ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Comparison and classification of dunhuang murals and medieval European musical instrument images from a cross-cultural perspective

Haixiang Gao[1,*]

[1] School of Contemporary Music and Technology, Nanjing University of Arts, NanJing, 210000, China

SUMMARY: *This paper proposes a cross-cultural classification method for images of Dunhuang murals and medieval European Musical Instruments. A dataset of 9216 images covering 16 musical instrument categories was constructed, with 4768 Dunhuang samples and 4448 European samples. The framework uses region clipping, label normalization and block-level coding, introduces hierarchical visual semantic alignment and adaptive gated fusion, and maps heterogeneous cultural cues into the discriminant space. A two-branch classifier is designed to extract the organ-shaped commonalities while preserving the details. The focus classification loss, contras alignment loss and domain constraint regularization term are jointly optimized in training to stabilize class boundaries. Experimental results show that the proposed method achieves 94.8% Accuracy, 93.6% Macro-F1 and 92.4% CDR, which are better than the ResNet50, Swin-T and CLIP-Linear baselines. The test results show that the misjudgment proportion of harp and harp is less than 6.3%, and the confusion rate of harp and harp is reduced to 8.1%, which indicates that the model can provide computational support for cross-cultural instrument analysis, cataloguing and knowledge organization.*

KEYWORDS: *Dunhuang mural; Medieval European musical instrument images; Cross-cultural image classification; Visual semantic alignment*

1 Introduction

Dunhuang murals preserve continuous image evidence of the shape, playing posture, and ceremonial scenes of Chinese Musical Instruments in the Middle Ages, while medieval European religious paintings, manuscript illustrations, and architectural decorations record another tradition of musical instrument images. There are significant differences between the two types of images in composition, color, scale and context, while retaining the common structural clues of string music, wind music and percussion music. Putting the two types of images into the unified computing framework for comparative classification can not only support knowledge collation in digital humanities, but also provide a computational basis for cross-cultural image retrieval, artifact image annotation and cultural heritage database organization. With the development of high-resolution acquisition, image slice annotation and visual coding technology, cultural heritage image analysis is shifting from manual description to deep feature learning and fine-grained discrimination.

Focusing on the intelligent analysis of cultural heritage, Yu T et al. constructed the Dunhuang cultural heritage protection project and dataset, which provided a basic sample organization method for subsequent computer vision research [1]. Zhou Z et al. proposed a

*ghx30599@163.com
https://doi.org/10.65102/is2026089 Published: 30 April, 2026

structure-guided deep network for digital repair of Dunhuang murals [2]. Madhu P et al. applied perceptual constrained style transfer to human pose estimation in ancient pottery vase paintings [3]. Jiang H and Yang T studied the method of drawing style feature extraction based on convolutional neural network [4]. Zhong Y and Huang X proposed an improved CNN painting style recognition system [5]. Liu S proposed an art painting image classification method based on naive Bayes [6]. Tan Y uses lightweight deep learning for painting image feature recognition and style transfer [7]. Dinesh Kumar R et al. used deep convolutional networks to learn categorical representations in artistic style transfer [8]. Cascone L et al. carried out the artistic style recognition of fragment images [9]. Fan T et al. proposed an intangible cultural heritage image classification method based on multi-modal attention and hierarchical fusion [10].

In terms of cognitive modeling of traditional art images, Geng J et al. proposed a multi-channel color fusion network for cognitive classification of traditional Chinese paintings [11]. Croce V et al. implemented the semi-automatic classification of digital heritage on an open 2D and 3D labeling platform [12]. Valencia J et al. summarized the research trend and agenda of machine learning for predicting artistic style [13]. Sha S et al. applied convolutional networks to classification and restoration of ancient fabric images [14]. Cheng J et al. proposed a painting style and emotion recognition method combining multi-feature fusion and style transfer [15]. Li H and Zhu W studied artistic image style transfer based on multi-scale feature fusion network [16]. Schaerf L et al. applied vision transformers to identify the authenticity of artworks [17]. Zhang X proposed ResNet-NTS oil painting style recognition model [18]. Liu Y et al. used convolutional transformers for art image recognition [19]. Xiang J et al. constructed a multi-scale convolutional network for adaptive classification of artistic images [20]. Existing research has formed a technical chain from convolutional representation, attention fusion to Transformer discrimination, but there is still a lack of a unified semantic alignment path for the cross-cultural comparative classification of Dunhuang murals and medieval European musical instrument images.

From the perspective of method evolution, existing art image classification models have been able to extract texture, color, layout and local shape features. However, cross-cultural musical instrument images still face three differences at the computational level. Second, the image naming of the same instrument in the two cultures does not completely coincide with the visual prototype. Thirdly, the simple feature splicing is easy to mix the cultural style signal with the general signal of the device category, and weaken the boundary of the subdivision category. Existing research shows that hierarchical heterogeneous modal alignment and adaptive gated fusion mechanism can provide a clear method skeleton for cross-cultural image classification. Based on this idea, we further translate it into a hierarchical visual semantic alignment structure for instrument image classification, which is used to maintain the similarities of instrument shapes and cultural image differences at the same time. In order to more clearly illustrate the connection between the existing research and the proposed method, this paper summarizes the related work from four aspects: sample source, classification skeleton, cross-feature collaborative modeling and cross-domain discrimination, as shown in Table 1.

Table 1: Technical summary of related studies

Research Direction	Representative References	Main Methods	Relevance to This Study
Dunhuang Digital Preservation	[1]–[2]	Dataset construction, structure-guided networks	Provides the sample source and the heritage-scene foundation
Artistic Style Recognition	[4]–[9], [18]–[20]	CNN, Bayesian methods, Transformer	Provides the classification backbone and experience in fine-grained recognition
Multimodal Heritage Classification	[10]–[12], [15]–[16]	Hierarchical fusion, color fusion, style transfer	Provides ideas for cross-feature collaborative modeling
Trustworthy Visual Discrimination	[13], [17]	Review analysis, ViT-based discrimination	Provides references for cross-domain discrimination and high-level semantic modeling

Based on the above research vein, this paper constructs a comparative classification framework for cross-cultural musical instrument images, and completes category unification, region cropping and visual coding on Dunhuang mural samples and medieval European image samples. Furthermore, a hierarchical visual semantic alignment and adaptive gated fusion model are introduced to learn the correspondence between class structure, performance posture and cultural style in the shared feature space. And the fine-grained classification boundary is strengthened through the discrimination mechanism under the constraint of cross-cultural differences. This study extends the image classification task from a single cultural domain to a heterogeneous cultural domain, which helps to forma more stable computational representation of musical instrument images, and also provides a new implementation path for computer vision, pattern recognition and digital cultural heritage analysis focused on by Informatica. Compared with the scheme that only relies on single-scale texture statistics, this method emphasizes more on establishing a stable mapping between local shape, global composition and semantic labels, so that the classification results can be used not only for automatic annotation, but also for subsequent knowledge graph association and cross-collection retrieval.

2 Methods and materials

2.1 Image sample construction and visual feature extraction of Dunhuang murals and medieval European Musical Instruments

There are obvious differences between Dunhuang murals and medieval European musical instrument images in color system, instrument scale, description method and preservation state. The sample construction cannot stay at the image collection level, but needs to organize source screening, category unification, region cropping, quality control and visual coding into continuous processing links. In order to ensure the consistency of the input, the sample sources are limited to five categories: high-definition mural local images, digital copy images, European religious painting local images, manuscript illustrations and architectural decoration images. Each image is annotated with the musical instrument entity as the center, and the samples with severe occlusion, subject missing and semantic ambiguity are removed, and the cross-cultural category mapping table is established under the premise of retaining the original cultural

context information. The category layer adopts a three-level organization method of string sound, qi sound and stroke sound, and the instance layer is refined into 16 sub-categories such as pipa class, Konghou class, Ruan Xian class, lute class, harp class, flute class, horn class and drum class, so that the isomorphic images in different cultures can enter a unified label space.

As shown in Fig. 1, the sample construction process consists of five links: original image input, musical instrument region annotation, cross-cultural category mapping, quality screening, and dataset division. The original image sources include high-definition images of Dunhuang cave murals, digital copy images of Dunhuang, local images of European religious paintings, illustrations of medieval manuscripts, and local images of architectural decorations. After the source images are imported into the original image database, they are first processed by manual correction and weakly supervised detection to form a unified annotation box output. Then they are entered into the cross-cultural category mapping module, and the instrument images in different cultural contexts are merged into string sound, chi sound, stroke sound and their 16 subcategories. After completing the category mapping, the system performs quality screening according to completeness, clarity, contour discrimination and occlusion rate, retains the samples that meet the requirements, and divides them into training set, validation set and test set after normalization.

