IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 1, February, 2026, pp. 159-170
Journal homepage: http://iieta.org/journals/ts

Intelligent Recognition and Life Prediction of Bridge Fatigue Cracks Based on Multimodal Visual Fusion and Spatiotemporal Convolutional Networks

Jianzhen Wu[1], Qi Yang[1], Lifeng Song[1], Qu Wang[2*]

[1] Fujian Vocational College of Agriculture, Fuzhou 350303, China
[2] College of Civil Engineering, Fuzhou University, Fuzhou 350108, China

Corresponding Author Email: wangqu1989@fzu.edu.cn

https://doi.org/10.18280/ts.430112

Received: 2 August 2025
Revised: 20 November 2025
Accepted: 15 December 2025
Available online: 28 February 2026

Keywords:
bridge fatigue cracks, multimodal visual fusion, Graph Convolutional Network, Long Short-Term Memory, 3D Convolutional Networks, structural health monitoring

ABSTRACT

Bridges, as critical components of transportation infrastructure, require early identification of fatigue cracks and accurate life prediction to ensure structural safety. However, existing detection methods suffer from low efficiency, insufficient feature representation in single-modal data, and inadequate modeling of spatiotemporal feature coupling. To address these limitations, this study proposes an intelligent bridge fatigue crack recognition model based on multimodal visual fusion and spatiotemporal convolutional networks. The model innovatively designs a feature transformation unit that embeds a Graph Convolutional Network (GCN) into a Long Short-Term Memory (LSTM) network, enabling the collaborative extraction of spatial topology and temporal dependencies from crack evolution sequences. A dual-stream architecture with independent 3D Convolutional Networks (C3D) is constructed to process spatiotemporal features from RGB and depth video modalities, respectively. During the feature fusion stage, a "cascade fusion–high-dimensional mapping" strategy is employed to enhance the abstract representation capability of multimodal features. Finally, classification results are produced through a Softmax layer, and supervised training is conducted using a multi-class cross-entropy loss function. To evaluate the performance of the proposed model, ablation studies, comparative experiments, and feature fusion analyses are conducted, using both crack recognition and life prediction as evaluation metrics. Experimental results demonstrate that the proposed model achieves a crack recognition accuracy of 96.2% and an F1-score of 95.6%, significantly outperforming conventional approaches. The Mean Absolute Error (MAE) of life prediction is only 28.3 days, representing a reduction of 34.4 days compared with the traditional Paris-law-based method. Moreover, the model exhibits strong robustness in micro-crack detection and under challenging conditions such as varying illumination and occlusion, validating the advantages of the integrated "recognition–prediction" framework. This research provides an efficient and accurate technical solution for bridge fatigue crack detection and promotes the intelligent development of structural health monitoring.

1. INTRODUCTION

Bridges, as the core hubs of the transportation infrastructure network, have structural safety that directly relates to regional traffic accessibility and public safety [1, 2]. Under the coupled effects of long-term loads, environmental corrosion, and material aging, fatigue cracks have become the primary cause leading to sudden failure of bridge structures [3, 4]. According to statistics from the International Association for Bridge Maintenance and Safety, approximately 35% of bridge defects worldwide originate from early fatigue crack propagation that is not detected in time. Such defects not only result in maintenance costs reaching up to 40% of construction costs, but may also cause serious casualties [5, 6]. Therefore, achieving early and accurate identification of fatigue cracks and life trend prediction plays an irreplaceable key role in improving the proactiveness of bridge operation and maintenance and reducing life-cycle costs.

The development of bridge crack recognition technology has undergone an important transition from contact-based detection to non-contact visual detection. Among traditional contact-based methods, ultrasonic testing identifies internal cracks through the propagation characteristics of acoustic waves, but its detection range is limited by the coverage area of the probe [7]. Magnetic particle testing is only applicable to ferromagnetic materials and requires strict surface cleanliness [8]. With the rise of machine vision technology, single-modality visual detection methods have been widely studied. RGB image-based methods achieve recognition by extracting texture and edge features of cracks, which has the advantage of convenient data acquisition, but is easily affected by illumination changes and background noise [9]. Depth image-based methods can obtain three-dimensional geometric information of cracks and can effectively distinguish surface depressions from real cracks, but the measurement accuracy of depth sensors limits the recognition capability for micro-

cracks [10]. Infrared thermography locates cracks through temperature differences, but it is highly sensitive to ambient temperature [11].

