Application of Multispectral Image Analysis Based on Transformer Model in the Restoration of Fine Fiber Materials in Ancient Murals

Ning Wang
College of Architecture and Art, Taiyuan University of Technology, Jinzhong 030600, Shanxi, China
Corresponding author: Ning Wang (e-mail: jzwxgc_007@163.com)

ABSTRACT Accurate identification of micro-damage in fiber-based mural materials is essential for quantitative restoration and non-destructive cultural heritage conservation. Since multispectral imaging exploits the interaction between electromagnetic waves and material surfaces to reveal spectral responses that are invisible in conventional imaging, it provides an effective foundation for fine structural analysis. This study proposes a Transformer-based multispectral image analysis framework for the restoration of ancient mural fiber materials. The method constructs a spectral–spatial joint representation through channel-level fusion and positional encoding and employs multi-head self-attention to capture weak cross-band fracture features and long-range contextual dependencies. An end-to-end decoding strategy with cross-scale feature fusion is further designed to achieve pixel-level segmentation and quantitative characterization of fiber microcracks. Experiments conducted on 128 multispectral mural images demonstrate an Intersection over Union of 90.2% for longitudinal crack detection, a damage density estimation error of 0.05, and over 88.3% of crack width predictions with an absolute error below 0.15 mm. The proposed framework enables reliable nondestructive assessment of mural fiber degradation while providing a transferable spectral analysis strategy for electromagnetic imaging, intelligent material inspection, and related engineering applications involving multispectral sensing.

INDEX TERMS multispectral image analysis, vision transformer, spectral–spatial feature fusion, fiber damage detection, electromagnetic imaging

I. INTRODUCTION

THE fiber-based materials in ancient murals are exposed to complex environments for long periods, often developing microcracks, fractures, and loose structures that weaken stability and accelerate deterioration. Among these, microcracks and fractures are the clearest signs of fiber degradation, as their propagation leads to loosening and decay [1,2]. Analyzing crack morphology and density provides essential quantitative indicators for evaluating deterioration and guiding restoration. Scientific diagnosis before restoration requires high-resolution, non-destructive imaging for quantitative assessment [3]. Multispectral images, with crossband information, reveal differential fiber responses under various wavelengths, offering a richer basis for microdamage identification [4].

However, mural fiber microcracks usually exhibit weak texture, low contrast, and blurred boundaries, with inconsistent multi-band responses that hinder spatial–spectral coupling. High-dimensional data contain noise and redundancy [5,6]. Traditional edge detection, filtering, and shallow convolution methods struggle to extract deep texture, often causing false or missed detections. Convolutional Neural Network (CNN)- and U-Net-based studies have achieved initial progress but remain limited to single-channel or pseudo-color inputs, lacking effective inter-band texture modeling [7,8]. Although multi-scale and attention mechanisms improve complex texture processing, they still fail to capture weak cross-band responses, restricting precise microdamage identification [9,10]. Therefore, a method with global modeling capacity and integrated spectral information is still needed [11,12].

To overcome discontinuous features and unstable localization in multispectral images, mural bands are organized into a wavelength-ordered tensor. Fixed-size blocks are linearly embedded with positional encoding to retain spatial consistency. A multi-layer self-attention mechanism models weak cross-band dependencies and long-range context, while skip connections and decoder layers fuse features to produce pixel-level segmentation and direction maps. This approach achieves fine-scale damage modeling, improves diagnostic

accuracy, and supports restoration with quantitative structural data.

II. RELATED WORKS

Multispectral imaging has become essential for nondestructive mural damage detection due to its differential response to surface and subsurface features. Wu et al. [13] proposed a Multi-scale Boundary and Region Feature Fusion model integrating boundary–region features and global–detail generators, achieving precise segmentation of Dunhuang and Yunnan murals. Its reverse attention mechanism enhanced crack responses and reduced background noise, inspiring the Transformer encoding used here. Adamopoulos et al. [14] built a supervised system based on multispectral reflectance imaging, using regression, decision tree, and ensemble methods to annotate damaged regions, motivating this study to retain band characteristics for differential modeling. Es Sebar et al. [15] developed a low-cost 3D reconstruction scheme combining visible-reflected, UV-induced luminescence, UV-reflected imaging, and photogrammetry to generate a unified multichannel radiation model for relic microstructure analysis. Despite this progress, cross-band weak feature modeling and deep texture representation remain limited under heterogeneous backgrounds [16,17]. Transformer networks show strong capability in modeling high-resolution details, long-range dependencies, and multi-scale features. Zhou et al. [18] proposed a Dam Crack Segmentation Transformer with an improved Swin module and weighted attention, achieving fine segmentation under low contrast and complex backgrounds. Shamsabadi et al. [19] combined Transformer and encoder–decoder architectures in Transformer U-Net, outperforming CNNs under strong texture noise. Ye et al. [20] integrated ResNet101 and Transformer modules in a Fissure Network, improving segmentation accuracy under illumination and structural variation. These methods perform well in engineering crack detection but lack spectral consistency modeling across multi-channel images, limiting their applicability to cultural heritage microstructure analysis.

III. MODEL METHOD DESIGN

MULTISPECTRAL IMAGE PREPROCESSING

The original multispectral mural images differ in band number, spectral range, and resolution, requiring standardization before model input. Illumination correction and dark current removal are applied to all channels, followed by normalized least squares fitting to align spectral responses across images. After band reordering by central wavelength, normalization and mirror padding ensure consistent spectral alignment. Principal Component Analysis (PCA) is then used for redundancy reduction with a fixed compression ratio of 0.5, retaining 15 principal components. Ratios below 0.4 cause weak spectral loss, while those above 0.6 increase tensor volume without accuracy gain. The selected ratio preserves spectral–spatial correlation and supports stable model convergence. The processed data are finally reconstructed into a three-dimensional tensor $I \in \mathbb{R}^{H\times W\times C}$, where H and W denote image height and width, and C represents the unified number of spectral bands.

