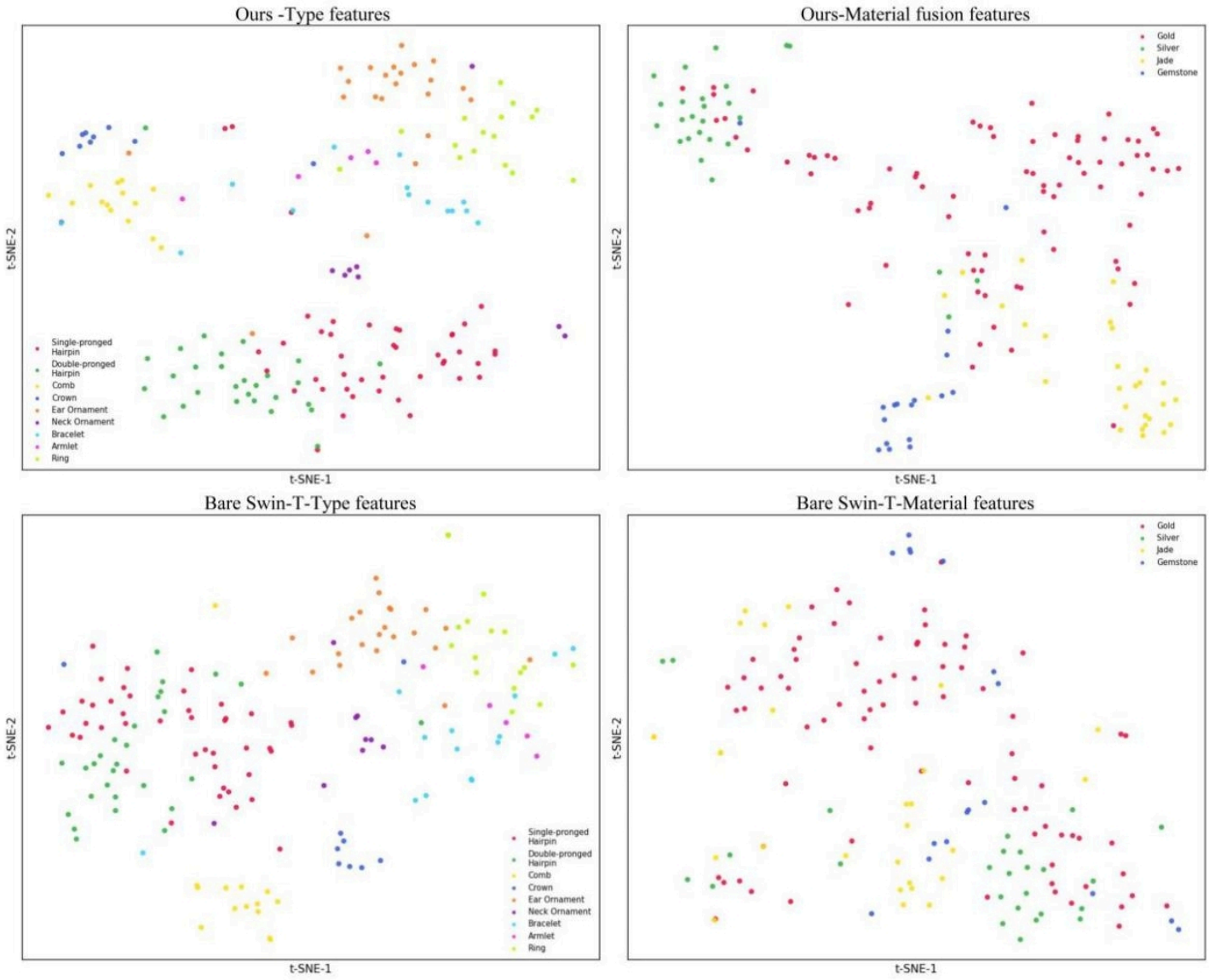


Figure 7

t-SNE comparison of feature distributions between the proposed method and the Swin-T baseline.