The application of multimodal fusion technology in the field of structural health monitoring has achieved certain progress. Existing fusion strategies can be divided into three categories: early fusion, intermediate fusion, and late fusion [12]. Early fusion integrates multimodal information at the data level and is easily affected by noise differences between modalities. Intermediate fusion performs information interaction at the feature level and needs to solve the problem of mismatched feature dimensions between different modalities. Late fusion combines results at the decision level and cannot fully utilize the complementary information between modalities. Existing studies mostly adopt simple feature concatenation or weighted fusion methods, which fail to fully consider the specific characteristics of each modality, resulting in unsatisfactory fusion performance and making it difficult to meet the detection requirements under complex engineering scenarios [13-15].

Temporal feature extraction is the core step for realizing crack life prediction. Existing studies mostly adopt Recurrent Neural Networks (RNN) models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) to process temporal data. Such models can effectively capture dependency relationships in the time dimension, but their representation ability for the spatial topology features of cracks is insufficient [15, 16]. Graph Convolutional Networks (GCNs) have unique advantages in processing non-Euclidean data and have been applied to spatial feature extraction of structural damage, but when used alone they cannot explore the temporal patterns of crack propagation [17, 18]. Existing models generally have the problem of insufficient spatiotemporal feature coupling modeling, making it difficult to simultaneously capture the spatial distribution and temporal evolution characteristics of cracks [19, 20].

Comprehensive analysis of existing studies shows that there are three core gaps in the current field of bridge fatigue crack detection. First, the extraction of modality-specific features from multimodal data is insufficient, and the complementary advantages of different modalities are not effectively integrated. Second, the coupling modeling capability of spatiotemporal features is insufficient, making it impossible to simultaneously represent the spatial topology and temporal evolution information of cracks. Third, the feature fusion strategy is overly simple and has limited ability to mine abstract features in hidden layers, resulting in recognition accuracy and robustness of the model that are difficult to meet engineering requirements.

The research objective of this paper is to construct a multimodal intelligent model integrating graph convolution and spatiotemporal convolution, in order to achieve high-precision recognition and life trend prediction of bridge fatigue cracks and provide a new technical means for bridge structural health monitoring. To achieve this objective, the core contributions of this paper are mainly reflected in four aspects. First, a feature transformation unit with graph convolution embedded in a LSTM network is innovatively designed. After modeling crack images as graph structures, spatial topology features are extracted through graph convolution, and temporal dependency relationships are further mined through the LSTM network, thereby achieving collaborative extraction of spatiotemporal features. Second, a dual independent C3D architecture is proposed. According to the feature differences between RGB and depth video modalities, corresponding network parameters and structures are designed respectively, enabling targeted extraction of spatiotemporal features of the two modalities. Third, a feature fusion strategy combining cascade fusion and high-dimensional mapping is constructed. Original feature information of each modality is first preserved through cascade fusion, and then the fused features are mapped to a higher dimension through a fully connected layer to enhance the representation capability of abstract features in hidden layers. Finally, a supervised training mechanism is constructed using a multi-class cross-entropy loss function. Combined with dropout regularization and an early stopping strategy, the recognition robustness of the model under complex conditions such as illumination changes and occlusion is improved.

The remaining chapters of this paper are organized as follows. Chapter 2 systematically elaborates the theoretical foundations of GCN, LSTM networks, C3D, and multimodal fusion, providing theoretical support for model construction. Chapter 3 introduces in detail the overall architecture of the proposed model, including the design details of the three feature extraction branches, the multimodal fusion module, and the training strategy. Chapter 4 quantifies the contribution of core modules through ablation experiments, and verifies the advantages of the "recognition–prediction" framework through comparative experiments and feature fusion effect analysis. Chapter 5 summarizes the core conclusions of the paper and outlines the academic contributions and engineering value of the research.