In view of the redundancy and noise caused by the high channel dimension in multispectral images, a linear projection function set by the principal component preservation rate is used to perform spectral compression on the input tensor. The process is formalized as minimizing the following objective function:

$$P = \arg\min_{P} \left\| I - IPP^{\top} \right\|_F^2 \tag{1}$$

In formula (1), $P \in \mathbb{R}^{C\times C'}$ retains the first C' principal components after compression to form a compact representation, satisfying $C' \ll C$. The pixel values in the spatial dimension are uniformly normalized to the interval [0, 1]. The numerical distribution of the entire image tensor is standardized through batch normalization to ensure that each channel has equal importance in subsequent calculations.

After the tensor is constructed, combined with the recognition requirements of the region of interest, the image edge area is filled with a mirror image to unify the image size, and the boundary cropping errors in some images are linearly interpolated based on adjacent valid spectral regions and accompanied by a binary mask to exclude these edge areas from subsequent feature learning. Finally, the input data format with consistent shape and standardized physical band order is generated:

$$\hat{I} = \mathrm{Compress}(\mathrm{Pad}((\mathrm{Norm}I))) \tag{2}$$

In formula (2), Compress represents spectral compression; Norm represents normalization operation; Pad represents mirror padding and mask processing. In the process definition of Equation 2, Norm and Pad are first executed to ensure that each image has a uniform statistical distribution and shape in the band and spatial dimensions. Then Compress is executed to perform principal component dimensionality reduction on the aligned physical band sequence. To analyze the impact of the spectral compression ratio on the input tensor structure, different principal component retention ratios are set, and key parameters such as the number of channels and tensor volume are measured. The tensor volume is defined as $H \times W \times C'$, where H and W denote the spatial dimensions (256 $\times$ 256 pixels), and C' denotes the number of channels after spectral compression. For numerical readability, the total element count is scaled by 10^6 and presented in Table 1 as "Tensor volume ($\times 10^6$)." This scaling does not alter the relative magnitude of tensor variation but ensures a consistent representation across compression ratios. The value of C' corresponds to the fixed number of principal components retained under each compression ratio, representing a constant channel dimensionality defined by the PCA compression process. As the principal component retention ratio increases, the number of image channels increases significantly, and the input

tensor scale expands linearly, imposing higher demands on memory allocation and computational stability during model training.

TABLE 1. Statistics of model input tensors under different spectral compression ratios

Compression ratio (C'/C)	Number of channels C'	Image size ($H \times W$, pixels)	Tensor volume ($H \times W \times C'$, $\times 10^6$ elements)
0.2	6	256 × 256	0.39
0.3	9	256 × 256	0.59
0.4	12	256 × 256	0.78
0.5	15	256 × 256	0.98
0.6	18	256 × 256	1.18

IMAGE BLOCK DIVISION AND FEATURE EMBEDDING

The multispectral image tensor $\hat{I} \in \mathbb{R}^{H\times W\times C'}$ after preprocessing is regarded as a high-dimensional data structure with uniform spectral order and spatial resolution. To adapt to the input requirements of the Transformer architecture, the image needs to be divided into several non-overlapping sub-regions of fixed size in the spatial dimension. Let the partition window size be $P \times P$, a two-dimensional sliding window operation is performed on the tensor, and a local image block sequence is constructed. Each image block $x_i \in \mathbb{R}^{P\times P\times C'}$ represents the local texture–spectral structure. By linearly transforming $\phi : \mathbb{R}^{P\times P\times C'} \to \mathbb{R}^D$ block by block, the three-dimensional sub-blocks are mapped to vector representations of fixed dimensions, thereby constructing an embedding sequence $\{z_i\}_{i=1}^N$ of length $N = \frac{HW}{P^2}$, where D is the feature dimension required by the Transformer encoder.

To ensure that spatial information remains distinguishable during sequence modeling, absolute spatial position encoding needs to be injected into each image block position. The coordinate embedding matrix $E_{\text{pos}} \in \mathbb{R}^{N\times D}$ generated by the two-dimensional sine–cosine function is element-wise summed with the initial embedding sequence to form the final input sequence:

$$Z = \{z_i + E_{\text{pos},i}\}_{i=1}^N \tag{3}$$

Positional encoding introduces spatial patterns of different frequencies in each dimension, effectively enhancing the model's ability to perceive the local structure position and improving the modeling accuracy of the position distribution of the fracture area.

To suppress the problem of uneven value distribution across blocks in the input embedding, residual normalization is used to normalize the embedding vector so that each image block embedding satisfies the following constraints:

$$\tilde{z}_i = \frac{z_i - \mu_z}{\sigma_z}, \mu_z = \frac{1}{N}\sum_{i=1}^N z_i, \sigma_z = \sqrt{\frac{1}{N}\sum_{i=1}^N (z_i - \mu_z)^2} \tag{4}$$

The standardized input sequence $\tilde{Z} \in \mathbb{R}^{N\times D}$ significantly improves the numerical stability of the model in the early stage of training, avoids gradient oscillation caused by the drift of embedded values in the deep structure, and enhances the separability of multi-scale textures in the initial input stage.

This experiment evaluates the effect of image partition granularity on model performance by varying block size and measuring the total number of blocks, reconstruction error, and inference time (Table 2). Finer partitions reduce reconstruction error but sharply increase inference cost, reflecting a trade-off between block density and computational efficiency. Reconstruction error is quantified by mean squared error (MSE) within the normalized input tensor, calculated as the average squared pixel deviation between reconstructed and reference tensors after intensity normalization to [0, 1], ensuring a consistent evaluation scale across partition settings.

TABLE 2. Analysis of embedding number and reconstruction error at different image partitioning granularities

Image block size (pixels)	Total number of image blocks N	Reconstruction error Mean Squared Error (normalized)	Inference time (ms)
32 × 32	64	0.016	45
16 × 16	256	0.011	68
8 × 8	1024	0.007	124
4 × 4	4096	0.005	207
2 × 2	16384	0.004	392

SPECTRAL–SPATIAL JOINT ENCODING MODULE

The tensor $\tilde{Z} \in \mathbb{R}^{N\times D}$ formed by the embedded input sequence is sent to the spectral–spatial joint encoding module for cross-band long-range texture modeling and local fracture feature capture. This module adopts a stacked multi-layer Transformer encoder structure, and each layer contains a multi-head self-attention sublayer and a feedforward network sublayer, allowing extraction of global context and cross-channel detail response. To enhance the model's perception of fine-grained fracture textures between spectral bands, a multi-head attention mechanism is introduced to construct a cross-channel correlation matrix, and dynamic weighted modeling is performed on each position embedding. Its self-attention mechanism is defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{5}$$

In formula (5), $Q = \tilde{Z}W^Q$, $K = \tilde{Z}W^K$, $V = \tilde{Z}W^V$, $W^Q, W^K, W^V \in \mathbb{R}^{D\times d_k}$ are trainable parameters, and d_k is the attention dimension of each head.

