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.

Research on Mold Spot Detection for Paper Cultural Relics Based on Multimodal Feature Fusion and HNFT Network

zhao ya
Chongqing Jiaotong University
huang guoshi
Chongqing University of Technology
deng xuexu
Chongqing University of Technology
song qiankun
songqk@cqjtu.edu.cn

Chongqing Jiaotong University
song tao
Chongqing University of Technology
tang bin
Chongqing University of Technology
tang huan
State Administration of Cultural Heritage
wu qian
Chongqing University of Education
wang jianxu
Chongqing University of Technology
qin dan
State Administration of Cultural Heritage

Article

Keywords:

Posted Date: June 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9975833/v1

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

Additional Declarations: No competing interests reported.

Research on Mold Spot Detection for Paper Cultural Relics Based on Multimodal Feature Fusion and HNFT Network

Zhao Ya[1,2*], Huang Guoshi [3], Deng Xuexu[3], Song Qiankun[1*], Song Tao[3*], Tang Bin[3], Tang huan[4], Wu Qian[2,1], Wang Jianxu[3], Qin Dan[4]

1. School of Economics and Management, Chongqing Jiaotong University, Chongqing 400074, China; zhaoya@cque.edu.cn (Y.Z.);
2. Department of Tourism and Service Management, Chongqing University of Education, Chongqing 400065, China; songqk@cqjtu.edu.cn (Q.S.)
3. Chongqing Key Laboratory of Optical Fiber Sensing and Photoelectric Detection, Chongqing University of Technology, Chongqing 400054, China; tsong@cqut.edu.cn (T.S.)
4. Key Scientific Research Base of Pest and Mold Control of Heritage Collection (Chongqing China Three Gorges Museum), State Administration of Cultural Heritage, Chongqing 400060, China

✉ *zhaoya@cque.edu.cn, Correspondence: songqk@cqjtu.edu.cn (Q.S.); tsong@cqut.edu.cn (T.S.)

Abstract: Driven by the demands of preventive conservation and risk assessment for cultural heritage collections, the prevention and mitigation of mold deterioration on paper-based relics have emerged as a core challenge. Among these, the precise identification of molds is particularly crucial for early protection and the formulation of differentiated restoration strategies. To address issues such as the insufficiency of single-modal information and the difficulties in small-sample recognition during mold identification, this paper proposes a multimodal feature fusion method for mold spot detection based on a Hyperspectral and Near-infrared Feature fusion Transformer (HNFT) network. The proposed HNFT integrates three data modalities: RGB images, hyperspectral images (HSI), and near-infrared (NIR) images. Specifically, differentiated feature extraction networks are designed for each modality: the RGB branch employs a 2D convolutional neural network (CNN) combined with spatial attention to extract multi-scale spatial textures; the HSI branch utilizes a combination of 3D CNNs and heterogeneous convolutions to extract joint spectral-spatial features; the NIR branch adopts a multi-layer 2D CNN coupled with spectral attention to

capture deep infrared spectral information. Furthermore, an adaptive token allocation mechanism is introduced to dynamically focus on key regions, and a gated cross-attention module is designed within the Transformer architecture to achieve deep cross-modal semantic fusion. Concurrently, to enhance model robustness, a missing-modality configurator based on learnable placeholders is employed to replace the traditional zero-padding strategy. Experimental results demonstrate that the proposed HNFT achieves an Overall Accuracy (OA) of 98.51% and a Kappa coefficient of 97.97%. Notably, the recognition accuracy for the Aspergillus niger category, utilizing a mere 25 training samples, reaches 98.20%, significantly outperforming the baseline models. Additionally, the detail recognition accuracy exhibits an improvement of approximately 3.73% over the second-best model. This study not only provides robust technical support for the high-precision detection of mold spots on paper-based cultural relics but also enables the further utilization of fine-grained deterioration categories and distribution information to construct a quantitative and dynamic risk assessment model for museum paper collections.

Introduction

Paper-based cultural relics not only bear witness to the glorious history of the Chinese nation but also promote the exchange and development of global civilization. Ranging from ancient paintings and calligraphy to classical texts, letters, and archives, these artifacts provide invaluable research materials for future generations with their rich content and exquisite craftsmanship. However, due to the fragility of their materials, paper-based relics are vulnerable to various threats during preservation, posing severe challenges to their conservation and restoration. In recent years, a consensus has emerged within the international cultural heritage conservation community: under the concept of preventive conservation, establishing an objective and multi-dimensional deterioration risk assessment system for museum collections is the foundation for implementing precise prevention and control[1,2]. According to a survey on the types of deterioration of paper-based relics in the Zhengzhou Museum conducted by Jin Hui et al. [3], approximately 88% of paper artifacts suffer from mold infestation to varying degrees, making mold spot contamination a highly destructive and high-risk factor. Exploiting the organic nature and high hygroscopicity of paper fibers, molds not only destroy the physical structure but can even trigger irreversible chemical changes, severely compromising the historical and artistic value of the relics[4,5].

Currently, mold eradication methods for cultural relics are primarily classified into three categories: mechanical removal, physical methods, and biochemical methods. Mechanical methods involve using tools such as brushes and vacuum cleaners to remove pathogenic microorganisms; however, they cannot completely eradicate all microbes[6]. Physical methods utilize technologies like ultraviolet radiation and laser cleaning for sterilization and disinfestation on paper artifacts, yet they exhibit limited efficacy against most bacteria and black fungi[7]. Biochemical methods are relatively more effective, but due to the vast diversity of mold species, a single strategy struggles to achieve comprehensive eradication. Precise restoration necessitates the accurate identification of mold species to prevent secondary damage. Therefore, the accurate identification of mold species is not only a prerequisite for eradication efforts but also the core key to achieving hierarchical risk management of deterioration and constructing a biological risk assessment model for paper-based relics. Traditional mold detection technologies for paper-based artifacts encompass morphological identification, molecular biology techniques[8], and biochemical assays[9].

However, these approaches are largely limited to utilizing only a single modality—either solitary imaging or spectral technology—failing to comprehensively identify molds. Consequently, they struggle to satisfy the data demands of modern cultural heritage conservation for fine-grained risk assessment, thus exhibiting inherent limitations.

Hyperspectral and near-infrared (NIR) spectroscopy technologies are primarily applied in fields such as food quality evaluation[10,11] and crop disease detection [12,13,14]. Furthermore, initial progress has been made in the identification and evaluation of mold contamination within the cultural heritage conservation domain. For instance, Mengzhen Lu et al.[15] proposed an automated labeling method based on hyperspectral images for detecting mold deterioration on murals, effectively addressing the issues of being time-consuming, highly subjective, and lacking consistency associated with traditional manual annotation. By fusing the spatial and spectral information of hyperspectral images, precise identification and labeling of molds on mural surfaces were achieved, demonstrating substantial practical value and research significance. Similarly, Sa Wang et al.[16] analyzed the spectral characteristics of mold deterioration and utilized NIR spectra along with spatial dimensional information to reconstruct visible light reflectance. They constructed a virtual restoration method based on 3D-CNNs, which effectively resolved problems such as color deviation common in traditional RGB padding methods and severe information loss in inverse transformation mold-removal methods, providing reliable technical support for the digital preservation and inheritance of cultural relics. The aforementioned studies provide vital technical support for the precise detection and scientific management of mold spot contamination, further expanding the application potential of hyperspectral and NIR spectroscopy technologies in cultural heritage conservation.