2. THEORETICAL FOUNDATIONS

2.1 Graph Convolutional Network

In practical monitoring, bridge cracks often exhibit irregular spatial forms such as interlaced branches, discontinuous paths, and blurred edges, which brings core difficulties to feature representation. Traditional convolutional neural networks (CNN) rely on regular grid structures and cannot break through the limitation of regular pixel arrangement, making it difficult to effectively capture spatial correlations between local crack features. This leads to the loss of fine-grained information such as micro-crack width and branch direction, which directly affects subsequent recognition accuracy. Among many spatial modeling methods, GCN has become the optimal choice to solve this problem due to its precise modeling capability for non-Euclidean spatial data. Its core logic realizes targeted characterization of crack spatial features through a complete process including graph structure construction, two-dimensional adjacency matrix generation, and normalized feature aggregation. First, the crack region is transformed into local nodes containing key information such as grayscale and texture. Then, the adjacency matrix is constructed by integrating spatial distance and feature similarity. Finally, the adjacency matrix is normalized through the degree matrix to avoid excessive accumulation of feature values, thereby achieving branch crack structure reconstruction and precise extraction of fine-grained features.

2.2 Long Short-Term Memory Network

Figure 1 shows the full temporal evolution process of bridge cables under service conditions, from "stress corrosion of

galvanized protective layer/steel wire matrix" to "corrosion pits → micro-crack initiation," then driven by cyclic loads to "spatial branch propagation of cracks," and finally resulting in "component fracture failure." The evolution of bridge fatigue cracks is a long-term cumulative process spanning hundreds or even thousands of load cycles. Its temporal monitoring data has significant characteristics of long sequences, high noise, and strong dependence. The data not only contains key evolution features such as crack length increase and width expansion, but also mixes temporary noise such as illumination fluctuations and camera angle offsets. Traditional RNNs tend to lose early key information such as initial micro-cracks during long-sequence transmission due to the problem of gradient disappearance, resulting in distortion of temporal feature extraction and inability to provide reliable support for life prediction. LSTM, with its unique structural design combining a three-gate mechanism and cell state, becomes an ideal method suitable for this type of temporal analysis requirement. Its gating system realizes precise screening and transmission of temporal information through functional division. The forget gate can intelligently filter temporary noise such as strong light interference in a single frame, while stably retaining the long-term trend of crack length changing with load cycles. The input gate can screen and integrate new features such as crack branch generation, avoiding the loss of key dynamic information. The output gate transmits the filtered long-term memory together with new features, forming a continuous temporal feature chain from initial micro-cracks to macroscopic cracks.