Layer normalization and residual connections are inserted between the outputs of each layer to suppress gradient dissipation, so that the deep encoder has stronger stability and semantic retention ability. The feedforward network of each layer of the Transformer contains two linear transformations and an activation function, and its structure is defined as:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{6}$$

In formula (6), $W_1 \in \mathbb{R}^{D\times D_{\text{ff}}}$, $W_2 \in \mathbb{R}^{D_{\text{ff}}\times D}$, D_{ff} are the intermediate dimensions of the feedforward network,

and the activation function uses ReLU to achieve nonlinear transformation. The model gradually builds a hierarchical representation of the image crack structure based on the output of each encoder layer and maintains the ability to restore original texture details through the fusion of information residuals between layers. In the tensor evolution of the modeling process, a set of intermediate state tensors $\{H^l\}_{l=1}^{L}$ are introduced to represent the encoding output of each layer, and their evolution relationship is as follows:

$$\begin{aligned} H^l &= \mathrm{TransformerLayer}(H^{l-1}) \\ &= \mathrm{FFN}(\mathrm{Attention}(H^{l-1})) \end{aligned} \tag{7}$$

The initial state is set to $H_0 = \tilde{Z}$. This structure ensures that the model can respond to local micro-texture differences while maintaining the global receptive field, especially improving discrimination accuracy near fracture and non-fracture boundaries.

To demonstrate the hierarchical organization and feature modeling path between the modules of this method in the multispectral image processing task, Figure 1 presents the complete structural relationship of the model from input construction, spectral alignment, compression, embedding conversion, cross-band encoding, to decoding output. The functional modules of each stage are organized hierarchically, and the mapping and transmission mechanisms between multispectral images in preprocessing, image block embedding, Transformer encoding, feature fusion, and multi-branch output are described from bottom to top, reflecting the overall process and local dependency coupling mode for modeling microcrack structures in the spectral–spatial dimension. The preprocessing stage is organized as a sequential pipeline, starting with illumination correction and dark current removal, followed by band reordering according to the central wavelength of each band, channel-level normalization, and mirror fill to unify the statistical distribution and spatial shape, and then performing principal component compression on the aligned physical band sequence to reduce spectral redundancy while maintaining spectral consistency across images.

CROSS-SCALE FEATURE FUSION AND DECODING

The spectral–spatial joint encoding process in the model backbone is based on the Transformer structure, and its multi-layer encoder establishes long-range contextual dependencies and cross-band texture responses on the embedded sequence. Figure 2 shows the feature evolution path of the image block after the multi-head attention mechanism, residual fusion, and nonlinear mapping in the Transformer encoder. Each layer's encoding output strengthens the directionality and continuity of the fractured texture while maintaining spatial order consistency, providing multi-scale semantic support for the subsequent decoding stage.

The high-dimensional global feature tensor from the Transformer encoder preserves rich spectral–spatial semantics but loses local detail and continuity. To achieve pixel-level segmentation of microcracks, shallow texture and deep semantic features are integrated through an end-to-end cross-scale decoding structure. A cascaded decoder with multi-level skip connections aligns and fuses intermediate features, while upsampling and convolution modules perform spatial expansion, channel compression, and shallow feature fusion to enhance edge restoration.

Assuming that the decoded output of the lth layer is $F^l \in \mathbb{R}^{\frac{H}{2^l} \times \frac{W}{2^l} \times D_l}$, it performs channel alignment and element-by-element fusion with the corresponding output $E^l \in \mathbb{R}^{\frac{H}{2^l} \times \frac{W}{2^l} \times D_l}$ of the encoder's lth layer, which is defined as:

$$\hat{F}^l = \sigma(\mathrm{BN}(\mathrm{Conv}_{1\times1}(F^l + E^l))) \tag{8}$$

In formula (8), $\mathrm{Conv}_{1\times1}$ represents the channel-by-channel convolution operation; BN is the batch normalization operation; σ is the activation function. The fusion result is then input into the upsampling module to restore higher spatial dimensions. To enhance the decoder's sensitivity to local detail changes, the residual path and spatial attention mechanism are introduced at each level to construct a locally sensitive feature response map, further improving edge positioning accuracy and weak structure restoration capabilities.

The upsampling process adopts a parallel fusion strategy of deconvolution and nearest neighbor interpolation to double the spatial dimension and maintain a smooth transition of feature distribution. Assuming the target upsampling ratio is r and the reconstructed image size is $H' = rH$, $W' = rW$, the upsampling output tensor is expressed as:

$$U = \mathrm{Upsample}(F^l) = \mathrm{Concat}(\mathrm{Deconv}(F^l), \mathrm{Interp}(F^l)) \tag{9}$$

The dual-path upsampling structure effectively reduces grid artifacts introduced by deconvolution and enhances edge smoothness through interpolation compensation. In the final stage of decoding, the output tensor is mapped to a single-channel probability map and compressed into a normalized damage response map through a sigmoid function, which serves as the basis for subsequent crack segmentation and direction estimation.