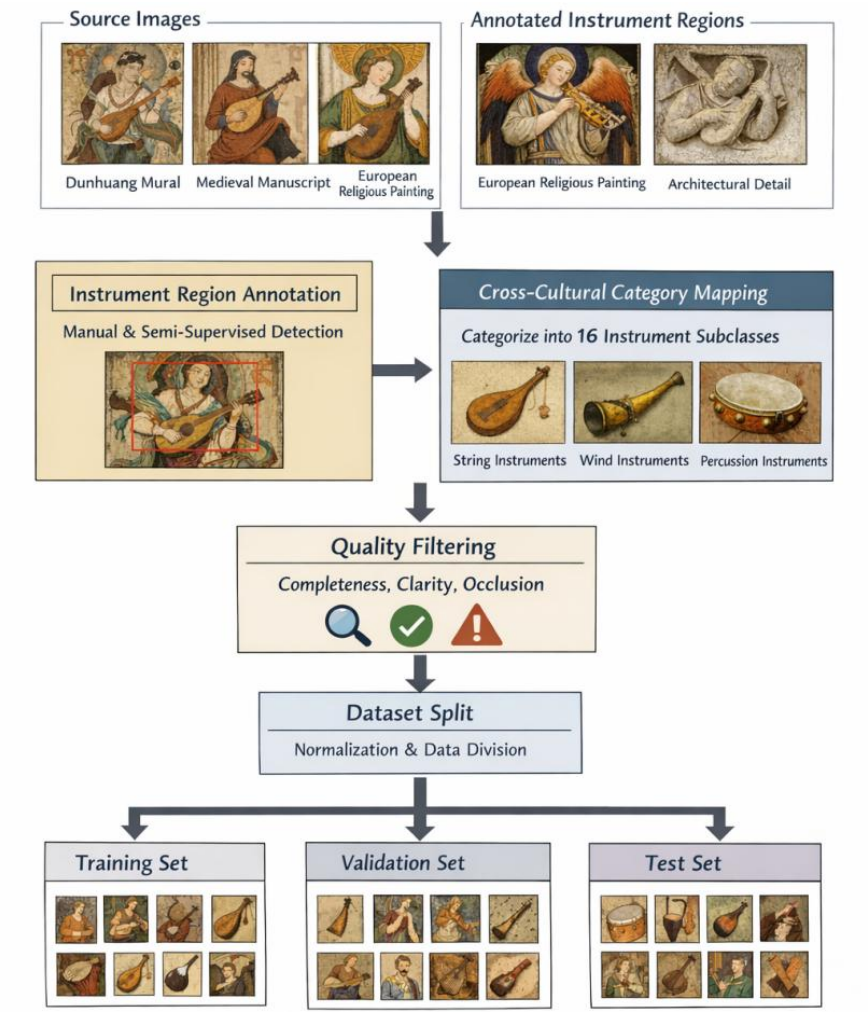

Figure 1: Sample construction and screening process

In order to avoid these factors interfering with feature learning, samples should be scored for completeness, clarity and contour discrimination before entering the database. The cropped

region uses normalized mapping to generate fixed-size input, and the transformation is written as follows.

$$\tilde{p}_{u,v} = \begin{bmatrix} \frac{u - x_{min}}{x_{max} - x_{min}} \\ \frac{v - y_{min}}{y_{max} - y_{min}} \end{bmatrix}, \quad I^c(i,j) = I(x_{min} + i\Delta_x, y_{min} + j\Delta_y) \tag{1}$$

where $(x_{min}, y_{min}, x_{max}, y_{max})$ represents the coordinates of the labeled box, $\tilde{p}_{u,v}$ represents the normalized pixel position, I^c represents the trimmed instrument image, and Δ_x and Δ_y represent the sampling step size. The function of this formula is to project the main body of the instrument in different resolutions and different compositions into a consistent scale, so that the key parts such as the body, string, mouthpiece and hitting surface are stably retained.

After cropping, the image is divided into fixed-size visual blocks and fed into a precoder to generate a block-level representation. The block embedding process is written as follows.

$$z_n^0 = W_p\, vec(I_n^c) + e_n, \quad H^0 = [z_1^0; z_2^0; \cdots; z_N^0; z_{cls}] \tag{2}$$

where I_n^c represents the n image block, W_p is the projection matrix, e_n is the position encoding, z_{cls} is the global labeling, and H^0 is the initial sequence representation. This formula is used to transcribe local texture, contour boundary and component distribution into visual tokens of uniform dimension.

It is difficult to completely cover the color attenuation and structural deformation in artistic images by only relying on block embedding. Therefore, three kinds of features are extracted in parallel: color statistics, edge response and contour topology. The color statistics vector is defined as follows.

$$c = [\mu_R, \mu_G, \mu_B, \sigma_R, \sigma_G, \sigma_B, \rho_{RG}, \rho_{RB}, \rho_{GB}]^\top, \quad \rho_{ab} = \frac{1}{|\Omega|} \sum_{(u,v)\in\Omega} \frac{(I_a - \mu_a)(I_b - \mu_b)}{\sigma_a \sigma_b} \tag{3}$$

Here, μ represents the channel mean, σ represents the channel standard deviation, ρ represents the channel correlation coefficient, and Ω is the set of pixels in the instrument area. This formula not only retains the color intensity, but also describes the coupling relationship between channels, so that the material, light and shade and color mode can be entered into the calculation expression.

The edge branch establishes the direction response around the body contour and component direction, which is calculated as follows.

$$g_k = \sum_{(u,v)\in\Omega} \omega_{u,v} \|\nabla I^c(u,v)\|_2 \exp\left(-\frac{(\theta(u,v) - \phi_k)^2}{2\tau^2}\right), \quad g = [g_1, g_2, \ldots, g_K]^\top \tag{4}$$

Here, $\omega_{u,v}$ represents the position weight, $\|\nabla I^c(u,v)\|_2$ is the gradient magnitude, $\theta(u,v)$ is the edge orientation Angle, ϕ_k is the orientation kernel center, and τ is the orientation bandwidth. This formula is used to compress the parallel relation of strings, the direction of mouthpiece arrangement and the closed boundary of drum frame into the direction spectrum description.

The contour topology branch further maps the organic-shaped connection relationship into an adjacency graph, which can be calculated as follows.

$$A_{mn} = \exp\left(-\frac{\|p_m - p_n\|_2^2}{\eta^2}\right) 1(d_{geo}(m,n) < \delta), \quad t_m = \sum_{n=1}^{M} A_{mn}\, r_n \tag{5}$$

where p_m and p_n represent the skeleton key point coordinates, $d_{geo}(m,n)$ represents the geodesic distance, δ is the connection threshold, η is the distance attenuation coefficient, and r_n is the local topological response. This formula preserves the spatial relationship between the body arc, the resonator and the support rod, so that the similar categories have a clearer separation in the detail structure.

As shown in Fig. 2, the visual feature extraction framework adopts the combination of multi-branch parallel and unified convergence. After entering the system, the normalized musical instrument image is fed into Patch Embedding, color statistics branch, edge response branch and contour topology branch simultaneously. Patch Embedding is used to generate block-level sequence features, the color statistics branch outputs the color feature vector, the edge response branch forms the direction spectrum feature, and the contour topology branch generates the adjacent structure feature. The four types of features are then jointly modeled in the local-global interaction module and projected into a unified visual representation space. In order to reduce the disturbance caused by brightness shift, blur, occlusion rate and contour integrity difference, the unified representation is further re-calibrated by quality perception, and the stable visual feature representation is output.

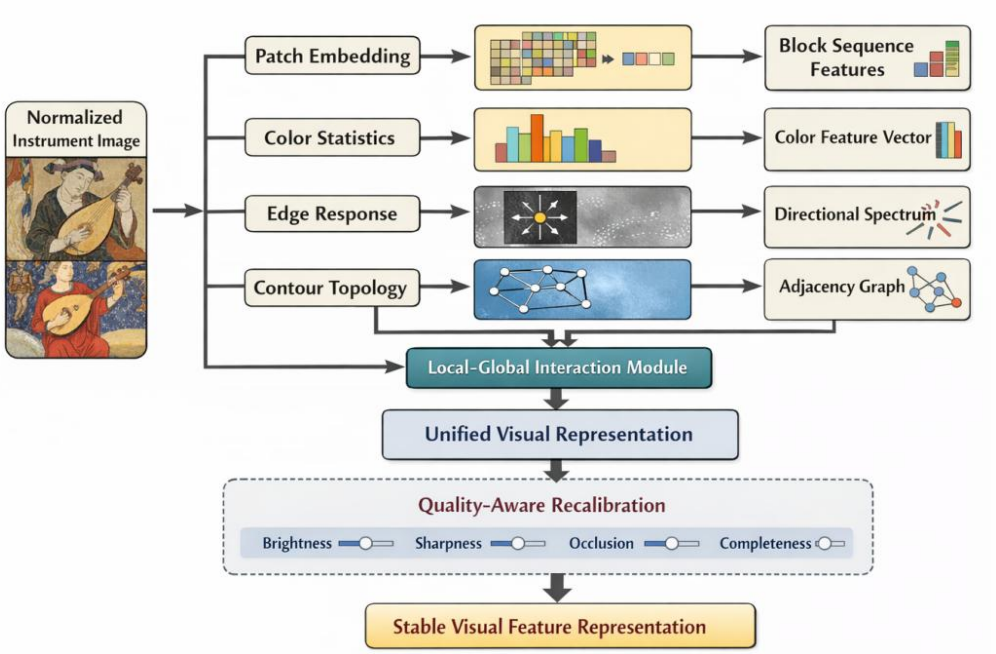

Figure 2: Framework for visual feature extraction

The update process between the local block and the global tag is written as follows.

$$h_i' = \mathrm{Softmax}\left(\frac{Q_i K_g^{\top}}{\sqrt{d}} + B_{rel}\right) V_g + h_i \tag{6}$$

Here, h_i' represents the ith local block feature, Q_i, K_g and V_g represent the query, key and value mapping results, respectively,B_{rel} is the relative position bias, and d is the feature

dimension. This formula enables the local texture to be re-updated under the constraints of the global performance scene, and avoids the isolated local dominant category judgment.