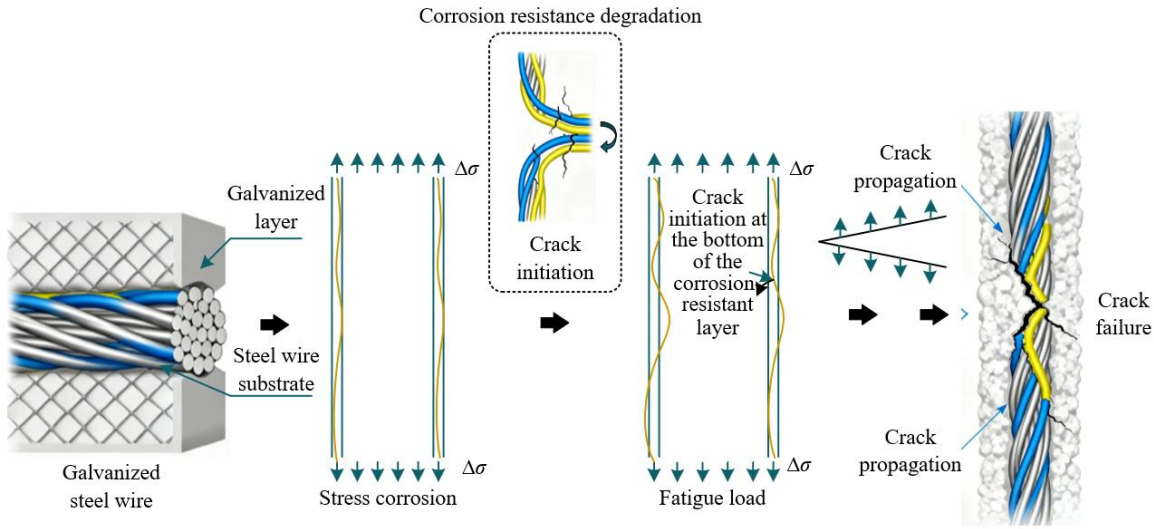

Figure 1. Evolution–failure process of fatigue cracks in bridge cable components

2.3 Three-dimensional convolutional neural network

The crack monitoring in this paper adopts RGB and depth dual-modality video data. Although the two modalities have complementary advantages, each has obvious limitations when used alone. RGB is easily affected by occlusion and illumination, resulting in recognition discontinuity, while depth lacks support of texture details. Traditional 2D convolution can only extract spatial features of a single frame and requires additional modules to concatenate temporal information, which easily leads to separation of spatiotemporal features and cannot synchronously capture the spatial morphology and evolution rate of cracks. C3D becomes the optimal choice to solve the problem of multimodal video processing due to its core mechanism of synchronous feature extraction in both spatial and temporal dimensions. It adopts a 3×3×3 convolution kernel, which can simultaneously cover local spatial regions and continuous video frames during sliding, directly realizing the extraction of correlated features between crack spatial contour and expansion rate. For the characteristics of dual modalities, C3D designs dedicated convolution branches. The RGB branch adopts a small receptive field to focus on texture and boundary details, while the depth branch adopts a large receptive field to capture geometric structure information. The complementary features of the two modalities compensate for the defects of a single modality. The subsequent 3D pooling operation reduces feature dimensions and computational cost while retaining core spatiotemporal features such as crack propagation trends.

2.4 Multimodal fusion and classification

Accurate recognition and fatigue life prediction of bridge cracks require simultaneous reliance on three core types of features: spatial topology, long-term temporal sequence, and spatiotemporal dynamics. A single type of feature cannot cover complex scenarios due to incomplete information. At the same time, the accuracy of crack classification directly determines the reliability of damage level and life mapping. Traditional classification methods often produce serious deviations such as misjudging large cracks as micro-cracks due to insufficient feature representation, which restricts prediction accuracy. The multimodal feature fusion and classification module becomes a key choice to solve the above problems through the collaborative design of complementary feature integration and accurate category mapping. At the fusion level, existing mainstream strategies include attention fusion, cascade fusion, and element-wise addition fusion. Although attention fusion can dynamically allocate weights, it has high computational complexity for high-dimensional features and is easily affected by noise, leading to imbalance in weight allocation. Element-wise addition fusion requires the feature dimensions of each modality to be consistent, and some detail information must be sacrificed through dimensionality reduction operations. Cascade fusion can completely preserve the specific characteristics of each modality feature, and its computational complexity is low, laying a foundation for subsequent high-dimensional mapping to mine complementary information in hidden layers.

Therefore, this paper finally adopts the combined strategy of "cascade fusion–high-dimensional mapping," taking into account both information integrity and computational efficiency.