MICROCRACK AND DIRECTION MAP OUTPUT MODULE

After completing the spatial information restoration, the high-resolution feature map $F_{\mathrm{seg}} \in \mathbb{R}^{H\times W\times C}$ output by the decoder needs to be further mapped into a continuous response map with crack heat, fiber density, and direction characteristics. This module introduces a multi-branch structure while maintaining spatial consistency. Each branch designs a dedicated convolution mapping path for different output targets to ensure that the discrimination signal of the fracture area and structural direction is not weakened. The crack heatmap prediction branch uses a double-layer convolution structure with superimposed sigmoid mapping

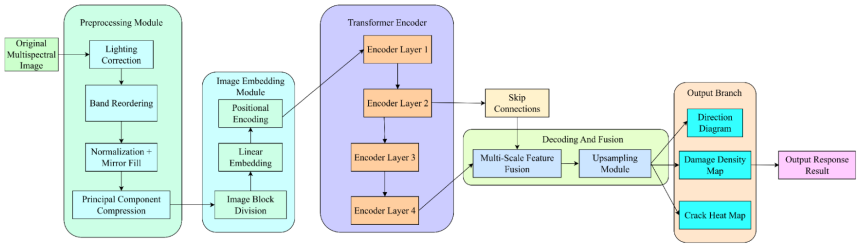

FIGURE 1. Hierarchical structure diagram of multispectral image analysis under Transformer architecture

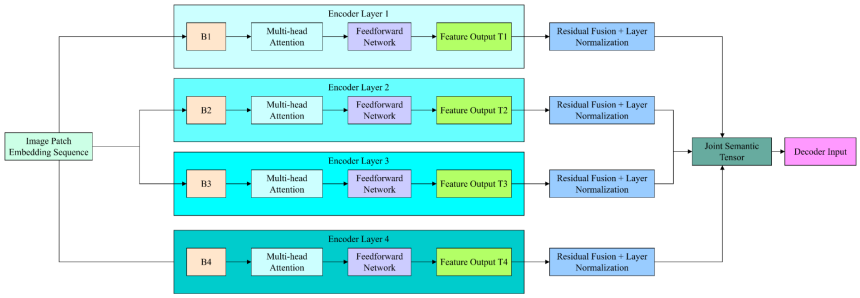

FIGURE 2. Schematic diagram of the image block feature evolution process in the spectral–spatial joint encoder

to form a pixel-level probability map. The output tensor $P_{\text{crack}} \in \mathbb{R}^{H\times W}$ represents the response intensity of the fracture area. The prediction function can be expressed as:

$$P_{\text{crack}}(x,y) = \sigma(W_2 * \delta(W_1 * F_{\text{seg}}(x,y))) \tag{10}$$

In formula (10), W_1, W_2 are learnable convolution kernels; δ is the activation function; $*$ represents the convolution operation. The damage density map is generated by the local statistics module, calculating the weighted sum of the thermal response in each fixed window $\Omega_i \subset \mathbb{R}^{s\times s}$ and constructing the density response tensor $D_{\text{crack}} \in \mathbb{R}^{H'\times W'}$, which is defined as follows:

$$D_{\text{crack}}(i)=\frac{1}{|\Omega_i|}\sum_{(x,y)\in\Omega_i} P_{\text{crack}}(x,y) \tag{11}$$

This method depicts local damage clustering areas while maintaining the smoothness of the overall response. Direction map prediction is based on feature field gradient regression, extracting the dominant texture direction of each pixel, using the direction regression module to output the angle tensor $\theta_{\text{pred}} \in \mathbb{R}^{H\times W}$, and learning the direction angle by minimizing the direction consistency loss. Its loss function is set as:

$$L_{\vartheta}=\frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W}\left[1-\cos(\vartheta_{\text{pred}}(x,y)-\vartheta_{\text{gt}}(x,y))\right] \tag{12}$$

This structure effectively reduces the gradient discontinuity problem caused by angle ambiguity in direction prediction and improves the stability of the main axis judgment of the fracture texture. The parameters between branchesshare the underlying feature representation, and throughindependent target optimization, complementary perception is formed to construct the output domain of multi-angle fiber damage expression.

IV. EXPERIMENTAL SETUP AND DATA PREPARATION

Experiments were conducted on a GPU-based workstation using the PyTorch framework. Model optimization adopted AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 1×10^{-4}) with an initial learning rate of 1×10^{-3} decayed to 1×10^{-6} via cosine annealing. Training used a batch size of 16 for 200 epochs with early stopping after 15 stagnant epochs. Gradient clipping (threshold 1.0) stabilized optimization, and Xavier plus truncated normal initializations were applied to CNN and Transformer layers. Cross-entropy and Dice losses were combined for structural consistency, while random rotation ($\pm10°$), mirroring, and $\pm3\%$ spectral jitter enhanced spectral–spatial diversity and reduced overfitting.

The dataset comprises 128 multispectral mural images obtained with a calibrated band-scanning system. Geometric calibration used a 1 mm grid aluminum plate and least-squares fitting to map pixels to physical dimensions, achieving a mean reprojection error of 0.036 mm. At a working distance of 0.8 m, the spatial resolution reached 0.046 mm/px based on a 118 mm $\times$ 118 mm field of view and 2560 $\times$ 2560 px image size. Validation using a micrometer ruler confirmed deviations below 0.05 mm, ensuring traceable millimeter-scale precision for crack width measurement.

The acquisition covers 24 spectral bands ranging from 420 nm to 950 nm, each with a full width at half maximum of 20 nm. The two sites differ in substrate material composition, pigment mineral ratio, and surface aging condition, forming a natural spectral variability for cross-site generalization analysis. The images from the primary site are used for model training and validation, while those from the secondary site serve as an independent extrapolation set

to examine the stability of spectral–spatial encoding under heterogeneous materials and aging patterns. The center wavelength of every band is determined by monochromator-based calibration to ensure spectral accuracy. During image capture, the base exposure duration is maintained at 120 ms under standard illumination. The camera system employs an adaptive electronic integration gain that adjusts according to the measured surface reflectance to achieve uniform radiometric response without altering the physical exposure duration. The average signal-to-noise ratio of the system is 62 dB under the controlled illumination intensity. The imaging geometry follows a perpendicular configuration with the optical axis normal to the mural surface at a working distance of 0.8 m. Illumination is provided by two diffuse halogen sources positioned symmetrically at a 45° incidence angle to minimize specular reflection. The original multispectral images were acquired at a resolution of 2560 $\times$ 2560 pixels and geometrically calibrated to ensure alignment between the physical acquisition scale and the model input tensor format. After spectral and spatial normalization, each image is downsampled to 256 $\times$ 256 pixels to form the standardized input size of the model. This downsampled resolution provides a standardized spatial scale for network input and stabilizes spectral–spatial feature alignment. Although some high-frequency microstructural details are reduced during resampling, the essential morphological patterns of cracks and fiber directions are retained through multispectral feature reconstruction and positional encoding in the model. The image size is consistent with the model input structure to ensure that no interpolation error is introduced during subsequent modeling. After preprocessing and fixed-window partition, each 256 $\times$ 256 image is divided into non-overlapping 16 $\times$ 16 pixel patches, yielding 256 patches per image and 32,768 patches total across 128 images, which are used as the effective training and validation samples.