Multi-source visual signals are uniformly projected into a shared representation space, which is calculated as follows.

$$s_i = \mathrm{LN}\left(W_h h_i' + W_c c + W_g g + W_t t_i + b_s\right) \tag{7}$$

where W_h,W_c,W_g,W_t are the projection matrices, b_s is the bias term, $\mathrm{LN}(\cdot)$represents the layer normalization, and s_i represents the unified visual representation. This formula compresses color, edge, topology and block-level context into the same coordinate system, so that different source images have a comparable basic representation.

Considering the differences in clarity, occlusion rate and pigment integrity between mural samples and painting samples, a quality-aware weight is further constructed to recalibe the unified representation dimension by dimension:

$$q_i = \sigma\left(W_q[l_i;\, b_i;\, o_i;\, r_i] + b_q\right), \quad \tilde{s}_i = q_i \odot s_i \tag{8}$$

where l_i represents brightness shift, b_i represents ambiguity, o_i represents occlusion rate, r_i represents contour integrity, W_q and b_q are quality mapping parameters, $\sigma(\cdot)$ is Sigmoid function, $\odot$ represents element-wise multiplication. This formula can suppress the invalid fluctuations introduced by low-quality regions, so that the final visual representation points more stably to the instrument-shaped structure, the component combination, and the performance feature itself.

The samples were also processed for cross-cultural distribution balance after storage. The proportions of sitting performance, lateral holding and group image embedding in Dunhuang samples are high, and the proportions of standing performance, frontal display and single display in European samples are high. If trained directly, the model is easy to regard posture and scene as the main signals of the category. To this end, only four types of operations such as rotation, brightness perturbation, partial occlusion simulation and scale dithering are retained in data augmentation, and strong geometric distortions that will change the shape topology are not introduced. The enhanced samples still retain the original source labels, era labels and scene labels, but these labels are only used for stratified sampling and batch equalization before training, and do not enter the visual encoding link, so as to ensure that the feature extraction stage always centers around the image subject itself.

In the storage organization, each sample consists of five parts: original image, crop image, edge image, skeleton image and annotation file, which are uniformly stored as traceable sample units. In addition to the category number, auxiliary attributes such as body length-width ratio, string salience, mouthpiece visibility and hitting surface integrity are recorded in the annotation file for quality statistics and abnormal review. In this way, the organized samples can not only support batch training, but also facilitate the tracing back to specific source images and specific structural parts during error analysis. After sample construction, region cropping, multi-branch feature extraction, unified projection and quality recalibuation, the input data is transformed into a stable, comparable visual representation with fine-grained structural information. The representation preserves the local component texture, the global shape contour and the scene constraint relationship at the same time, so that the same kind of foreign images and different kind of near-shape images can obtain a clearer basis for distinguishing in the computational space, and also provides support for subsequent retrieval and archiving.

2.2 Cross-cultural Contrastive classification Model Based on hierarchical visual semantic Alignment and adaptive gated fusion

After sample construction and visual feature extraction, block-level textures, color statistics, orientation spectra, and topological structures have been organized into the same batch of input, but these representations still carry obvious cultural domain differences. The images of Musical Instruments in Dunhuang murals are characterized by flat coloring, linear contours and partial damage, while the images of medieval Europe emphasize more on the level of light and dark, volume relationship and scene perspective. If the multi-channel features are directly processed by single-layer splicing or simple weighting, the model is easy to miswrite the cultural style difference as the category boundary, resulting in the distribution of the same kind of Musical Instruments in different cultural domains is too scattered. Based on this, this paper constructs a hierarchical visual semantic alignment and adaptive gated fusion model, which divides the cross-cultural comparative classification process into five steps: modal projection, hierarchical region alignment, category prototype aggregation, cross-domain semantic interaction and dynamic gated fusion, so that the common structure of similar Musical Instruments forms a stable aggregation in the shared space. At the same time, the discriminative distance between different classifier classes is maintained.

As shown in Fig. 3, the model adopts a sequential structure from left to right. The leftmost input simultaneously receives block-level visual features, color statistical vectors, directional spectrum features, and topological structure features. The modal projection layer is responsible for dimension unification and initial semantic compression, the hierarchical region alignment layer further corrects the source offset between local regions, and the category prototype aggregation layer extracts the common center of the musical instrument category from the aligned region representation. The semantic memory of Dunhuang domain and the semantic memory of European domain are introduced in the cross-cultural semantic interaction layer, so that the two types of cultural images can exchange the shape related information under the premise of retaining their own style boundaries. The adaptive gated fusion layer on the far right does not average the multi-channel results, but dynamically generates the fusion proportion according to the quality weight, domain difference weight and structure consistency weight, and finally outputs a stable representation vector that can be used for classification decision.

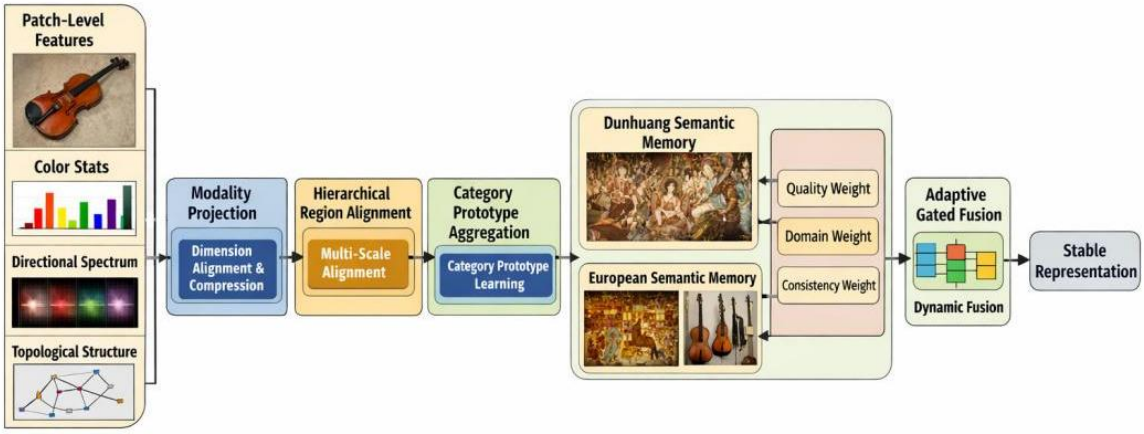

Figure 3: Cross-cultural contrastive classification model with hierarchical visual semantic alignment and adaptive gated fusion

To eliminate the inconsistency of multi-channel features in dimension and scale, the modal projection layer first maps four types of inputs into a unified semantic space, which is calculated

as follows.

$$z_i = \mathrm{LN}\left(W_h h_i + W_c c_i + W_g g_i + W_t t_i + b_p\right) \tag{9}$$

Here, h_i represents block-level visual features, c_i represents color vectors, g_i represents directional spectrum features, t_i represents topological structure features, b_p is the bias term, and z_i is the initial representation under a unified dimension. The function of Equation (9) is to compress visual signals with different sources and granularities into the same representation coordinate system, which provides a stable entry for region-level alignment.

After obtaining the initial representation, the model rewrites the context relationship between local blocks using a hierarchical region alignment mechanism, which is computed as follows.

$$\hat{z}_i = \sum_{j=1}^{N} \mathrm{Softmax}\left(\frac{Q_i K_j^{\top}}{\sqrt{d}} + B_{ij}\right) V_j + z_i \tag{10}$$

Here, Q_i,K_j,V_j represent the query vector, key vector, and value vector respectively, B_{ij} represents the relative position bias, d represents the feature dimension, and $\hat{z}_i$ represents the representation after region alignment is completed. Equation (10) puts the local structures such as string train, resonator, mouthpiece and drum frame back into the context of the overall region, so that the local contour is no longer separated from the performance posture and scene relationship and judged separately.

As shown in Table 2, the model is processed by layer-by-layer mapping and layer-by-layer updating. The modal projection layer is responsible for unifying the multi-channel feature dimensions, the region alignment layer is responsible for compressing the local source offset, the prototype aggregation layer is responsible for extracting the class common center, the semantic interaction layer is responsible for establishing the dual-domain correspondence, and the gated fusion layer is responsible for generating the stable final representation. Such a hierarchical arrangement makes the feature flow clearer and the classification representation more stable.