In this study, we propose a deep learning framework based on multi-source multimodal feature fusion and gated attention. This framework extracts spatial textures, joint spectral-spatial features, and deep infrared spectral information from RGB, hyperspectral, and near-infrared images, respectively. It achieves deep cross-modal feature fusion through adaptive token allocation and gated cross-attention modules, and introduces a missing-modality configurator with learnable placeholders to enhance robustness in practical applications. This method not only provides an efficient means for the early detection and pathogen tracking of mold spots on paper-based artifacts but also, by extracting multi-dimensional deterioration feature

indicators, broadens the technological pathway for subsequently establishing a comprehensive risk assessment system for museum paper collections based on multi-source heterogeneous data, bearing significant scientific and practical implications.

The proposed method

In this section, we first introduce the proposed HNFT model. Subsequently, we detail the design logic of feature extraction under three different paths within the model and their roles in multimodal feature extraction and fusion. Finally, we summarize the advantages and potential of the HNFT model. Here, the hyperspectral image features input after data patch partitioning are $\mathbf{X} \in \mathrm{R}^{B \times H \times W \times D}$, where B represents the batch size, H and W represent the spatial dimensions, and D represents the feature dimension.

HNFT Model

This model aims to enhance the accuracy and robustness of mold detection through the efficient fusion of multimodal features. The core design philosophy of the HNFT model is to fully exploit the characteristics of three different data modalities by separately extracting features from the spectral-spatial dimensions of HSI data, the spatial and spectral dimensions of SWIR data, and the spatial dimensions of RGB images. Through a deep fusion mechanism, it achieves the joint representation of global features. The specific network architecture is illustrated in Fig. 1.

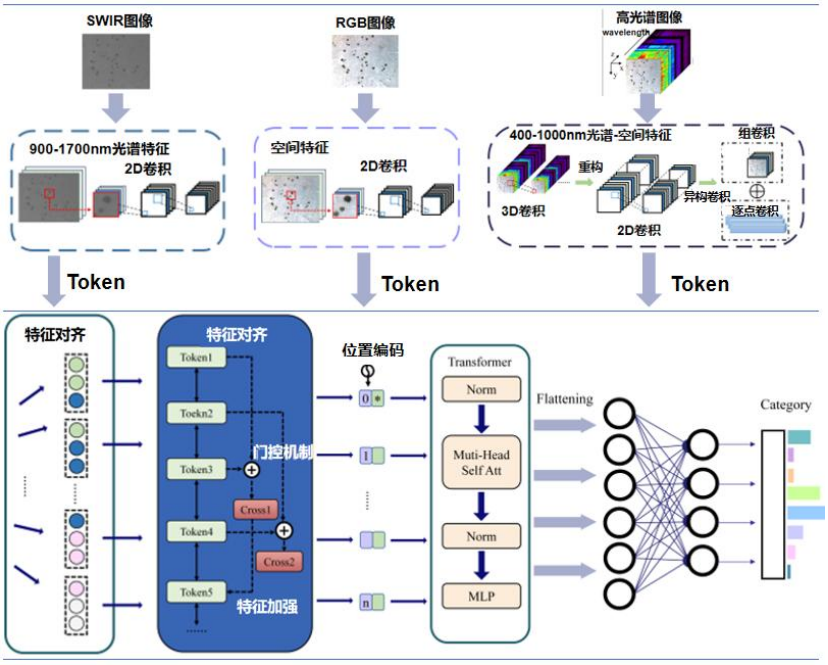


Fig. 1 HNFT Network Structure Diagram

In the HNFT model, the three RGB bands are first extracted from the input hyperspectral image to generate an RGB digital image. Next, distinct extraction methods are applied to different modal features. For the RGB digital modality, a 2D-CNN is utilized to capture spatial characteristics such as edge morphology and color variations of the pseudo-color images. For the SWIR data modality, a 2D-CNN is employed to extract the spatial distribution features and spectral combination information within the 900-1700nm infrared band. For the HSI modality, hybrid convolution (3D-CNN and 2D-CNN) and heterogeneous convolution (group convolution and pointwise convolution) are adopted to extract joint spatial-spectral features within the 400-1000 nm range, deeply analyzing the spatial distribution and spectral sequence variations of mold spots in hyperspectral data[17]. For the aforementioned three types of features, an adaptive token generation module is used to dynamically focus on crucial regions within the feature maps and aggregate these regions into feature tokens, thereby strengthening the representational capacity of the tokens. Finally, a built-in missing modality configurator ensures model robustness under non-ideal inputs. Combined with a modality alignment module and a Gated Cross-Attention mechanism, deep interaction and semantic fusion of cross-modal features are achieved within a unified Transformer architecture[18], significantly improving the accuracy and sensitivity of mold detection in complex environments.

Missing Modality Configurator

In multi-source hyperspectral detection tasks, asymmetric modality missing frequently occurs due to limitations in sensor conditions, illumination environments, and data transmission links. To address this issue, this study introduces a missing modality configurator[19]. By incorporating learnable placeholders, it replaces traditional “zero-padding” or “direct discarding” strategies, pre-setting a globally shared learnable parameter matrix for each modality $\mathrm{P} \in \mathrm{R}^{B \times N \times D}$ (where N is the number of generated tokens). The core value of this mechanism lies in preserving the continuity of the feature space, ensuring that the model maintains consistent decision-making logic even when the input data is incomplete[20].

Specifically, the model presets a globally shared and parameterized quantitative matrix template $\mathrm{P} \in \mathrm{R}^{B \times N \times D}$ for each modality. During the initialization phase, this template discards static zero initialization in favor of random initialization using a Gaussian normal distribution $N(0, \sigma^2)$. To prevent the model from over-relying on missing modalities in the early stages of training, the initial weights are multiplied by

an extremely small scaling factor α (0.02). When the absence of a certain input feature is detected, the model keeps the computational graph topology constant and utilizes a Tensor Expand mechanism to extend the unitary template P_m along the batch dimension as a view, thereby constructing a dynamic completion tensor:

$$\mathrm{P}_{batch} = \mathrm{Expand}(\mathrm{P}_m, [B, N, D]) \tag{1}$$

This operation incurs almost zero computational overhead because it only logically alters the tensor's stride without copying actual values in the GPU memory, achieving highly efficient completion under large-scale parallel computing. The generated batch placeholders must then pass through a Layer Normalization (LN) process. The purpose of this step is to forcibly constrain the numerical distribution of the pseudo-features onto a hypersphere similar to that of the real features. This ensures that when calculating correlation weights, there will be no severe imbalance in attention allocation due to modality missing. The specific structure is shown in Fig. 2.