The division of training, validation, and testing data is strictly performed at the image level prior to patch generation. The original 128 images are first partitioned into independent subsets, and no patch extracted from a single mural image appears simultaneously in multiple subsets. All patches used for training, validation, and testing are generated exclusively within their respective image groups, ensuring that model evaluation is free from data leakage and overestimation of performance metrics. The scale of the patch-level dataset is chosen to balance the requirement of fine-grained texture modeling and computational stability. The image is processed by illumination correction, normalization, and rearrangement to construct a high-dimensional tensor of unified format and is divided into non-overlapping sub-blocks embedded in a sequence according to a fixed window. Manual annotation is completed by professionals based on the multi-band response results to complete the crack delineation. This annotation process is performed entirely on multispectral image data without any direct intervention on the mural surface, ensuring that the construction of ground truth does not alter or damage the original material. The annotation content covers crack direction, width range, and spatial position. The label format is standardized to support the direction regression and density statistics module call. The training set, validation set, and test set are divided at the image level with a ratio of 75:15:10 to ensure an even distribution of crack types across the validation subset. The partitioning procedure guarantees that all spectral–spatial data derived from a single image remain confined to the same subset, maintaining complete independence among training, validation, and testing groups. The spectral and spatial masks are retained consistently within each subset to preserve the continuity of cross-band texture modeling without any overlap between sets. The evaluation uses intersection over union ratio, width absolute error, and direction recognition accuracy, while the density estimation error is calculated by regional response normalization deviation. The regional response normalization deviation (RRND) serves as the quantitative metric for evaluating the accuracy of density estimation. It is mathematically defined as the square root of the mean of the squared difference between the normalized local response and the mean response within the same evaluation region. In the weighted formulation, each pixel response within the region is assigned a relative contribution weight w_i, and the weighted mean deviation is computed as:

$$E_{RRND} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} w_i (R_i - \bar{R})^2}, \quad w_i = \frac{R_i}{\sum_{j=1}^{N} R_j} \tag{13}$$

In formula (13), where R_i denotes the normalized response value of pixel i, $\bar{R}$ represents the regional mean, w_i is the response-based normalization weight reflecting the relative intensity contribution of each pixel within the evaluation region, and N is the total number of pixels in the evaluated area. The E_{RRND} value is non-negative and decreases toward zero as the predicted and reference distributions become consistent. Its magnitude is constrained by the reciprocal square root of the pixel number within the evaluation region, reflecting the normalized nature of this weighted deviation metric. A value of 0 represents perfect alignment, while a value near 1 denotes severe deviation. RRND focuses on local density uniformity and complements global error indicators such as mean absolute error (MAE), root mean square error (RMSE), and peak signal-to-noise ratio (PSNR). The subsequent section presents comparative evaluations using these indicators to demonstrate the sensitivity and reliability of RRND in microcrack density analysis.

For the CNN model used in the damage density estimation error comparison, a baseline three-layer convolutional neural network is constructed. Each layer contains 32, 64, and 128 filters, respectively with 3 $\times$ 3 convolution kernels, followed by batch normalization and ReLU activation. Max-pooling layers with a stride of 2 are applied after the first

two convolutional layers, and a fully connected layer with 256 units is appended before the final regression layer. The network parameters are initialized using Xavier initialization and optimized with Adam at a learning rate of 0.001. To ensure fairness, the CNN model is trained under the same dataset division and preprocessing pipeline as the Transformer model, and its convergence is monitored with early stopping based on the validation loss. All models are compared uniformly under the same data partitioning and preprocessing conditions to verify the modeling stability and performance advantages of the Transformer structure in the microcrack recognition task. For the U-Net + CNN model, the CNN module refers to a three-layer convolutional block with 3×3 kernels, batch normalization, and ReLU activation, which is connected to the U-Net output to strengthen feature aggregation before classification. For ResNet with Conditional Random Field (ResNet + CRF), a ResNet-50 backbone is modified in its first convolutional layer to accept multispectral channels through weight replication along the channel dimension, and the extended kernel weights are retrained jointly with the network to restore spectral response balance. All baseline networks are subjected to identical preprocessing and normalization, where each input tensor is standardized to zero mean and unit variance on a per-channel basis to eliminate inter-band statistical bias. For DeepLabv3+ and other encoder–decoder architectures, the first convolutional kernel group is expanded to match the multispectral channel dimension, and all network parameters are initialized using variance-preserving initialization prior to training from scratch on the multispectral dataset. After adaptation, the total parameter counts and computational complexities are quantified to ensure comparability. The ResNet + CRF model comprises 25.6 million parameters with 41.2 GFLOPs, the U-Net + CNN model comprises 18.3 million parameters with 29.7 GFLOPs, DeepLabv3+ comprises 27.9 million parameters with 38.4 GFLOPs, and the proposed Transformer model comprises 30.2 million parameters with 42.8 GFLOPs. The adaptation process maintains a comparable magnitude of parameter scale and computational load across models, ensuring a fair evaluation of multispectral feature modeling performance.

Samples of different crack types are weighted according to their pixel proportions in 128 images in the Intersection over Union (IoU) and error statistics, and index offset correction is performed in combination with the label continuity in the validation set. Table 3 summarizes the weight ratio of each crack type in the test set according to the number of pixels.