Table 2: How the model hierarchies are organized

Module	Input	Output	Function
Modality Projection Layer	Modality Projection Layer	Initial representation(z)	Unifies dimensionality and scale
Region Alignment Layer	Initial representation(z)	Region representation($\hat{z}$)	Corrects local source shift
Prototype Aggregation Layer	Region representation($\hat{z}$)	Category prototype(p)	Compresses shared class characteristics
Semantic Interaction Layer	Semantic Interaction Layer	Interaction representation(u)	Establishes cross-domain configurational relations
Gated Fusion Layer	Gated Fusion Layer	Fused representation(f)	Performs dynamic recalibration and stabilizes output

In order to extract class commonality centers from region representations, the prototype aggregation layer uses attention allocation to form category prototypes, and the calculation

process is written as follows.

$$\alpha_{ik} = \frac{\exp(\hat{z}_i^{\top} \mu_k)}{\sum_{r=1}^{K} \exp(\hat{z}_i^{\top} \mu_r)}, \quad p_k = \sum_{i=1}^{N} \alpha_{ik} \hat{z}_i \tag{11}$$

Here, μ_k represents the k learnable prototype center, α_{ik} represents the weight assigned by the regional features to this prototype, and p_k represents the aggregated category prototype. Equation (11) can compress locally scattered instrument shape information into stable category centers, so that similar Musical Instruments can form comparable structural anchors in different cultural images.

The cross-cultural semantic interaction layer exchanges prototype information under the constraint of dual-domain semantic memory, which is calculated as follows.

$$\beta_{kl} = \frac{\exp\left((W_D p_k^D)^{\top} (W_E p_l^E)\right)}{\sum_r \exp((W_D p_k^D)^{\top} (W_E p_r^E))}, \quad u_k^D = \sum_{l=1}^{K} \beta_{kl} \left(W_D p_k^D + W_E p_l^E + m^D + m^E\right) \tag{12}$$

Here, p_k^D and p_l^E represent the Dunhuang domain and European domain prototypes, m^D and m^E represent the semantic memory of the two domains, and β_{kl} represents the cross-domain matching weight u_k^D represents the updated interaction result. Equation (12) enables similar Musical Instruments to complete cross-domain proximity in the shared space, while retaining boundary information brought by cultural style differences.

To explicitly incorporate the differences between cultural domains into the model, the system further constructs the domain difference vector, which is in the form of:

$$d_i = \tanh\left(W_\Delta \left(\bar{p}_{y_i}^D - \bar{p}_{y_i}^E\right) + b_\Delta\right) \tag{13}$$

where $\bar{p}_{y_i}^D$ and $\bar{p}_{y_i}^E$ represent the mean prototype of the category to which the sample belongs in the two domains, W_Δ is the difference mapping matrix, b_Δ is the bias term, d_i represents the domain difference vector. Equation (13) transforms cultural style differences into learnable representations, so that the model will not smooth out the actual differences in visual organization of images in the two domains when aggregating similar Musical Instruments.

After completing the difference modeling, the adaptive gated fusion layer calculates the dynamic weight according to the mass vector, domain difference vector and structure consistency vector, which is written as follows.

$$\gamma_i = \sigma\left(W_q q_i + W_d d_i + W_s s_i + b_\gamma\right) \tag{14}$$

Here, q_i represents the mass vector, d_i represents the domain difference vector, s_i represents the structural consistency vector, W_q,W_d,W_s are mapping matrices, b_γ is the bias term, and γ_i represents the gating weight. Equation (14) can dynamically determine the retention ratio in the feature dimension, so that visual information with clear, complete and stable configuration can obtain higher weight.

After the gating weights are generated, the multiway results are rewritten into a unified fusion representation, which is calculated as follows.

$$f_i = \gamma_i \odot u_i + (1 - \gamma_i) \odot (\eta_1 \hat{z}_i + \eta_2 p_i) \tag{15}$$

Here, u_i represents the cross-cultural interaction results, $\hat{z}_i$ represents the region

alignment results, p_i represents the prototype features,η_1,η_2 are the balance coefficients, and f_i is the final fusion representation. Equation (15) is not a simple sum, but dynamically adjusts the proportion between local information, common center and cross-domain interaction results along the feature dimension.

In order to ensure the compact distribution of similar instruments in the shared space and maintain the discrimination distance of different instruments, the hierarchical consistency constraint is added to the representation layer:

$$L_h = \sum_{i,j} y_{ij} \|f_i - f_j\|_2^2 + \sum_{i,j} (1 - y_{ij}) \left[m - \|f_i - f_j\|_2\right]_+^2 \tag{16}$$

Here, y_{ij} denotes the homogeneous or heterogeneous label, m denotes the interval threshold, and L_h denotes the hierarchical consistency loss. Equation (16) makes the representation of the same class samples more compact and the representation of the different class samples remain sufficiently separated, thus stabilizing the class boundary in the representation space.

In the model training phase, the alignment term, interaction term, consistency term and gating regularization term are combined as a unified optimization objective, which is in the following form:

$$L_{fuse} = \lambda_a \sum_i \|\hat{z}_i - p_{y_i}\|_2^2 + \lambda_c \sum_i \|u_i - p_{y_i}\|_2^2 + \lambda_h L_h + \lambda_g \|\Gamma\|_1 \tag{17}$$

Here,λ_a,λ_c,λ_h,λ_g are the loss weight coefficients, Γ represents the set of sector-wide control vectors, and L_{fuse} represents the joint objective in the fusion stage. Equation (17) constrains the model to simultaneously take into account local alignment, cross-domain interaction and structural compactness during the training process to avoid a certain branch dominating the overall representation direction during the optimization process.

After the above calculation process, the fusion representation output by the model simultaneously contains three types of information: the first is the common configuration of similar Musical Instruments in different cultural domains, the second is the cross-cultural relationship cues provided by the dual-domain semantic memory, and the third is the quality adaptive weight generated by the dynamic gating. This representation neither directly presses the plane contour in Dunhuang murals onto the volume representation of European images, nor miswrites the light and dark structure in European images as new musical instrument category features. The final representation space is consistent in three aspects: intra-class aggregation, inter-class separation and cross-domain stability, so as to provide a clear structure, error controllable and interpretable computational basis for the comparative classification of cross-cultural instrument images.

2.3 Instrument category discrimination and decision-making mechanism under the constraints of cross-cultural differences

The mechanism continues to complete category discrimination, discrepancy constraint and confidence decision after the fusion representation is formed. Although Dunhuang murals and similar Musical Instruments in medieval European images share the basis of instrument shape, they still have stable differences in color setting, composition position and local decoration. If the class probability is only output by the linear classification head, the model is easy to directly map the cultural style deviation to the class boundary, which leads to the overlap between the

near-shape heterogeneous class and the same foreign domain samples near the decision surface. Therefore, this paper organizes the discrimination process into four continuous steps: structure prototype matching, domain difference suppression, confidence calibration and joint decision making, so that the final classification result satisfies the classifier class consistency, cultural separability and output stability at the same time.

As shown in Fig. 4, after the fusion representation vector enters the discrimination module, it is not directly fed into a single fully connected layer, but first performs similarity matching with the category prototype library, and then calculates the difference offset according to the domain statistical center. Subsequently, the system uses an independent confidence estimation layer to generate sample confidence, and jointly maps the confidence and category scores into the final decision space. In this way, the classification score and sample reliability are modeled separately, which can reduce the disturbance of complex background, local defects and foreign near-shape samples on the final output.

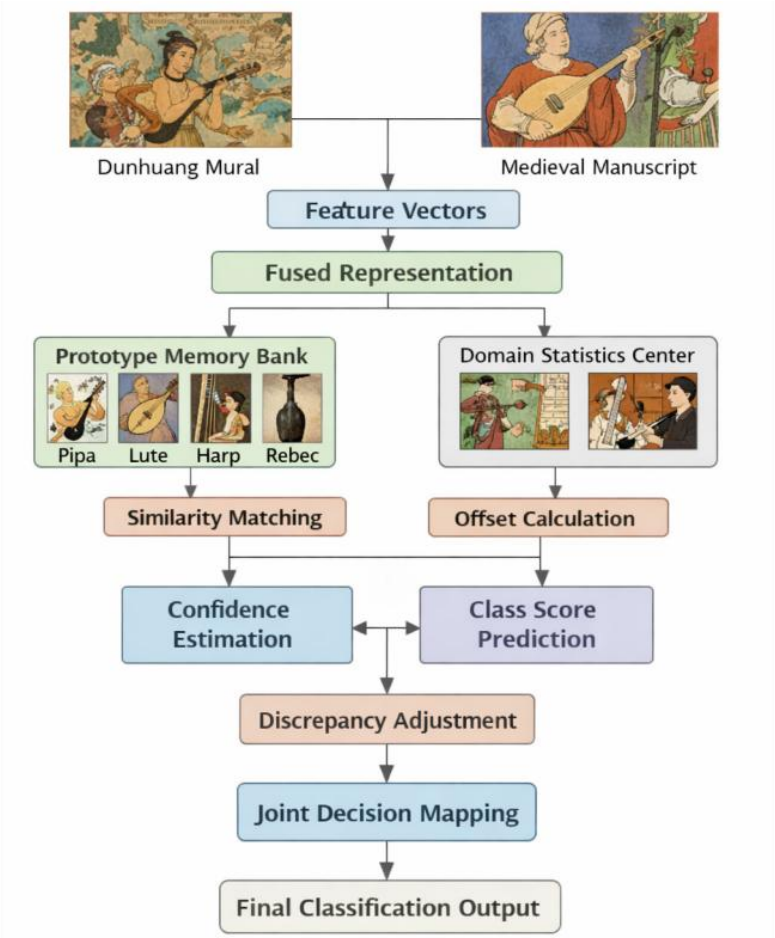

Figure 4: Instrument category discrimination and decision-making mechanism under the constraint of cross-cultural differences