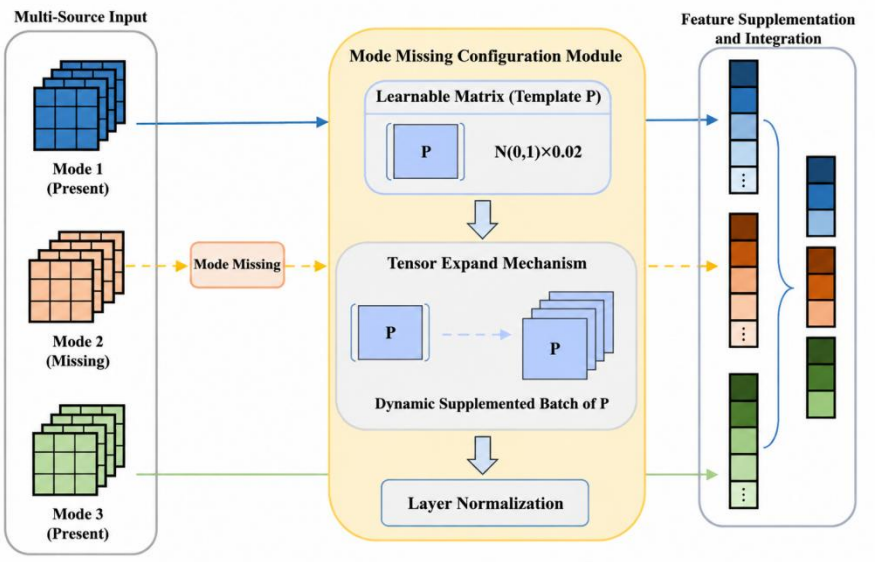


Fig. 2 Structure Diagram of Modal Absence Configurator

Modality Alignment Module

Due to significant heterogeneity in spatial resolution, spectral coverage, and channel dimensions among RGB, SWIR, and HSI data, traditional feature concatenation often causes the model to fall into local optima, tending to overfit a dominant modality while ignoring complementary cross-modal information. Therefore, this study designs a Modality Alignment Module aimed at constructing a shared high-dimensional semantic manifold, projecting heterogeneous features into a unified metric space.

Assume that two input feature sequences in the model are $X_1 \in R^{N_1 \times D}$ and $X_2 = R^{N_2 \times D}$. First, a feature projection layer is utilized to uniformly map the original RGB spatial features, SWIR infrared features, and HSI spatial-spectral features into a high-dimensional linear semantic manifold, eliminating dimensional discrepancies across different sensors:

$$Z_1 = \mathrm{GELU}\left(\mathrm{LN}\left(X_1 W_{p1} + b_{p1}\right)\right) \tag{2}$$

$$Z_2 = \mathrm{GELU}\left(\mathrm{LN}\left(X_2 W_{p2} + b_{p2}\right)\right) \tag{3}$$

where Z_1 and Z_2 are the aligned projected features, and W_p and B_p are the learnable weight matrices and bias variables. Subsequently, a dot product probe is used to calculate the semantic correlation between each Token of the input features X_1 and X_2, constructing an association matrix S:

$$S = \frac{Z_1 Z_2^{\cdot}}{\sqrt{D}} \tag{4}$$

where $S \in R^{N_1 \times N_2}$, and $\sqrt{D}$ is a scaling factor used to mitigate the issue of excessively large score values caused by high feature dimensions. A larger value of S indicates higher feature consistency between the two modalities in that local region. A bidirectional attention mechanism is then used to construct a cross-modal similarity matrix S, performing bidirectional Softmax normalization on the correlation matrix to determine the weight distribution of information transfer. This allows each modality to absorb the advantageous information of the other in terms of spatial morphology or spectral components:

$$A_1 = \mathrm{Softmax}_{row}(S) \tag{5}$$

$$A_2 = \mathrm{Softmax}_{col}(S^{\cdot}) \tag{6}$$

Based on the calculated weights, complementary contextual information is extracted from the counterpart modality to enhance the representation of the current modality:

$$R_1 = A_1 \cdot X_1 \tag{7}$$

$$R_2 = A_2 \cdot X_2 \tag{8}$$

Where R_1 is the refined feature of Modality 1. Even if X_1 itself lacks certain spatial details, it can acquire R_2 compensation from X_2.

Finally, a dynamic Gated Fusion mechanism is introduced to generate a dynamic

gate, adaptively balancing the retention ratio between the original features and the aligned enhanced features. The gate value is calculated as follows:

$$g_1 = \sigma(\text{Linear}([X_1; R_1])) \tag{9}$$

where σ is a function whose output ranges between (0, 1).

The linear weighted feature calculation formula is:

$$X_1 = g_1 \odot X_1 + (1 - g_1) \odot R_1 \tag{10}$$

By adaptively adjusting the fusion weights, this mechanism ensures that the model can precisely lock onto the evolutionary features of mold spots under complex background interference. The specific structural diagram is shown in Fig. 3. This design not only circumvents the stringent physical constraints of pixel-level alignment but also effectively eliminates spatial shifts caused by sensor noise through semantic alignment, laying a semantically consistent foundation for subsequent deep fusion.

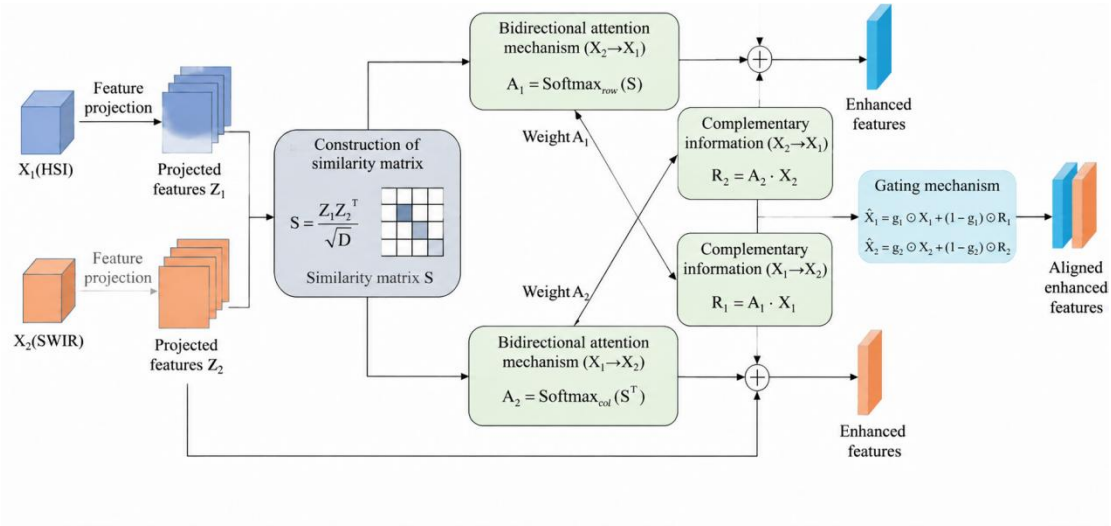


Fig. 3 Structure Diagram of Modal Alignment Module

Gated Cross-Attention Mechanism

The Gated Cross-Attention block (GatedCrossBlock) serves as the core unit of this model for achieving cross-modal high-order semantic interaction. Following the modality alignment layer, this module adopts a symmetric dual-branch architectural design aimed at achieving deep feature fusion by coupling intra-modal and inter-modal information flows. Each processing branch consists of three core components: a self-attention mechanism, a cross-attention mechanism, and a dynamic gating network[21].