TABLE 3. Statistical table of weight distribution of crack type samples in the test set

Crack types	Sample weight	Frequency in the test set (pictures)	Total number of annotated pixels (millions)
Longitudinal cracks	0.28	15	0.82
Transverse cracks	0.24	12	0.71
Oblique cracks	0.18	10	0.55
Local cracks	0.22	11	0.68
Random cracks	0.08	5	0.34

V. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

MICROCRACK RECOGNITION ACCURACY

This experiment compares model performance in microcrack recognition and examines the influence of crack size and type on segmentation accuracy. Three models are evaluated: Transformer, U-Net, and DeepLabv3+. Crack sizes are divided into five ranges (0–1 mm, 1–2 mm, 2–3 mm, 3–4 mm, and >4 mm), and types include longitudinal (C1), transverse (C2), oblique (C3), local (C4), and random (C5) cracks. The Transformer integrates self-attention and cross-band feature extraction for fine and complex textures, U-Net captures local details through its encoder–decoder design, and DeepLabv3+ applies dilated convolution with a feature pyramid for multi-scale analysis. Figure 3 illustrates IoU scores across crack sizes and types, revealing each model's relative strengths.

The Transformer model in this paper performs better than the other two comparison models for fine cracks (such as 0–1 mm) and complex crack types (such as oblique and local cracks). In the identification of fine longitudinal cracks, the IoU score reaches 90.2%. Although U-Net performs better on larger cracks (such as 2–3 mm and above), it is inferior when dealing with microcracks and complex cracks, with a highest IoU of 86.2%. DeepLabv3+ has relatively stable performance on large crack sizes (such as 3–4 mm and >4 mm), but its performance declines for the segmentation of smaller crack sizes, with the highest IoU being 85.3%. The model in this paper performs well on microcracks and complex crack types, indicating significant advantages in processing fine textures and extracting cross-band features.

ERROR ANALYSIS OF DAMAGE DENSITY ESTIMATION

Traditional image processing struggles to identify complex cracks, while deep learning has achieved notable advances. To ensure a fair comparison in Figure 4, the baseline CNN uses a lightweight architecture with convolutional and pooling layers for hierarchical feature extraction and a regression head for density estimation, trained under the same preprocessing and parameters as the Transformer. This experiment compares three models—Transformer, CNN, and Support Vector Machine (SVM)—to analyze damage density estimation errors across crack types and training stages. Figure 4 presents the errors of each model in early, middle, and late training stages, where crack type labels C1–C5 correspond to longitudinal, transverse, oblique, local, and random cracks.

In the initial stage, the Transformer model outperforms CNN and SVM on all crack types. The errors on random and longitudinal cracks are 0.15 and 0.12, respectively; the errors of the CNN model are 0.22 and 0.18; those of the SVM are 0.40 and 0.35, respectively. After entering the mid-stage of training, the Transformer model continues to maintain its advantage, with the error significantly reduced to 0.11 and 0.08, while the errors of CNN and SVM remain high, at 0.18 and 0.14, and 0.35 and 0.32, respectively. In

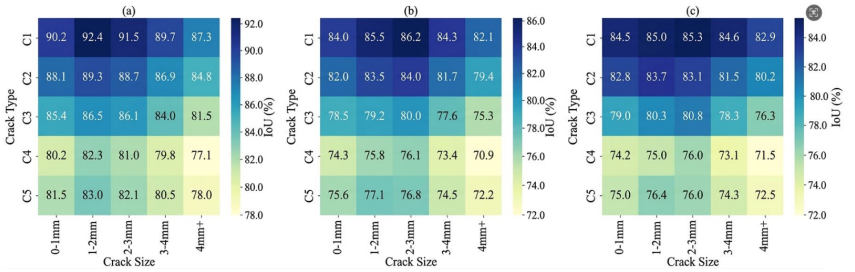

FIGURE 3. Comparison of IoU scores of different models under crack size and crack type (a) Transformer, (b) U-Net, (c) DeepLabv3+

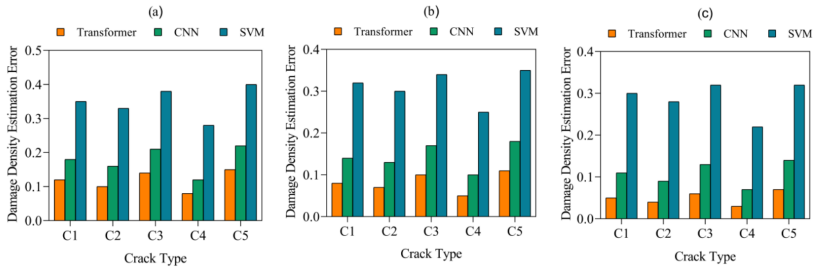

FIGURE 4. Comparison of damage density estimation errors of different models at different training stages (a) Initial stage, (b) Midstage, (c) Late stage

the later stages, the error of the Transformer model is further reduced to 0.07 and 0.05, highlighting its advantages in deep learning, especially in the refinement of crack identification. However, the errors of CNN and SVM converge slowly and remain large in the later stages, indicating their limitations in dealing with complex crack types. Overall, the convergence speed and error reduction of the Transformer model under different crack types are better than those of the other two models, proving its high efficiency and accuracy in microcrack identification and damage density estimation.

To quantitatively compare the performance of regional response normalization deviation with standard evaluation metrics, Table 4 summarizes the RRND, MAE, RMSE, and PSNR results for the Transformer, CNN, and SVM models under the same dataset and training conditions. The Transformer model achieves an RRND of 0.05, an MAE of 0.041, an RMSE of 0.063, and a PSNR of 28.7 dB, all of which are superior to the baseline models. The CNN and SVM models exhibit larger RRND values of 0.11 and 0.18, corresponding to more uneven local response distributions. In summary, the Transformer outperforms the control models in all indicators, demonstrating higher spatial uniformity and estimation accuracy.

TABLE 4. Comparative evaluation of density estimation indicators among models

Model	RRND (↓)	MAE (↓)	RMSE (↓)	PSNR (dB, ↑)
Transformer	0.05	0.041	0.063	28.7
CNN	0.11	0.082	0.121	22.3
SVM	0.18	0.136	0.187	18.6

EVALUATION OF CRACK WIDTH EXTRACTION CAPABILITY

This section evaluates the model's fine-scale response in crack width extraction using annotated and predicted data from 128 multispectral mural images. The error distribution illustrates local deviation trends, while the width comparison verifies linear consistency across continuous intervals. Figure 5 presents the correspondence between error frequency and predicted–true values. Crack widths were converted from pixels to millimeters using the calibrated spatial resolution of 0.046 mm/px defined in Section "EXPERIMENTAL SETUP AND DATA PREPARATION", ensuring that reported errors represent actual geometric deviations.