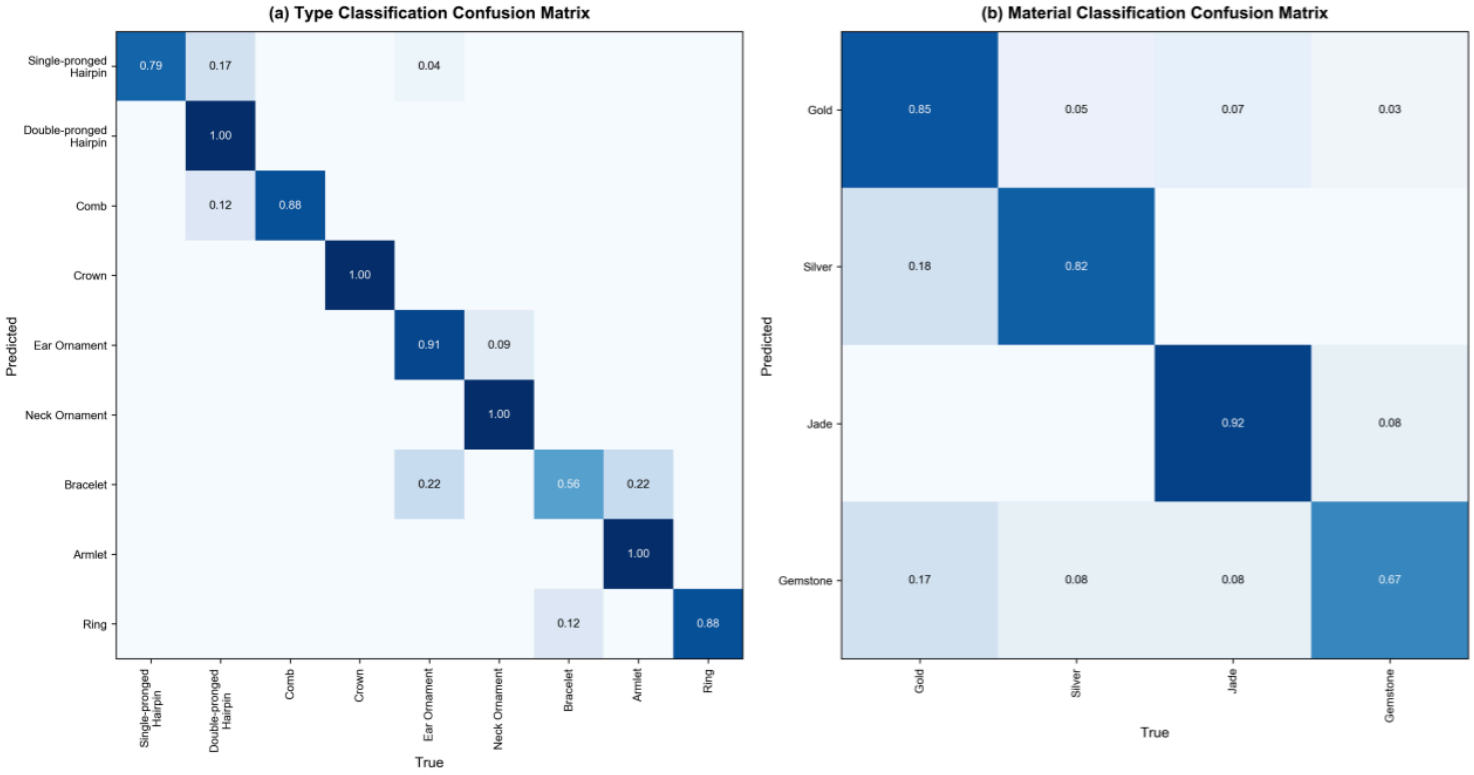


Figure 8

Confusion matrices: (a) Type classification Confusion Matrix; (b) Material Classification Confusion Matrix.