Prior to cross-modal exchange, the module first injects local positional encodings into the features of each modality via positional embedding. It then calculates

intra-modal contextual associations through a self-attention mechanism to strengthen the feature representation of each modality itself. Finally, through residual connections and normalization, a solid foundation is established for subsequent cross-modal comparisons. Taking Modality 1 as an example, the specific calculation formulas are as follows:

$$Q_1 = X_1 + E_{loc1} \tag{11}$$

$$K_1 = X_1 + E_{loc1} \tag{12}$$

$$V_{self1} = \mathrm{Softmax}\left(\frac{Q_1 K_1^{\cdot}}{\sqrt{D}}\right) X_1 \tag{13}$$

$$V_1^{(1)} = \mathrm{LayerNorm}(X_1 + \mathrm{Dropout}(V_{self1})) \tag{14}$$

Through this stage of processing, the model effectively extracts structured information within a single modality, laying the groundwork for precise cross-modal alignment. Building upon feature enhancement, the module achieves semantic-level information retrieval by exchanging Query (Q) and Key/Value (K/V). The model utilizes the features of one modality as a guide to retrieve relevant complementary information from another modality. For instance, it leverages the sensitive spectral features of the HSI modality to locate mold spot regions, and accordingly retrieves high-resolution spatial edge details from the SWIR image. This translates abstract spectral features into a composite representation with clear spatial boundaries. The specific calculation formulas are as follows:

$$M_{12} = \mathrm{Softmax}\left(\frac{Q_1 K_2^{\cdot}}{\sqrt{D}}\right) V_2^{(1)} \tag{15}$$

$$M_{21} = \mathrm{Softmax}\left(\frac{Q_2 K_1^{\cdot}}{\sqrt{D}}\right) V_1^{(1)} \tag{16}$$

Finally, to ensure the stability of the fusion process, the module introduces a dynamic gating mechanism. The system adaptively generates a gate coefficient g within the range of (0,1) based on the feature distribution state of the current modality. This coefficient acts as an information filter. By weighting and adjusting the cross-referenced features, it effectively filters out irrelevant background noise or interference from invalid modalities, preventing cross-modal information pollution. The specific calculation formulas are as follows:

$$g_1 = \sigma(\text{Linear}(\text{GlobalAvgPool}(V_1^{(1)}))) \tag{17}$$

$$\hat{M}_{12} = M_{12} \cdot g_1 \tag{18}$$

Ultimately, combined with residual connections and a feed-forward network, the module outputs highly condensed features that are semantically consistent and informationally complementary, significantly elevating the model's accuracy in identifying the evolutionary features of mold spots in complex environments. The specific structural diagram is shown in Fig. 4.

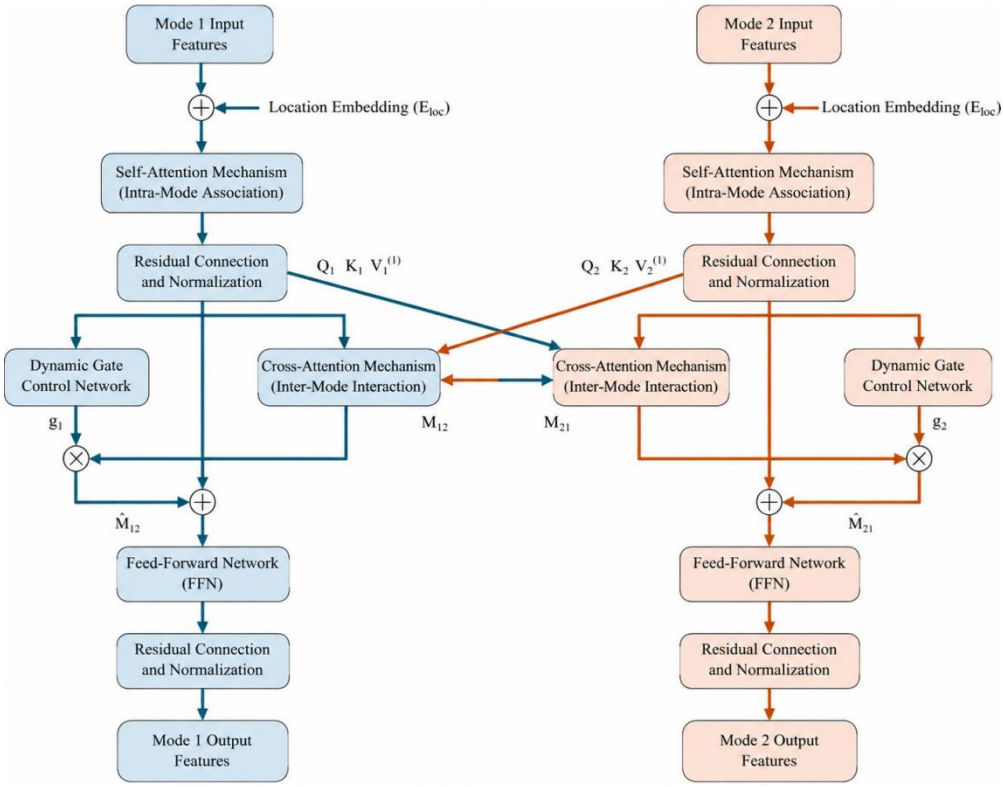


Fig. 4 Structure Diagram of Gated Cross-Attention Mechanism

Multimodal mold data collection

Simulated mold-infected sample

Cotton-hemp raw Xuan paper, cut into 8×8 cm squares, was selected as the paper medium for the simulated mold-infected samples. To ensure the diversity and accuracy of the experimental data, three representative mold samples from different genera were selected as the experimental subjects: Aspergillus niger from the genus Aspergillus, Penicillium citrus from the genus Penicillium, and Trichoderma longiformis from the genus Trichoderma. Belonging to distinct genera, these three molds reflect fungal diversity and ecological adaptability. They are widely distributed in natural environments and can significantly impact the preservation of cultural relics;

therefore, paying attention to molds from these different genera is crucial for the detection and conservation of artifacts. The relic paper and molds required for this experiment were provided by the Key Scientific Research Base of the National Cultural Heritage Administration for Pest Control of Collection Relics. Their physical photographs are shown in Fig. 5.

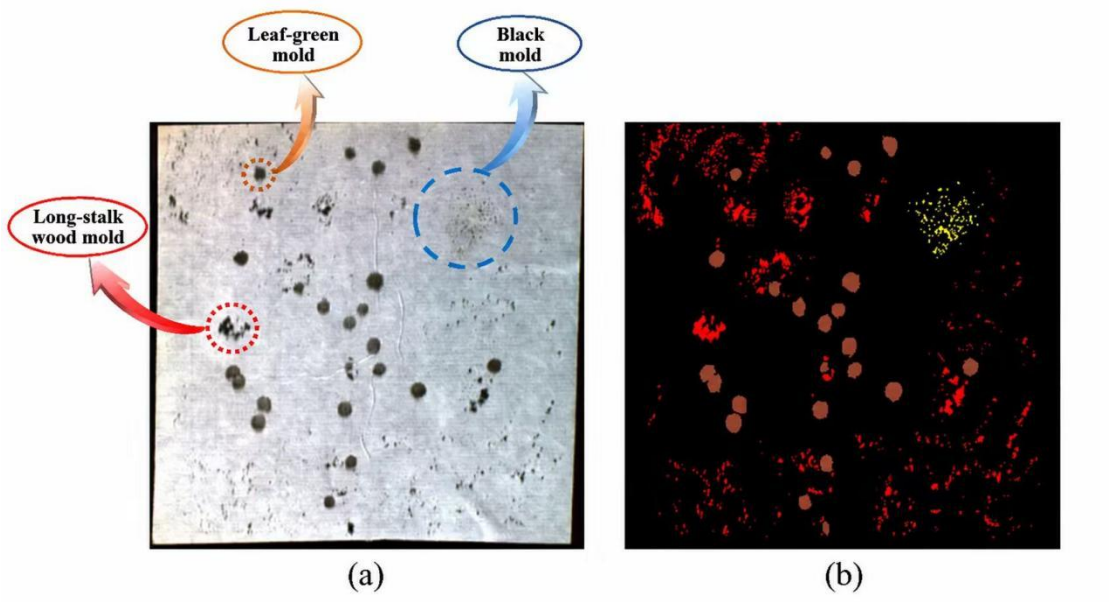


Fig. 5 Simulated mold spot samples of black oak wood. (a) False color image. (b) Mark truth image.