The structure shown in Table 3 shows that the discriminative process is not a single-step classification, but is jointly completed by prototype matching, difference correction and credibility recalibuation. The domain difference constraint layer is responsible for correcting the offset caused by the source domain. The confidence estimation layer is responsible for evaluating whether the current sample has a clear enough structure basis. The joint decision-making layer maps the three types of results into the final output. Considering the imbalance of sample density between musical instrument categories, the prototype matching layer does not

directly use the fixed category center, but uses the combination of batch update and momentum update to maintain the prototype library. In this way, the konghou class, Lyra class and Longhorne class with a small sample size will not be covered by the high-frequency class in the early stage of training, and the class center can also maintain a good stability. At the same time, the domain statistics center does not participate in the competition of category prototypes, but is only responsible for describing the global shifts in color distribution, contour closure and local texture density between the Dunhuang domain and the European domain, thus separating the category modeling and domain difference modeling. It receives input that has been prototypal compressed, difference corrected, and screened for credibility, so that the final output is closer to a constrained stable judgment than to an unscreened direct classification. Such a design is particularly necessary for cross-cultural image classification, as the lateral grip, group embedding, and local damage in Dunhuang images act on the visual representation at the same time as the frontal display, perspective shortening, and highlight reflection in European images.

Table 3: Hierarchical organization of discrimination and decision mechanisms

Module	Input	Output	Function
Category Prototype Matching Layer	Fused representation (f)	Similarity score (s)	Computes the degree of closeness to each category
Domain Difference Constraint Layer	Representation (f) and domain center (c)	Difference offset (Δ)	Suppresses cultural-domain shift
Confidence Estimation Layer	Fused representation (f)	Confidence value (r)	Evaluates sample reliability
Joint Decision Layer	(s), (Δ), (r)	Final score (o)	Produces a stable discrimination result

In order to establish the basic discriminant relationship between the class prototypes and the samples, the class prototype matching layer first calculates the normalized distance between the fusion representation and the center of each class:

$$D_{ik} = \|f_i - p_k\|_2^2, \quad s_{ik} = \frac{\exp(-D_{ik}/\tau)}{\sum_{j=1}^{K} \exp(-D_{ij}/\tau)} \tag{18}$$

Here, f_i represents the fused representation of the i sample, p_k represents the prototype of the k class instrument, D_{ik} represents the Euclidean distance of the sample to the center of the class, τ represents the temperature coefficient, and s_{ik} represents the prototype matching score of the i sample belonging to the k class. Equation (18) establishes the category judgment based on the structure prototype distance, so that the model can still complete the initial discrimination around the shape commonality when facing images from different cultural domains.

After prototype matching is completed, the domain difference constraint layer further estimates the offset of the sample with respect to the statistical center of the local domain, which is calculated as follows.

$$\Delta_i = \tanh\left(W_\delta(f_i - c_{d_i}) + b_\delta\right), \quad \hat{s}_i = s_i - \lambda_\delta \Delta_i \tag{19}$$

Here, c_{d_i} represents the statistical center of the cultural domain to which the sample belongs, W_δ and b_δ are learnable parameters, Δ_i represents the difference offset vector, $\hat{s}_i$

represents the modified category score, and λ_δ represents the domain difference suppression coefficient. Equation (19) can weaken the traction effect of cultural style on classification boundary, so that differences in plane color and shading modeling no longer directly dominate the category output.

In order to avoid high scores of low-quality samples in the discrimination stage, the system sets up an independent confidence estimation layer, and combines the representation strength and distribution uncertainty to generate confidence:

$$r_i = \sigma(W_r f_i + b_r), \quad q_i = -\sum_{k=1}^{K} \hat{s}_{ik} \log \hat{s}_{ik} \tag{20}$$

where r_i represents the sample confidence value, W_r and b_r are mapping parameters, $\sigma(\cdot)$ represents the Sigmoid function, and q_i represents the entropy value of the modified distribution. In Equation (20), the representation features and distribution uncertainty are included in the credibility evaluation at the same time, so that the samples with serious local defects, fuzzy boundaries or similar indirect close are automatically downgraded in the subsequent decision.

After obtaining the correction score and confidence value, the joint decision layer writes them into the final output space together, which is expressed as follows.

$$o_i = r_i \odot \hat{s}_i + (1 - r_i) \odot \frac{1}{K} 1, \quad \hat{y}_i = \arg\max(o_i) \tag{21}$$

where o_i represents the final decision score, 1 represents the all-one vector, K represents the total number of categories, and $\hat{y}_i$ represents the final predicted category. Equation (21) retains the prototype matching results on the high confidence samples and introduces a uniform smoothing term on the low confidence samples, thereby suppressing the extreme biased output and improving the discrimination stability.

The joint objective function is optimized synchronously in an end-to-end manner. The category supervision term ensures the basic discriminative power, the difference constraint term limits the source domain shift, and the confidence calibration term keeps the output score consistent with the sample reliability. After the combined effect of the three, the model will not blindly pursue a higher maximum category score, but tends to form a discrimination space with smoother boundaries and more balanced cross-domain distribution. The overall objective function for the training phase is written as follows.

$$L_{dec} = \lambda_c \sum_i CE(y_i, o_i) + \lambda_d \sum_i \| \Delta_i \|_2^2 + \lambda_r \sum_i (r_i - \max(o_i))^2 \tag{22}$$

Here, $CE(\cdot)$ represents the cross-entropy loss, y_i represents the true class label, and λ_c, λ_d, and λ_r represent the loss weights. Equation (22) enables the model to compress unnecessary domain shifts while ensuring the classification accuracy, and to keep the confidence output consistent with the final discriminant strength.

After the above processing, the discrimination and decision mechanism no longer relies on a single class probability, but forms the final output under the common constraints of structural prototype, domain difference and sample confidence. This design can keep the plane-outline Musical Instruments in Dunhuang murals and the volumetric near-shape Musical Instruments in European images separable in the shared space, and make the samples with local damage, blurred edges and strong background interference obtain more robust decision results. On the

whole, the final output not only maintains the clarity of the category boundary, but also maintains the stability and interpretability in the process of cross-cultural classification, and the overall discrimination is more stable.

3 Results

3.1 Experimental analysis of comparative classification of cross-cultural musical instrument images

The experimental platform uses AMD Ryzen 9 7950X processor, NVIDIA RTX 4090 graphics card, 24GB video memory and 64GB memory, the operating system is Ubuntu 22.04, and the deep learning framework is PyTorch 2.1. The training batch size is set to 32, the initial learning rate is set to 0.0003, the optimizer uses AdamW, the weight decay is 0.01, the learning rate scheduling uses cosine annealing strategy, and the minimum learning rate is set to 1e-6. The model is trained for 120 epochs and the iteration is stopped when the validation set Macro-F1 is not promoted for 12 consecutive epochs. The sample library consists of Dunhuang mural instrument images and medieval European instrument images, which contains 9216 samples in total, and is divided into training set, validation set and test set according to 7:1.5:1.5. Data augmentation uses cropping, brightness perturbation, mild rotation, occlusion simulation, and normalization, and does not introduce strong geometric distortions that would change the shape topology of the appliance. The evaluation metrics are Accuracy, Macro-F1, Cross-domain Recall and Expected Calibration Error, which are used to simultaneously investigate the overall classification ability, fine-grained balance, cross-domain recall ability and confidence stability.

Fig. 5 shows the variation trend of the Accuracy and Macro-F1 of different models on the validation set. The full model enters the stable lifting interval after the 18th epoch, and the convergence speed is faster than the three comparison models of ResNet50, SWI-T and CLIP-Linear. The Accuracy reaches 94.8% in the 42nd epoch, and the fluctuation is controlled within 0.3 percentage points thereafter. The hierarchical visual semantic alignment can complete the representation construction of the shape earlier. Macro-F1 reaches 93.6% in the 50th epoch, which is 4.7 percentage points higher than ResNet50 that relies on the convolution backbone, and 3.9 percentage points higher than CLIP-Linear that only performs linear mapping. The results show that after the reorganization of structure, vector and topological relations in the semantic space, it can provide clearer boundary support for Konghou class, especially for Konghou class, lute class and harp class samples.