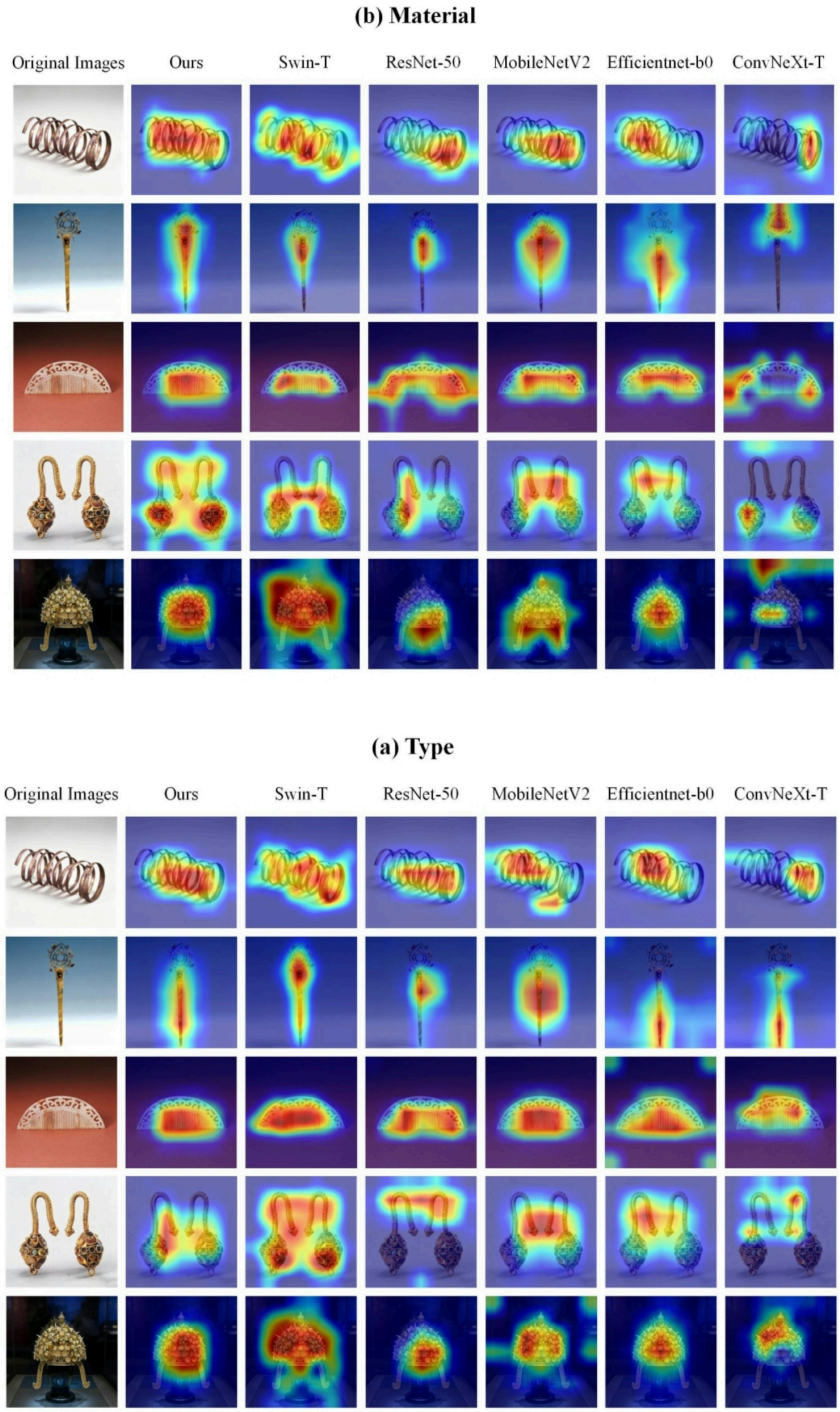


Figure 9

Grad-CAM visualization comparison of different models on the ancient Chinese jewelry classification task: (a) Type classification; (b) material classification.