Mold Image Acquisition System for Paper Relics

The hyperspectral mold image acquisition system built for paper relics in this study is illustrated in Fig. 6. It consists of an iSpecHyper-VS1000 portable hyperspectral imager from Lisen Optics, two halogen light sources with adjustable brightness and color temperature, and a Canon lens with a 24mm focal length, covering an acquisition range of 400-1000nm. Through multiple imaging experiments, the optimal parameters were determined to be a focal length of 4615mm, a frame rate of $16\,Hz\cdot s^{-1}$, and an integration time of $56338\,\mu s$. Images captured under these parameters exhibit excellent quality, featuring sharp edges and no obvious distortion.

The Short-Wave Infrared (SWIR) mold image acquisition system for paper relics is shown in Fig .7. It comprises a Bobcat GigE SWIR spectral camera from Xenics, Belgium, for SWIR image acquisition, a halogen light source, and a personal computer (PC), covering an acquisition range of 900-1700nm.



Fig. 6 Hyperspectral mold spot image acquisition system

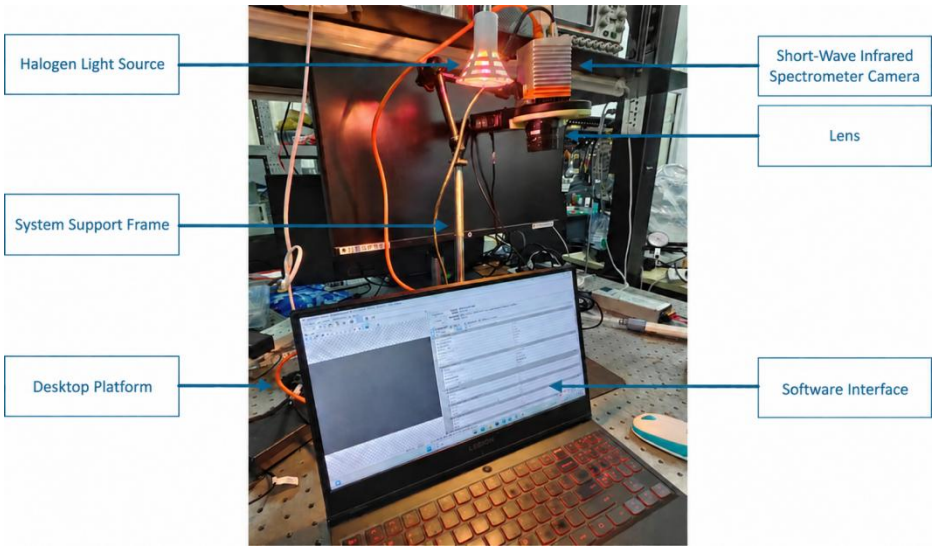


Fig. 7 Short-Wave Infrared spectroscopy image acquisition system

The hyperspectral image dataset used in this study encompasses mold spots generated by the aforementioned three types of mold infections. The sampling height was set to 40 cm. The images have dimensions of 326×333pixels, comprising a total of 2,433 effective pixels and 300 spectral bands. Obtain the multimodal sample dataset as shown in Fig .8, and divide the dataset into a 10% training set and a 90% testing set. The class names and sample quantities used for training and testing in the classification task are detailed in Table 1.

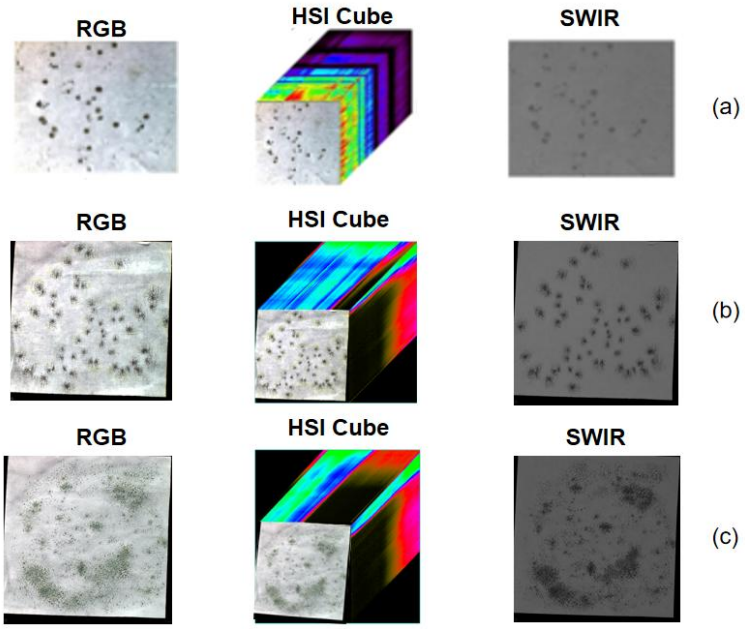


Fig .8 (a) *Penicillium citrinum* (b) *Aspergillus niger* (c) *Trichoderma longibrachiatum* multi-modal sample images

Table 1: Categories of mold spot coverage in the dataset and the number of training and testing samples for each category

Serial number	taxon	Number of training data samples	Number of test data samples
1	Penicillium citrus	115	1026
2	Aspergillus niger	25	217
3	Trichoderma longiformis	105	945
	total	245	2188

Evaluation Metrics

To verify the effectiveness of the model, this study adopts three commonly used classification performance evaluation metrics: Overall Accuracy (OA), Average Accuracy (AA), and the Kappa coefficient (Kappa), to quantitatively analyze the classification results of each model. Specifically, OA reflects the overall classification accuracy of the model across all categories, AA represents the mean classification accuracy of each category, and the Kappa coefficient measures the consistency between the classification results and random classification, thereby providing a more comprehensive evaluation of the model's performance. Concurrently, to further intuitively compare the classification capabilities of different models, qualitative analysis is conducted using classification maps to observe spatial distribution characteristics and the precision of classification boundaries. This allows for a

comprehensive assessment of model classification performance from both quantitative and qualitative dimensions.

Comparison of Different Models

This study selects several state-of-the-art models for comparative analysis with the proposed method, including 1D-CNN[22], 2D-CNN[23], 3D-CNN[24], ViT[25], SSFTT[26], and SpectralFormer[27]. Furthermore, to ensure the uniformity and validity of the experimental data, a standardized data preprocessing method is applied to all data. Specifically, a Savitzky-Golay (SG) smoothing filter is utilized to remove noise, and Principal Component Analysis (PCA) is employed to reduce the spectral bands to 30 dimensions. This process removes redundant information and highly similar features, thereby increasing recognition accuracy.