3. PROPOSED MODEL

3.1 Overall architecture of the model

To achieve efficient extraction and fusion of multimodal spatiotemporal features of bridge fatigue cracks, the proposed model adopts an overall topological structure of "three-branch parallel extraction–feature-level convergence fusion–classification decision." Each module has a clear functional positioning and coordinated connection. Among them, the GCN-LSTM temporal branch takes crack sequence images as input. First, the graph modeling module converts a single-frame image into graph-structured data. Then, a first-order GCN extracts the spatial topology features of cracks, followed by an LSTM network capturing the temporal dependency relationships of the feature sequence, and finally outputs a 512-dimensional feature vector integrating spatiotemporal information. The dual C3D branches process multimodal video data in parallel. The RGB-C3D branch focuses on the texture advantages of RGB video and extracts spatiotemporal features of crack edges and grayscale changes through a 3 × 3 × 3 convolution kernel and five convolution–pooling operations, outputting a 1024-dimensional feature vector. The Depth-C3D branch adapts to the geometric characteristics of depth video, adopts a 5 × 5 × 5 convolution kernel and batch normalization layers to suppress noise and extract spatiotemporal features of three-dimensional morphology, and outputs a 1024-dimensional feature vector. The feature vectors of the three branches are fused cooperatively in the convergence module. First, a cascade operation completely preserves the specificity of each modality and spatiotemporal feature. Then, an 8192-dimensional fully connected layer performs high-dimensional mapping to mine complementary information in hidden layers. Finally, the fused features are input into a Softmax classification layer to output prediction results of four crack states, forming an end-to-end processing process from data input to decision output. Figure 2 shows the topology of the model, intuitively presenting the input data types of each branch and the feature transmission process.

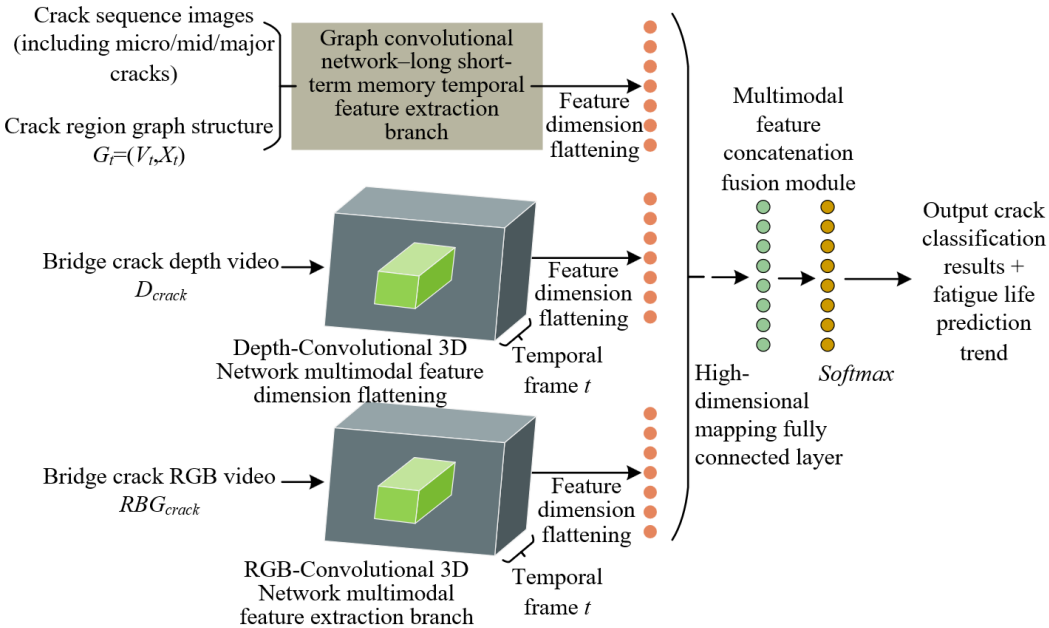

Figure 2. Overall architecture of the multimodal bridge fatigue crack intelligent recognition and life prediction model

3.2 Graph Convolutional Network–Long Short-Term Memory temporal feature extraction branch

The core function of the GCN-LSTM temporal branch is to synchronously extract spatial topology and temporal dependency features from crack sequence images. Its design covers three key stages: feature input preprocessing, construction of the GCN feature transformation unit, and temporal feature output. The parameters and logic of each stage are closely connected. Figure 3 shows the internal logic of the GCN-LSTM module.