The error distribution diagram shows that the prediction error is mainly concentrated in the range of -0.1 mm to +0.1 mm, and the frequency in the range of 0.000 to +0.050 mm is the highest, reaching 31 cases, accounting for 24.2% of the total samples. The positive error is slightly higher than the negative error, reflecting that the model has a slight overestimation trend at the wider crack boundary. The absolute value of the sample error for 88.3% of cases is less than 0.15 mm, indicating stable width approximation capability in most crack areas, and the standard deviation of the error is 0.10 mm, reflecting good discrete control in fine-scale prediction. In the width comparison diagram, the predicted and labeled values show a significant linear correlation. The samples are mostly distributed on both sides of the ideal line y = x and are evenly distributed, with no systematic deviation. The fitting line and the reference line have a high degree of overlap, further showing that the model has a stable response mechanism and accurate boundary perception in the continuous width range. Overall, the model has strong consistency and error convergence in

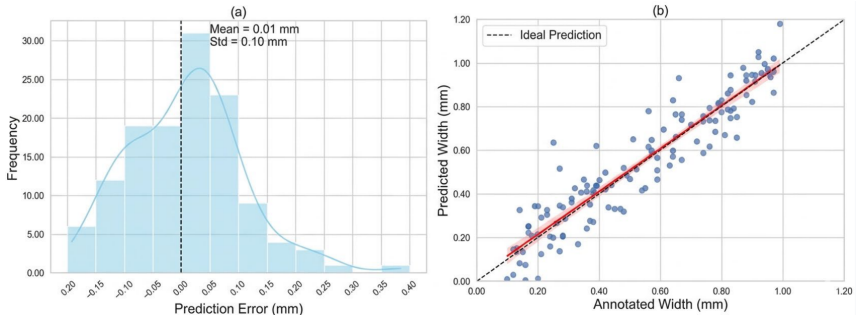

FIGURE 5. Crack width extraction error distribution and prediction consistency evaluation diagram (a) Crack width prediction error histogram, (b) Comparison diagram of model prediction values and manually marked widths

microscale crack width extraction.

To enhance the reliability of the evaluation, the crack width extraction performance of U-Net and DeepLabv3+ is analyzed under the same data partitioning and preprocessing pipeline. The results are summarized in Table 5. Compared with U-Net and DeepLabv3+, the Transformer maintains a higher proportion of accurate predictions within 0.15 mm, achieves the lowest mean absolute error, and exhibits the smallest error dispersion, confirming its advantage in quantitative width extraction.

TABLE 5. Crack width extraction performance comparison among different models

Model	Proportion of samples with absolute error < 0.15 mm	Mean absolute error (mm)	Standard deviation of error (mm)
Transformer	88.3%	0.08	0.10
U-Net	71.6%	0.14	0.17
DeepLabv3+	74.8%	0.12	0.15

FIBER ORIENTATION EXTRACTION ACCURACY

This experiment evaluates the orientation extraction performance of different image segmentation models under various crack types. Six representative deep learning architectures are compared: ResNet + CRF (M1), U-Net combined with Convolutional Neural Network (U-Net + CNN, M2), Densely Connected Convolutional Network (DenseNet, M3), DeepLabv3 Plus (M4), Feature Pyramid Network (FPN, M5), and Segmentation Network (SegNet, M6). Performance variations result from differences in feature extraction and contextual modeling. Directional accuracy is defined as the proportion of pixels whose predicted orientation deviates by less than $\pm 10^\circ$ from the reference, calculated within a normalized circular domain $[-90^\circ, +90^\circ]$ to remove axial symmetry ambiguity. Crack type labels C1–C5 follow the same definitions as in Figure 3 to ensure consistency in annotation and evaluation, as illustrated in Figure 6.

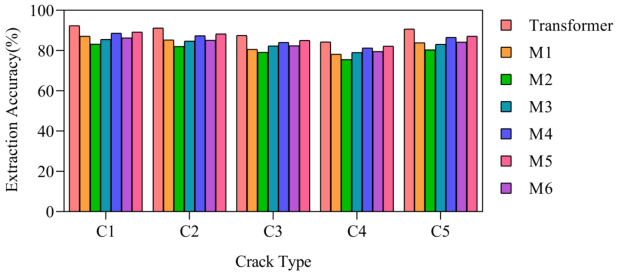

FIGURE 6. Comparison of the direction extraction accuracy of each model under different crack types

The Transformer model shows high accuracy in all crack types, especially in the direction extraction of longitudinal (92.4%) and transverse cracks (91.2%), which is significantly better than other models. ResNet + CRF performs well in these two categories (87.1%, 85.3%) but is weak in identifying oblique and local cracks.

U-Net + CNN (local cracks: 75.5%) and DenseNet (oblique cracks: 82.3%) perform worse for complex cracks. Although DeepLabv3+ and FPN are close to ResNet + CRF in some types, they are still behind the Transformer (90.7%) in random crack recognition. Overall, the Transformer has higher accuracy and adaptability in diversified crack recognition due to its global feature extraction and context modeling capabilities, while other models remain insufficient in complex and fine crack processing.

VI. CONCLUSIONS

This paper proposes a Vision Transformer (ViT)-based multispectral image analysis method for detecting microcracks and fractures in the fiber base of ancient murals. Through channel fusion and positional encoding, the model constructs a spectral–spatial joint representation and employs self-attention for precise segmentation of weak cracks. Experiments achieve 90.2% IoU in longitudinal crack detection, a density estimation error of 0.05, and 92.4% accuracy in direction extraction. The method demonstrates strong accuracy and generalization across materials and aging states, achieving quantitative diagnosis and structural analysis within a non-destructive imaging framework. By correlating microcrack features with fiber degradation, it

provides reliable data for evaluating and reinforcing mural fiber materials.