Experimental Environment Configuration

To validate the proposed hyperspectral image classification algorithm for mold spots based on multimodal feature fusion, all experiments were implemented using the Python 3.12 programming language within the PyCharm 2024.3.1 integrated development environment. The experimental hardware environment was a high-performance personal computer equipped with an Intel® Core i5-14600KF processor, an NVIDIA GeForce RTX 4070 GPU, and 32 GB of RAM. During the experiments, all deep learning models utilized Cross-Entropy Loss as the optimization objective, combined with the Adam optimizer for parameter updates, and employed Stochastic Gradient Descent (SGD) to achieve training optimization. The learning rate was uniformly set to 0.0001, the batch size to 15, and the number of training epochs was fixed at 50 to ensure the stability and convergence of the models, thereby reliably evaluating their performance.

Mold Detection Experiment and Result Analysis

In this section, we first describe the hyperspectral image dataset utilized in our experiments. Subsequently, we analyze the impact of the three different modality extraction modules on the model's performance through ablation studies. Finally, we conduct comparative experiments between the proposed model and other mainstream models, evaluating the mold spot classification performance both quantitatively and visually.

Ablation Experiments

To fully validate the effectiveness of the proposed method, ablation studies on different data fusion mechanisms were conducted on the aforementioned dataset. Four

fusion strategies were considered: concatenation fusion, additive fusion, multiplicative fusion, and cross-attention fusion. The impact of different components on the overall model accuracy was analyzed from the perspective of classification precision. All experimental results are listed in Table 2. Specifically, when employing basic fusion methods, the overall accuracy of the model does not exceed 95%. This indicates that simple linear stacking struggles to fully excavate the deep complementary information among multi-source modalities. Particularly when processing data with complex nonlinear characteristics, such as HSI and SWIR, it is highly prone to causing the loss or dilution of key classification features. When utilizing the cross-attention fusion proposed in this study, the model performance achieves optimality; compared to the worst-performing additive fusion, the OA is improved by approximately 5.63%.

Table 2: Analysis of the multimodal fusion method of the proposed model(bolded data represent the optimal results)

Serial number	fusion mechanism				evaluation index		
	joint	additivity	Multiplicative	gated cross-attention	OA(%)	AA(%)	κ ×100
1	√	×	×	×	93.40	92.25	92.67
2	×	√	×	×	92.88	92.03	91.78
3	×	×	√	×	94.92	94.10	93.36
4	×	×	×	√	**98.51**	**98.36**	**97.97**

To further verify the necessity of multi-source heterogeneous data in the recognition of mold spots on paper relics, this study designed an input modality ablation experiment, as shown in Table 3. By training and testing different combinations of the three modalities (HSI, RGB, and SWIR), we deeply analyzed the impact of each component on the model's classification performance. The experimental results reveal that the model's performance is significantly constrained when only a single modality is used as input. Among them, the recognition performance based on HSI (OA = 85.84%) outperforms the results using solely SWIR (OA = 80.67%) or RGB images (OA = 78.51%). This demonstrates that the rich spectral-spatial features provided by hyperspectral data serve as the core foundation for identifying mold species, whereas a single RGB image performs the worst when dealing with complex mold-infected regions due to the lack of deep biochemical fingerprint information. Through modality combinations, the model's accuracy

exhibits a substantial enhancement. In particular, when HSI is combined with either RGB or SWIR, the OA metrics leap to over 93%. This proves that information supplementation across spectral dimensions can more effectively capture subtle differences in mold metabolic products, thereby enhancing the discriminability of the feature space. When the model integrates all three modalities simultaneously, all performance metrics reach their optimum, achieving a further improvement of approximately 4.11% compared to the best-performing dual-modality combination.

Table 3: Ablation experiments on the fusion of different modalities(bolded data represent the optimal results)

Serial number	Input mode			Evaluation index		
	HSI	RGB	SWIR	OA(%)	AA(%)	$\kappa \times 100$
1	√	×	×	85.84	82.76	83.34
2	×	√	×	78.51	75.19	73.85
3	×	×	√	80.67	78.26	77.31
4	√	√	×	93.82	92.86	88.61
5	√	×	√	94.40	94.15	90.73
6	√	√	√	**98.51**	**98.36**	**97.97**

Quantitative Analysis

Table 4 presents a detailed comparison of the performance of various classification algorithms mentioned in Section 3 on this dataset, encompassing OA, AA, the Kappa coefficient, and the specific classification accuracy for each mold type, with the optimal results highlighted in bold. The evaluation data clearly indicate that the proposed HNFT method significantly outperforms the baseline models across all core evaluation metrics, achieving the highest OA (98.51%), AA (98.36%), and Kappa coefficient (97.97%). Specifically, in terms of small-sample recognition, HNFT exhibits exceptionally strong feature extraction capabilities. Taking Class 2 (Aspergillus niger) as an example, this category contains only 25 samples in the training set, representing typical long-tailed distribution data. The recognition rate of the traditional 1D-CNN on this class is as low as 0.00%, and the 2D-CNN only achieves 14.29%. In stark contrast, by leveraging multimodal feature fusion and global context modeling, HNFT drastically boosts the accuracy for this category to 98.20%, fundamentally resolving the failure issue of single-modality models in recognizing complex biochemical features.

Although advanced algorithms such as SSFTT and SpectralFormer also achieve

an OA exceeding 92%, their balance across different categories remains slightly inferior to that of HNFT. HNFT maintains a high-level recognition accuracy of over 98% across all three mold categories. The high consistency between its AA metric (98.36%) and OA metric (98.51%) proves that the model possesses excellent generalization performance and robustness against various disease morphologies. Furthermore, this study explores the impact of varying proportions of training samples on the performance of each model; the experimental results are illustrated in Fig. 9. As the proportion of training samples gradually increases, the accuracies of all compared models improve. Although the performances of SSFTT and SpectralFormer closely follow, their convergence accuracies consistently remain slightly lower than those of the proposed method.

Table 4 Classification Accuracy of Different Classification Methods on the Dataset (Bolded data represent the best results for each category)

Serial number	1D-CNN	2D-CNN	3D-CNN	ViT	SSFTT	Spectralformer	HNFT
1	81.18	89.48	90.76	83.74	93.60	96.53	**98.73**
2	0.00	14.29	56.21	80.29	94.17	90.17	**98.20**
3	78.78	82.10	85.71	85.21	91.19	95.76	**98.15**
OA(%)	72.07	81.63	85.25	84.87	92.61	94.78	**98.51**
AA(%)	53.32	61.95	75.78	83.08	92.99	93.82	**98.36**
κ ×100	55.11	66.76	78.71	81.17	87.24	91.23	**97.97**