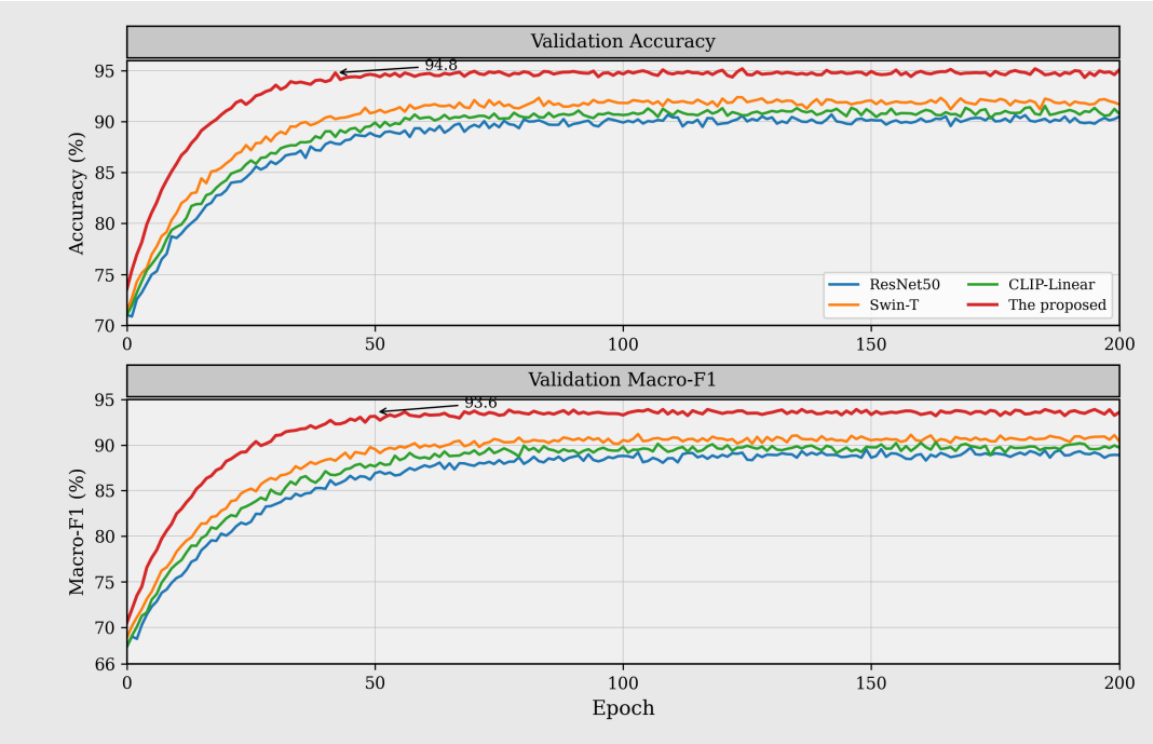

Figure 5: Accuracy versus Macro-F1 variation trend of different models on the validation set

Table 4 lists the overall classification results of the full model versus the mainstream baseline on the test set. The complete model achieves the best performance on the four indicators, and the Cross-domain Recall reaches 92.4%. The model maintains a consistent recognition ability between the Dunhuang domain samples and the European domain samples. ECE drops to 0.021, the deviation between the output score and the true correct rate is small, and the classification probability after gated fusion is stable. The improvement of contras classification performance does not depend on the backbone depth, but comes from the synergy of three types of mechanisms: alignment, interaction and fusion.

Table 4: Overall classification performance comparison

Model	Accuracy (%)	Macro-F1 (%)	Cross-Domain Recall (%)	ECE
ResNet50	90.1	88.9	85.7	0.046
Swin-T	91.8	90.4	87.6	0.038
CLIP-Linear	90.9	89.7	86.8	0.041
Full Model	94.8	93.6	92.4	0.021

Fig. 6 further presents the cross-domain confusion matrix of the full model on the test set. There is some overlap between the pipa class in Dunhuang images and the lute-like class in European images, but the proportion of misjudgments has been controlled within 6.3%. The confusion rate between the harp class and the harp class decreases from more than 14% in the comparison model to 8.1%, and the category archetype aggregation and cross-cultural interaction concompression the decision overlap of the juxomorphoid exotic samples. The diagonal distribution of drum class, horn class and flute class is more concentrated, indicating that the categories with sharper shape boundaries can form higher discrimination confidence after the contour topology and direction spectrum features are involved.

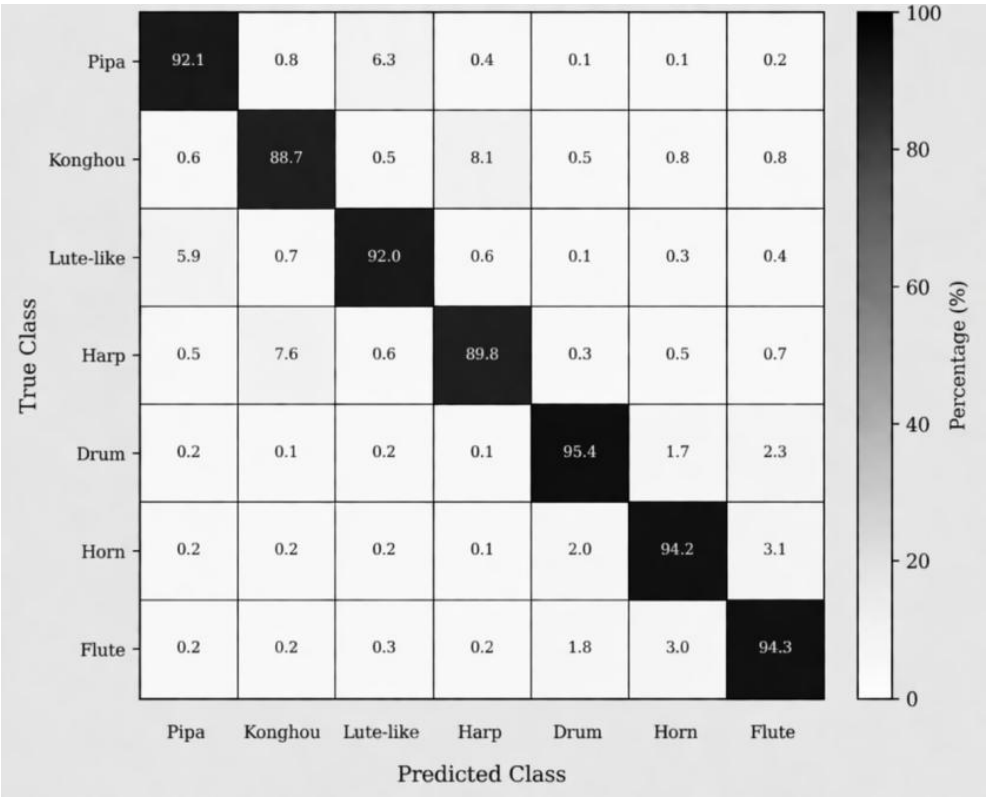

Figure 6: Confusion matrix for test set of cross-cultural musical instrument images

In order to verify the contribution of each module to the overall results, the experiment is organized by gradually adding components, and four groups of configurations are set up, Base, Base+VA, Base+VA+CI and Full. Base only retains block-level visual encodings and linear classification heads, VA represents hierarchical visual alignment, CI represents cross-cultural interaction, and Full adds adaptive gated fusion and disparity constraints. Table 5 shows that the Accuracy is improved from 90.7% to 92.6% after adding visual alignment, and the semantic space can reduce the representation offset between the two types of cultural images. After adding cross-cultural interaction, the Cross-domain Recall increases to 90.8%, and the dual-domain semantic memory enhances the connection of similar samples from foreign domains. The complete model improves Macro-F1 to 93.6%, and lowers ECE to 0.021. The gated fusion not only improves the classification accuracy, but also improves the output calibration.

Table 5: Results of module ablation experiments

Configuration	Accuracy (%)	Macro-F1 (%)	Cross-Domain Recall (%)	ECE
Base	90.7	89.2	84.9	0.051
Base + VA	92.6	91.1	88.7	0.036
Base + VA + CI	93.4	92.2	90.8	0.029
Full	94.8	93.6	92.4	0.021

From the overall experimental results, the advantage of the full model is mainly reflected in the synchronous convergence of the two types of samples. In the Dunhuang mural samples, local damage, color layer pealing and figure overlaying do not significantly weaken the extraction of the main features of the shape. In the European sample, perspective shortening, light and dark contrast, and decorative details also did not cause large-scale category drift. Table

4 shows that the Accuracy of the complete model reaches 94.8%, Macro-F1 reaches 93.6%, and Cross-domain Recall reaches 92.4%, indicating that the classification results are not only dominant in high-frequency categories, but also maintain a balanced recognition ability on cross-domain samples. Table 5 further shows that with the addition of hierarchical visual alignment, the model is already able to compress the local representation offset; After the addition of cross-cultural interaction, the correspondence between samples from the same class and foreign domains is more stable. After the gated fusion is completed, the ECE is reduced to 0.021, and the confidence of the final output is more consistent with the true classification results. Combining the convergence trend in Fig. 5 and the confusion distribution in Fig. 6, it can be judged that the classification of near-shape categories by this method no longer depends on a single texture difference, but is based on the combined effect of the body structure, component combination and cultural domain statistical constraints, so it can still maintain a relatively stable classification boundary in complex scenes.