AUTHOR CONTRIBUTIONS

Ning Wang designed, collected and analyzed the data, and drafted the manuscript. Ning Wang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Ning Wang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] V. Rajakumaran, S. Guessasma, A. D'Orlando, A. Melelli, M. Scheel, T. Weitkamp, et al., "Impact of Defects on Tensile Properties of Ancient and Modern Egyptian Flax Fibers: Multiscale X-Ray Microtomography and Numerical Modeling," Fibers, vol. 12, no. 12, pp. 111, 2024, doi: 10.3390/fib12120111.

[2] A. Parsa Sadr, S. Maraghechi, A. S. J. Suiker, and E. Bosco, "Chemo-mechanical ageing of paper: effect of acidity, moisture and microstructural features," Cellulose, vol. 31, no. 11, pp. 6923-6944, 2024, doi: 10.1007/s10570-024-06005-5.

[3] E. Grifoni, L. Bonizzoni, M. Gargano, J. Melada, N. Ludwig, S. Bruni, et al., "Hyper-dimensional Visualization of Cultural Heritage: A Novel Multi-analytical Approach on 3D Pomological Models in the Collection of the University of Milan," ACM Journal on Computing and Cultural Heritage (JOCCH), vol. 15, no. 2, pp. 34, 2022, doi: 10.1145/3477398.

[4] E. Adamopoulos, C. Colombero, C. Comina, F. Rinaudo, M. Volinia, M. Girotto, et al., "Integrating Multiband Photogrammetry, Scanning, and GPR for Built Heritage Surveys: The Façades of Castello del Valentino," ISPRS Annals of the Photogrammetry Remote Sensing and Spatial Information Sciences, vol. VIII-M-1-2021, pp. 1-8, 2021, doi: 10.5194/isprs-annals-VIII-M-1-2021-1-2021.

[5] X. Ju, X. Zhao, and S. Qian, "TransMF: Transformer-Based Multi-Scale Fusion Model for Crack Detection," Mathematics, vol. 10, no. 13, pp. 2354, 2022, doi: 10.3390/math10132354.

[6] K. Ren, W. Sun, X. Meng, and G. Yang, "MSFDN: multi-scale spatial-spectral-frequency joint denoising network for hyperspectral images," Geo-spatial Information Science, pp. 28(6), 2844–2863, 2025, doi: 10.1080/10095020.2025.2477552.

[7] E. Zhao, N. Qu, Y. Wang, and C. Gao, "Spectral Reconstruction from Thermal Infrared Multispectral Image Using Convolutional Neural Network and Transformer Joint Network," Remote Sensing, vol. 16, no. 7, pp. 1284, 2024, doi: 10.3390/rs16071284.

[8] T. Li and Y. Gu, "Progressive Spatial–Spectral Joint Network for Hyperspectral Image Reconstruction," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-14, 2022, doi: 10.1109/TGRS.2021.3079969.

[9] Q. Gu, H. Luan, K. Huang, and Y. Sun, "Hyperspectral Image Classification Using Multi-Scale Lightweight Transformer," Electronics, vol. 13, no. 5, pp. 949, 2024, doi: 10.3390/electronics13050949.

[10] X. Qiao, S. K. Roy, and W. Huang, "Multiscale Neighborhood Attention Transformer with Optimized Spatial Pattern for Hyperspectral Image Classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-15, 2023, doi: 10.1109/TGRS.2023.3314550.

[11] Y. He, H. Li, M. Zhang, S. Liu, C. Zhu, B. Xin, et al., "Hyperspectral and Multispectral Remote Sensing Image Fusion Based on a Retractable Spatial–Spectral Transformer Network," Remote Sensing, vol. 17, no. 12, pp. 1973, 2025, doi: 10.3390/rs17121973.

[12] X. Qin, H. Song, J. Fan, and K. Zhang, "Spatio-spectral Cross-Attention Transformer for Hyperspectral image and Multispectral image fusion," Remote Sensing Letters, vol. 14, no. 12, pp. 1303-1314, 2023, doi: 10.1080/2150704X.2023.2288069.

[13] X. Wu, Y. Yu, and Y. Li, "A damage detection network for ancient murals via multi-scale boundary and region feature fusion," npj Heritage Science, vol. 13, no. 1, pp. 82, 2025, doi: 10.1038/s40494-025-01546-9.

[14] E. Adamopoulos, "Learning-based classification of multispectral images for deterioration mapping of historic structures," Journal of Building Pathology and Rehabilitation, vol. 6, no. 1, pp. 41, 2021, doi: 10.1007/s41024-021-00136-z.

[15] L. Es Sebar, L. Lombardo, P. Buscaglia, T. Cavaleri, A. Lo Giudice, A. Re, et al., "3D Multispectral Imaging for Cultural Heritage Preservation: The Case Study of a Wooden Sculpture of the Museo Egizio di Torino," Heritage, vol. 6, no. 3, pp. 2783-2795, 2023, doi: 10.3390/heritage6030148.

[16] Y. Li, Y. Yu, and X. Wu, "A dual-encoder hierarchical feature fusion network for ancient mural disease detection," npj Heritage Science, vol. 13, pp. 284, 2025, doi: 10.1038/s40494-025-01844-2.

[17] Z. Xu, X. Zhang, W. Chen, J. Liu, T. Xu, and Z. Wang, "MuralDiff: Diffusion for Ancient Murals Restoration on Large-Scale Pre-Training," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 8, no. 3, pp. 2169-2181, 2024, doi: 10.1109/TETCI.2024.3359038.

[18] J. Zhou, G. Zhao, and Y. Li, "Vison Transformer-Based Automatic Crack Detection on Dam Surface," Water, vol. 16, no. 10, pp. 1348, 2024, doi: 10.3390/w16101348.

[19] E. A. Shamsabadi, C. Xu, and D. Dias-da-Costa, "Robust crack detection in masonry structures with transformers," Measurement, vol. 200, Art. no. 111590, 2022, doi: 10.1016/j.measurement.2022.111590.

[20] T. Ye, M. He, Y. Wang, J. Wang, L. Zhu, and J. Zhang, "An Optimized Intelligent Segmentation Algorithm for Concrete Cracks Based on Transformer," Electronics, vol. 14, no. 9, pp. 1720, 2025, doi: 10.3390/electronics14091720.