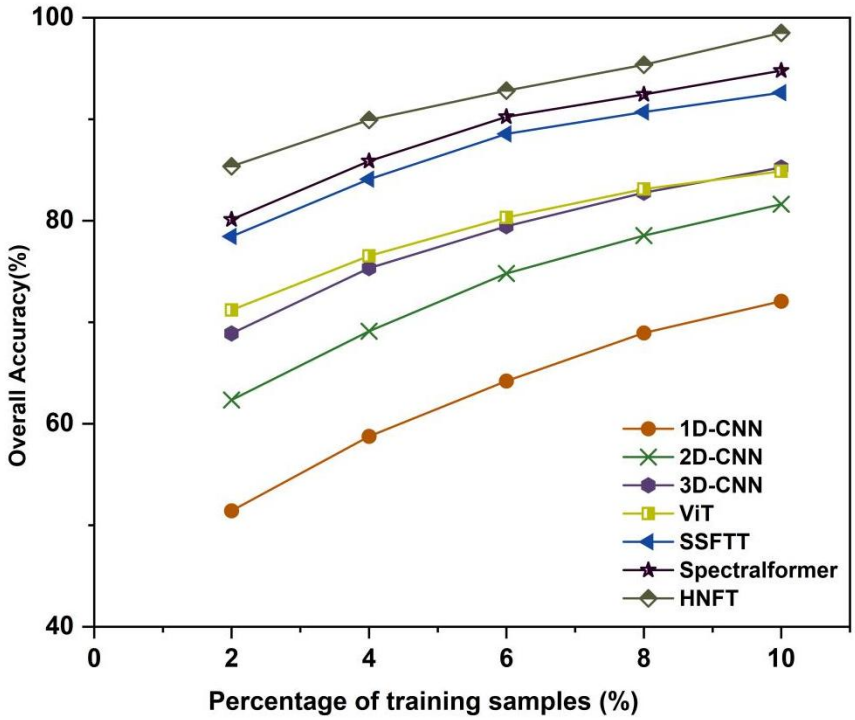


Fig. 9 Comparison chart of OA for different proportions of training samples

Visual Evaluation

To more intuitively evaluate the detection performance of each model on mold spots of paper relics, Fig. 10 displays the classification prediction maps of different algorithms on the testing set. By comparing the classification results of each model with the Ground Truth map (Fig. 10a), it can be observed that the visualization results highly align with the quantitative metrics in Table 4. The classification maps of 1D-CNN (Fig. 10b) and 2D-CNN (Fig. 10c) exhibit abundant noise and misclassified regions. Particularly, their capability to capture small-sample categories is extremely weak, resulting in an overall fragmented appearance of the images. As the model dimension increases, 3D-CNN (Fig. 10d) and ViT (Fig. 10e) show improvements in regional consistency; however, they still fail to precisely restore the edge details of the mold spots. SSFTT (Fig. 10f) and SpectralFormer (Fig. 10g) perform well in identifying larger areas of mold spots. Nevertheless, at the edges of mold spots and in complex regions where different fungal species interweave, varying degrees of “salt-and-pepper noise” and boundary blurring phenomena still emerge. This indicates that traditional hybrid networks or Transformer architectures still possess limitations in balancing spatial details and spectral features. In contrast, the classification map of the proposed HNFT method (Fig. 10h) is the closest to the Ground Truth map. The model not only accurately identifies all fungal categories but also performs exceptionally well in preserving mold morphology and spatial continuity. Even in

minuscule mold spot areas, HNFT achieves precise localization and fundamentally eliminates the misclassified patches common in other models.

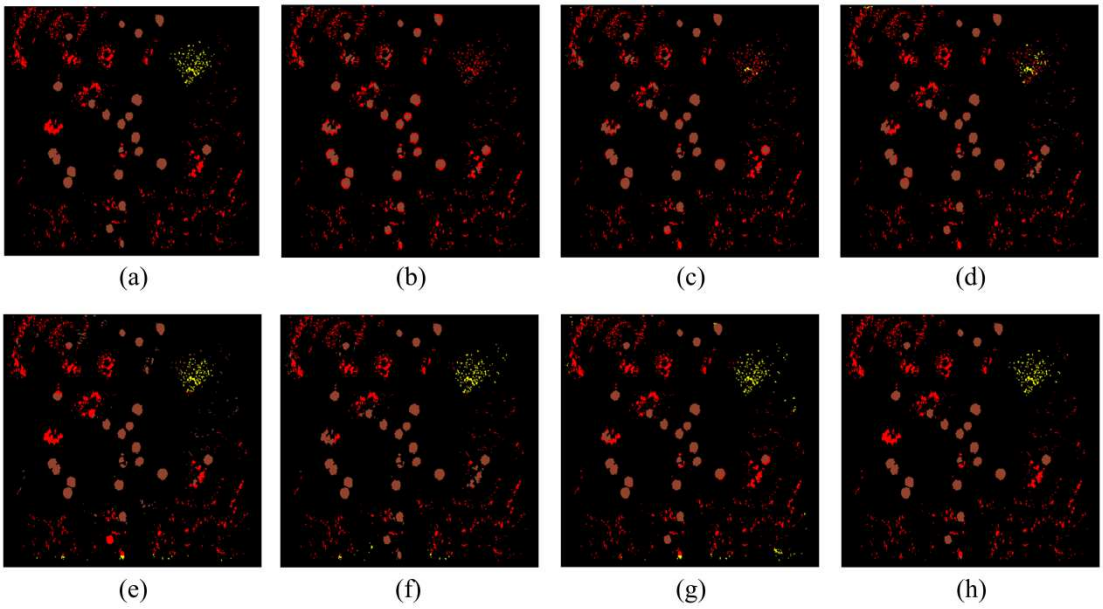


Fig. 10 Classification result graph. (a) Real value graph. (b) 1D-CNN (OA = 72.07%). (c) 2D-CNN (OA = 81.63%). (d) 3D-CNN (OA = 85.25%). (e) ViT (OA = 84.87%). (f) SSFTT (OA = 92.61%).(g) Spectralformer (OA = 94.78%). (h) HNFT (OA = 98.51%)

Conclusion

To address core pain points such as accuracy bottlenecks and the dimensional deficiency of single-modal information in the detection of mold spots on paper-based cultural relics, this paper proposes a multi-source multimodal feature fusion network based on hyperspectral and near-infrared (NIR) imaging. By integrating the spatial textures from RGB images, the fine-grained joint spectral-spatial information from hyperspectral images (HSI), and the deep spectral information from NIR images, the proposed method constructs a cross-dimensional collaborative feature representation mechanism. Ablation studies confirm that concurrently incorporating tri-modal data enhances feature discriminative power by approximately 4.11% compared to a single modality. Such deep integration of multi-source information successfully resolves the misclassification challenge caused by the phenomenon of “different materials exhibiting identical spectra” inherent in traditional detection, thereby enabling the model to maintain exceptional robustness even in complex environments. Experimental results demonstrate the outstanding performance of the HNFT model in the mold spot classification task for paper-based relics, achieving an Overall Accuracy (OA) of 98.51% and a Kappa coefficient of 97.97%. All metrics significantly outperform those of traditional classification methods and single-branch networks.

Particularly in the recognition of edge details and minute deterioration regions, its accuracy exhibits an improvement of approximately 3.73% over the second-best model. This realizes a leap from macroscopic localization to microscopic characterization, providing a robust empirical basis for accurately demarcating the affected scope of the deterioration and quantifying risk levels.

In future work, to accommodate the resource-constrained hardware environments of local museums and archaeological field sites, the HNFT network structure will be further optimized through techniques such as model pruning, knowledge distillation, and quantization. These optimizations aim to enhance computational efficiency while reducing memory footprint, thereby providing lightweight algorithmic support for the development of portable and real-time mold spot detection terminals for paper-based relics. More importantly, the high-precision deterioration classification model and the multi-source feature representation mechanism established in this study will be deeply integrated into the preventive conservation framework for museum collections. Building upon this foundation and incorporating environmental monitoring parameters (e.g., temperature and humidity), future research will conduct systematic and quantitative risk assessments for the biodeterioration of paper-based artifacts. This will propel cultural heritage conservation towards a comprehensive paradigm shift centered on early prediction and preventive mitigation.