3.2 Analysis of recognition results of different cultural domains and musical instrument subdivision categories

On the basis of the overall classification results have been determined, continuing to separate the recognition of different cultural domains and musical instrument subdivision categories can more clearly observe the discriminative boundaries of the model under cross-cultural conditions. After the test set is divided into Dunhuang domain and European domain according to cultural source, the complete model maintains high stability in both domains, but the feature dependence is not completely consistent. In the Dunhuang domain samples, the proportion of low saturation coloring, local damage and group image embedding is high, and the model relies more on contour topology and direction spectrum features to complete the judgment. In the European domain samples, volume shadow, background overlay and object decoration are more obvious, and the interaction between color statistics and local blocks has a more direct impact on the classification results. This difference does not cause significant performance imbalance, indicating that the hierarchical visual semantic alignment has compressed the source offset to a manageable range.

Fig. 7 shows the 3D scatter distribution of Dunhuang domain and European domain samples in the fused representation space. The horizontal axis is the structure representing principal component Z1, the vertical axis is the texture-color representing principal component Z2, and the vertical axis is the cross-domain alignment response value Z3. The scatter points are distinguished by cultural domain and category source, with the Dunhuang domain samples represented by circle points and the European domain samples represented by triangle points. In the figure, four groups of high-frequency proximal categories, namely, the pipa class, the konghou class, the lute class and the harp class, are selected as the observation objects. The alignment results show that the scatter centers of similar instruments in the two domains are significantly closer to each other. Taking the pipa class as an example, the two sample points in Dunhuang domain are distributed in (2.31,1.42,0.88) and (2.45,1.36,0.92), and the corresponding points in European domain are (2.18,1.61,0.79) and (2.27,1.54,0.84). The four points are relatively close in dimension Z1, but still retain hierarchical differences in dimension Z2 and Z3. The separation of Konghou class and harp class is mainly reflected in the Z2 dimension, with the samples of Konghou class concentrated in the range 2.44 to 2.72, and the samples of harp class concentrated in the range 1.97 to 2.15, indicating that the fusion representation has coded the instrumentality structure and cultural domain deviation separately.

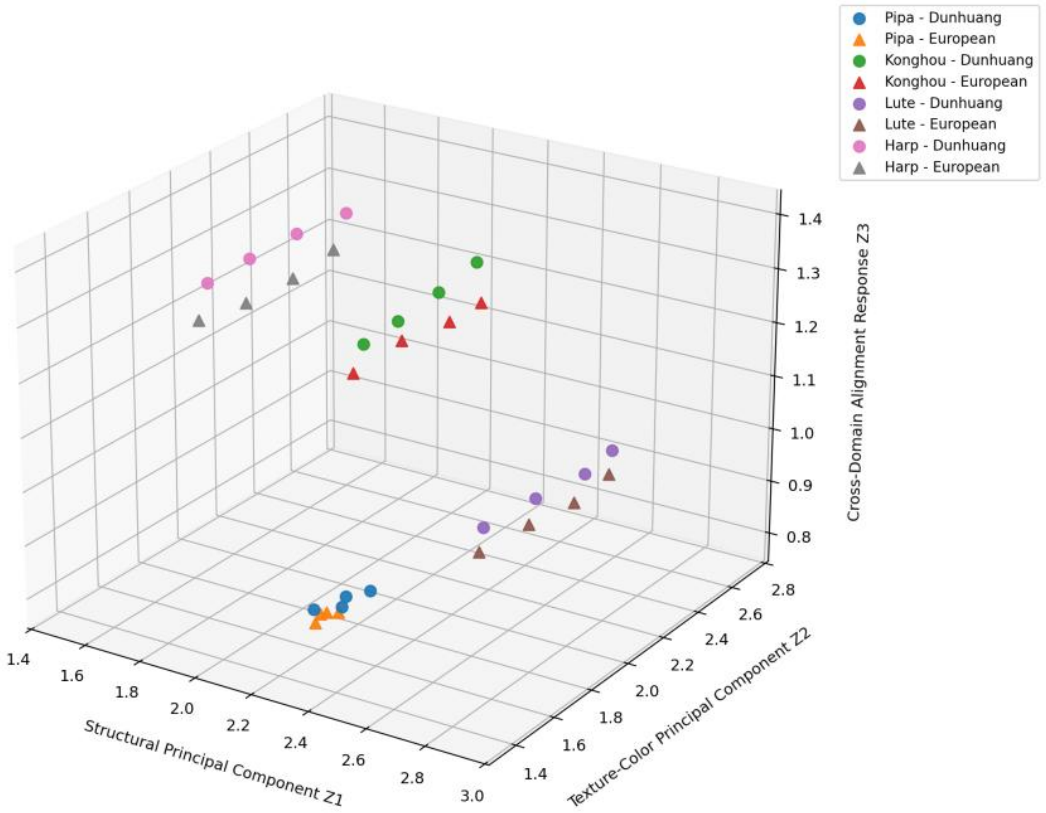

Figure 7: Scatter plot of the distribution of cross-cultural samples in the 3D fused representation space

Table 6 lists the recognition results of the full model on the two types of cultural domains. The Dunhuang domain Accuracy is 94.5%, Macro-F1 is 93.1%, Cross-domain Recall is 91.8%, and ECE is 0.020. The European domain Accuracy is 95.0%, Macro-F1 is 93.8%, Cross-domain Recall is 92.9%, and ECE is 0.023. The Accuracy difference between the two domains is only 0.5 percentage points, and the Macro-F1 difference is 0.7 percentage points, indicating that the model does not show significant single-domain bias. The Cross-domain Recall in Dunhuang domain is slightly lower, which is mainly related to local damage and figure occlusion. The ECE is slightly higher in the European domain, indicating that strong shading and decorative details introduce a slight confidence shift. However, from the perspective of the overall amplitude, the inter-domain discrimination ability maintains a good balance.

Table 6: Recognition results on different cultural domains

Cultural Domain	Accuracy (%)	Macro-F1 (%)	Cross-Domain Recall (%)	ECE
Dunhuang Domain	94.5	93.1	91.8	0.020
European Domain	95.0	93.8	92.9	0.023

In order to continue to observe the difficulty of fine-grained category recognition, this paper regrouped the 16 musical instrument subcategories according to the similarity of instrument shape, and counted the confusion within the group. The results show that the near-shape misjudgment still accounts for the main proportion in the chord-sound category, especially between the pipa and the lute, and the konghou and the harp. In contrast, the boundaries of the flute, horn, and drum classes are clearer, and the confusion mainly occurs in samples where the

instrument occupies only a small area of the picture or the components are occluded. This result indicates that the key of cross-cultural classification is not coarse category distinction, but fine-grained judgment of the same category but close neck proportion, support structure and opening direction.

Fig. 8 presents the 3D scatter results in discriminant space for pairs of highly confused musical instrument categories. The horizontal axis is the class recall, the vertical axis is the class precision, and the vertical axis is the average confidence score. Each scatter point corresponds to the comprehensive performance of a subdivision category on the test set. The closer the point position is to the high-value region on the upper right, the better the recognition accuracy and decision stability of the category are. The figure shows that the drum class, horn class and flute class are distributed in the high Recall, Precision and Confidence region. The recall, precision and confidence of the drum class are 95.0%, 95.4% and 0.961, respectively, and those of the horn class are 93.5%, 94.2% and 0.947. The classification of flute is 94.1%, 94.6% and 0.952, and the discrimination boundary is relatively stable. The three dimensional coordinates of harp, harp, harp and pipa are 91.9,92.8,0.931,91.2, 91.7,0.924, 89.7,90.8,0.908 and 90.4,91.2,0.914, respectively. Konghou class and harp class have the smallest scatter distance, indicating that this group of classes is still the main identification intensive area in the fine-grained classification, but its 3D position has been significantly higher than the distribution level under the basic model condition.