Data availability

The data in this study is a unique hyperspectral image dataset. Currently, it is not suitable to fully disclose this data. If necessary, the corresponding author can be contacted to discuss possible forms of data sharing.

Acknowledgements

This research was supported by grants from the Chongqing Talent Project (cstc2021ycjh-bgzxm0287),the Chongqing Special Program for Technological Innovation and Application Development (CSTB2024TIAD-KPX0101),the Chongqing Municipal Education Commission Science and Technology Research Program (KJQN202605773).

Author contributions

The study's conception and design were collaboratively developed by Zhao Ya, Huang Guoshi and Deng Xuexu contributed to the study conception and design. The preparation of materials, as well as data collection and analysis, were preform conducted by Song Qiankun,Tang Bin,Wu Qian,Wang Jianxu,Xiao Zhongyu,Luo

Xiling,Qin Dan.Song Tao and Tang Huan are responsible for oversaw the supervision of the research activities. The first draft of the manuscript was written by Huang Guoshi. All authors reviewed and approved the final version of the manuscript.

Competing interests

The authors declare no conflict of interest.

References

1 Lorusso S, Natali A, Palla F. Risk management in the cultural heritage sector: museums, libraries and archives[J]. 2022.

2 Elnaggar A, Said M, Kraševec I, et al. Risk analysis for preventive conservation of heritage collections in Mediterranean museums: case study of the museum of fine arts in Alexandria (Egypt)[J]. Heritage Science, 2024, 12(1): 1-17.

3 Jin H. Investigation and study on the preservation and protection of paper relics in Zhengzhou Museum[D]. Zhengzhou University, 2015.

4 Yang L G. Causes of damage to paper relics and measures for protection and restoration[J]. Collections, 2023, (12): 110-112.

5 Zhen C A, Zhao D P. Research progress on biological deterioration of paper relics[J]. Journal of Beijing Institute of Graphic Communication, 2019, 27(05): 32-37.

6 Wu F S, Li J, Li R X, et al. Research progress on the application of biocides and antibacterial nanomaterials in cultural heritage conservation[J]. Sciences of Conservation and Archaeology, 2023, 35(05): 115-127.

7 Rivas T, Pozo-Antonio J S, de Silanes M E L, et al. Laser versus scalpel cleaning of crustose lichens on granite[J]. Applied Surface Science, 2018, 440: 467-476.

8 Pfendler S, Einhorn O, Karimi B, et al. UV-C as an efficient means to combat biofilm formation in show caves: evidence from the La Glacière Cave (France) and laboratory experiments[J]. Environmental Science and Pollution Research, 2017, 24: 24611-24623.

9 Yan L, Hong W, Lu X F. Isolation and molecular identification of mold on collection paintings, calligraphy, and silk relics[J]. Identification and Appreciation to Cultural Relics, 2021, (10): 60-65.

10 Strycker B D, Han Z, Duan Z, et al. Identification of toxic mold species through Raman spectroscopy of fungal conidia[J]. PloS one, 2020, 15(11): e0242361.

11 Vashpanov Y, Heo G, Kim Y, et al. Detecting green mold pathogens on lemons using hyperspectral images[J]. Applied Sciences, 2020, 10(4): 1209.

12 Chun S W, Song D J, Lee K H, et al. Deep learning algorithm development for early detection of Botrytis cinerea infected strawberry fruit using hyperspectral fluorescence imaging[J]. Postharvest Biology and Technology, 2024, 214: 112918.

13 Zou Z, Zhen J, Wang Q, et al. Research on nondestructive detection of sweet-waxy corn seed varieties and mildew based on stacked ensemble learning and hyperspectral feature fusion technology[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2024, 322: 124816.

14 Siripatrawan U, Makino Y. Assessment of food safety risk using machine

learning-assisted hyperspectral imaging: Classification of fungal contamination levels in rice grain[J]. Microbial Risk Analysis, 2024, 27: 100295.
15 Lu M, Zhang Q, Li Y, et al. Research on Labeling Method for Fungal Diseases in Murals Based on Hyperspectral Images[C]//2024 3rd International Conference on Image Processing and Media Computing (ICIPMC). IEEE, 2024: 93-100.
16 Wang, Sa, et al. "Virtual restoration of ancient mold-damaged painting based on 3D convolutional neural network for hyperspectral image." Remote Sensing 16.16 (2024): 2882.
17 Hong D, Zhang B, Li X, et al. SpectralGPT: Spectral remote sensing foundation model[J]. IEEE transactions on pattern analysis and machine intelligence, 2024, 46(8): 5227-5244.
18 Roy S K, Deria A, Hong D, et al. Multimodal fusion transformer for remote sensing image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-20.
19 Zhan Y, Yang R, You J, et al. A systematic literature review on incomplete multimodal learning: techniques and challenges[J]. Systems Science & Control Engineering, 2025, 13(1): 2467083.
20 Saidi S, Idbraim S, Karmoude Y, et al. Deep-learning for change detection using multi-modal fusion of remote sensing images: A review[J]. Remote Sensing, 2024, 16(20): 3852.
21 Xing S, Wang Y. Cross-Modal Attention Networks for Multi-Modal Anomaly Detection in System Software[J]. IEEE Open Journal of the Computer Society, 2025.
22 Hu W, Huang Y, Wei L, et al. Deep convolutional neural networks for hyperspectral image classification[J]. Journal of Sensors, 2015, 2015(1): 258619.
23 Zhao W, Du S. Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(8): 4544-4554.
24 Kanthi M, Sarma T H, Bindu C S. Multi-scale 3D-convolutional neural network for hyperspectral image classification[J]. Indonesian Journal of Electrical Engineering and Computer Science, 2022, 25(1): 307-316.
25 Zhou H, Zhang X, Zhang C, et al. Vision transformer with contrastive learning for hyperspectral image classification[J]. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 1-5.
26 Sun L, Zhao G, Zheng Y, et al. Spectral–spatial feature tokenization transformer for hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-14.
27 Hong D, Han Z, Yao J, et al. SpectralFormer: Rethinking hyperspectral image classification with transformers[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-15.

Figure legends

Fig. 1 HNFT Network Structure Diagram

Fig. 2 Structure Diagram of Modal Absence Configurator

Fig. 3 Structure Diagram of Modal Alignment Module

Fig. 4 Structure Diagram of Gated Cross-Attention Mechanism

Fig. 5 Simulated mold spot samples of black oak wood. (a) False color image. (b) Mark truth image.

Fig. 6 Hyperspectral mold spot image acquisition system

Fig. 7 Short-Wave Infrared spectroscopy image acquisition system

Fig .8 (a) Penicillium citrinum (b) Aspergillus niger (c) Trichoderma longibrachiatum multi-modal sample images

Fig. 9 Comparison chart of OA for different proportions of training samples

Fig. 10 Classification result graph. (a) Real value graph. (b) 1D-CNN (OA = 72.07%). (c) 2D-CNN (OA = 81.63%). (d) 3D-CNN (OA = 85.25%). (e) ViT (OA = 84.87%). (f) SSFTT (OA = 92.61%).(g) Spectralformer (OA = 94.78%). (h) HNFT (OA = 98.51%)