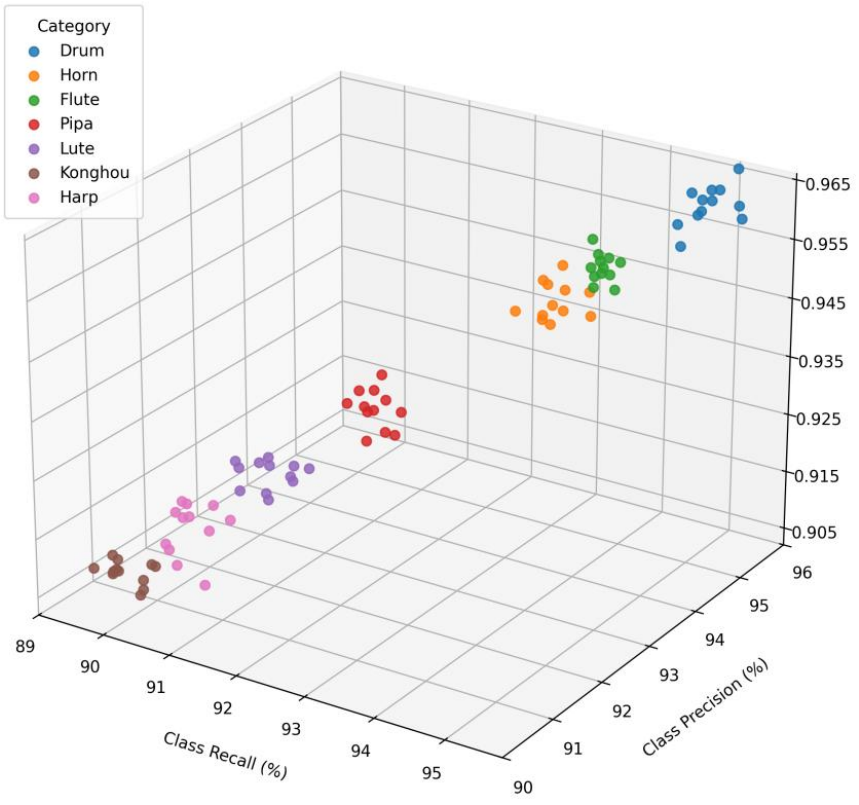

Figure 8: 3D performance scatter plot for pairs of high confusion musical instrument categories

Table 7 further lists the Precision, Recall, and F1 results for the four sets of key segmentation categories. Although both lute and lute contain pear-shaped resonator and chord sequence structures, the complete model has stabilized the F1 of the two classes at more than 91%, reaching 92.3% and 91.4% respectively. Konghou and harp are still the most difficult

groups to distinguish, but the Recall has reached 89.7% and 90.4%, and the F1 is 90.2% and 90.8%, respectively, indicating that after cross-cultural interaction, the model can make more detailed distinctions by using the direction of the support rod, the opening direction of the harp frame and the connection relationship of the components.

Table 7: Identification results for key segmentation categories

Category	Precision (%)	Recall (%)	F1 (%)
Pipa Category	92.8	91.9	92.3
Lute Category	91.7	91.2	91.4
Konghou Category	90.8	89.7	90.2
Harp Category	91.2	90.4	90.8

From the combination of the domain-level results and the subdivided category results, the advantage of the complete model is not only reflected in the overall accuracy, but also reflected in the synchronization and stability of the two types of cultural domains and multiple groups of near-shape categories. The low-saturation damaged areas in the Dunhuang domain samples did not weaken the main features of the shape, and the light and dark levels and local decorations in the European domain samples did not misrepresent the cultural style as the category boundary. At the category level, the most difficult groups of samples obtain relatively stable Precision and Recall, which indicates that the model has been able to process cultural origin, local texture and organ structure separately. At the same time, the confidence calibration in the decision layer does not significantly compress the scores of high-frequency categories, but controls the overconfidence of low-quality samples in a smaller range, making the cross-domain output more stable. On the whole, the classification results of cross-cultural musical instrument images maintain both the inter-domain balance and the clarity of fine-grained category boundaries.

4 Conclusion

Focusing on the task of cross-cultural comparative classification between Dunhuang murals and medieval European musical instrument images, this paper constructs a complete computational link consisting of visual feature extraction, hierarchical visual semantic alignment, adaptive gated fusion, and difference constraint discrimination. The experimental results show that the model achieves 94.8% Accuracy, 93.6% Macro-F1, 92.4% Cross-domain Recall and 0.021 ECE, which can extract the features under the conditions of low saturation, local damage and group image embedding in Dunhuang domain. It is also able to maintain clear category boundaries under conditions of European domain light and dark contrast, perspective shortening, and decorative overlays. Subdivision category analysis shows that the recognition results of highly confused samples such as pipa, lute, harp and harp improve synchronously, indicating that the proposed method has been able to model the cultural origin differences, local texture changes and organ structure relationship separately. The current model is still sensitive to few-sample categories and severely occluded samples, and the update efficiency of dual-domain semantic memory is also affected by the batch distribution. Future research can continue to introduce open vocabulary supervision, cross-collection transfer training, and lightweight deployment strategies, and combine knowledge graph constraints and retrieval enhancement mechanisms to improve the generalization ability and interpretation ability of cross-cultural instrument image classification. Through hierarchical mapping, cross-domain interaction and dynamic weight allocation, the model uniformly encodes and gradually updates the visual information of different granularities, and forms a stable output at the decision end.

Such a structure can compress the representation shift caused by cultural domain differences, weaken the interference of near-shaped heterogeneous samples on the classification boundary, and make the class distribution in the shared space clearer.

Funding

Jiangsu Postgraduate Research & Innovation Program

Project Title: A Study on Instrument Forms and the Aesthetic Ideology of Chinese Traditional Crafts

Project Number: KYCX25_2571

References

[1] Yu T, Lin C, Zhang S, et al. Artificial intelligence for Dunhuang cultural heritage protection: the project and the dataset[J]. International Journal of Computer Vision, 2022, 130(11): 2646-2673.

[2] Zhou Z, Liu X, Shang J, et al. Inpainting digital Dunhuang murals with structure-guided deep network[J]. ACM Journal on Computing and Cultural Heritage, 2022, 15(4): 1-25.

[3] Madhu P, Villar-Corrales A, Kosti R, et al. Enhancing human pose estimation in ancient vase paintings via perceptually-grounded style transfer learning[J]. ACM Journal on Computing and Cultural Heritage, 2022, 16(1): 1-17.

[4] Jiang H, Yang T. Research on the extraction method of painting style features based on convolutional neural network[J]. International Journal of Arts and Technology, 2022, 14(1): 40-55.

[5] Zhong Y, Huang X. A painting style system using an improved CNN algorithm[J]. IEIE Transactions on Smart Processing & Computing, 2022, 11(5): 332-342.

[6] Liu S. Research on the classification method of artistic painting image style based on naive Bayesian[J]. International Journal of Information and Communication Technology, 2022, 21(4): 398-411.

[7] Tan Y. Feature recognition and style transfer of painting image using lightweight deep learning[J]. Computational Intelligence and Neuroscience, 2022, 2022(1): 1478371.

[8] Dinesh Kumar R, Golden Julie E, Harold Robinson Y, et al. Deep convolutional nets learning classification for artistic style transfer[J]. Scientific Programming, 2022, 2022(1): 2038740.

[9] Cascone L, Nappi M, Narducci F, et al. Classification of fragments: recognition of artistic style[J]. Journal of Ambient Intelligence and Humanized Computing, 2023, 14(4): 4087-4097.

[10] Fan T, Wang H, Deng S. Intangible cultural heritage image classification with multimodal attention and hierarchical fusion[J]. Expert Systems with Applications, 2023, 231: 120555.

[11] Geng J, Zhang X, Yan Y, et al. MCCFNet: Multi-channel color fusion network for cognitive classification of traditional Chinese paintings[J]. Cognitive Computation, 2023, 15(6): 2050-2061.

[12] Croce V, Manuel A, Caroti G, et al. Semi-automatic classification of digital heritage on the Aïoli open source 2D/3D annotation platform via machine learning and deep learning[J]. Journal of Cultural Heritage, 2023, 62: 187-197.

[13] Valencia J, Pineda G G, Pineda V G, et al. Using machine learning to predict artistic styles: an analysis of trends and the research agenda[J]. Artificial Intelligence Review, 2024, 57(5): 118.

[14] Sha S, Li Y, Wei W, et al. Image classification and restoration of ancient textiles based on convolutional neural network[J]. International Journal of Computational Intelligence Systems, 2024, 17(1): 11.

[15] Cheng J, Yang L, Tong S. Painting style and sentiment recognition using multi-feature fusion and style migration techniques[J]. Informatica, 2024, 48(21): 127-138.

[16] Li H, Zhu W. Art image style conversion based on multi-scale feature fusion network[J]. Informatica, 2024, 48(10).

[17] Schaerf L, Postma E, Popovici C. Art authentication with vision transformers[J]. Neural Computing and Applications, 2024, 36(20): 11849-11858.

[18] Zhang X. Oil painting image style recognition based on ResNet-NTS network[J]. Journal of Radiation Research and Applied Sciences, 2024, 17(3): 100992.

[19] Liu Y, Bai H, Wang J. Fine-art recognition using convolutional transformers[J]. PeerJ Computer Science, 2024, 10: e2409.

[20] Xiang J, Yang Y, Bai J. Adaptive classification of artistic images using multi-scale convolutional neural networks[J]. PeerJ Computer Science, 2024, 10: e2336.