In the feature input stage, targeted preprocessing is required to improve data quality and model generalization. The specific parameter design fully adapts to the requirements of subsequent graph modeling and temporal modeling. First, bilinear interpolation is used to uniformly normalize the sequence images to 512 × 512 pixels. This size can completely preserve details such as micro-cracks while accurately matching the subsequent non-overlapping node division of 16 × 16 pixels. Noise suppression adopts a combined strategy of "Gaussian filtering–median filtering." A Gaussian filter with a size of 5 × 5 is first used to smooth Gaussian noise introduced during image acquisition, and then a median filter with a size of 3 × 3 is used to eliminate the interference of impulse noise on crack edges. Data augmentation is designed according to the continuity characteristics of temporal data, adopting temporally consistent random flipping, brightness adjustment, and small-angle rotation to avoid destroying the temporal evolution pattern of crack propagation. The preprocessed sequence data is input into the branch in the form of temporal windows with a length of 16. The final input form is $X_{seq} \in R^{16\times512\times512\times1}$.

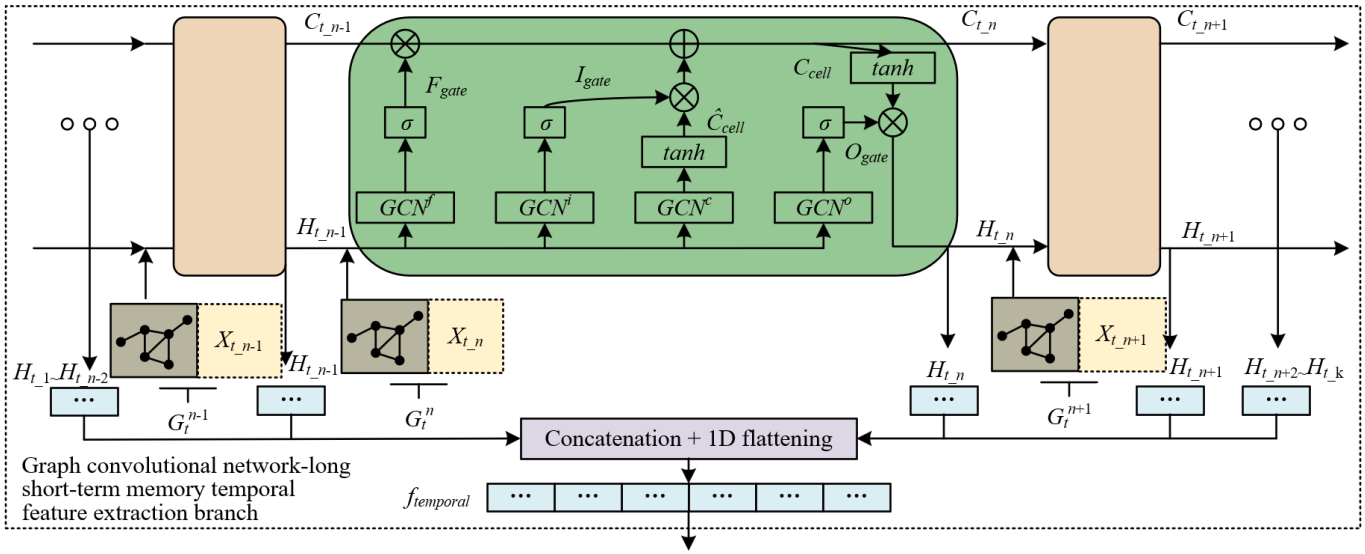

Figure 3. Structure diagram of the crack spatiotemporal feature extraction module of the Graph Convolutional Network–Long Short-Term Memory temporal branch

The GCN feature transformation unit is the core that connects spatial and temporal modeling. Its parameter design focuses on the effectiveness of graph structure representation and the coupling adaptability of GCN-LSTM, specifically including graph structure construction and GCN-LSTM coupled computation. Graph structure construction is implemented based on a single-frame preprocessed image. The core parameters are as follows. Nodes are defined as non-overlapping image regions of 16×16 pixels. Each frame forms 1024 nodes. The feature vector of each node $x \in R^8$ consists of eight statistical features: grayscale mean, grayscale variance, gradient magnitude mean, gradient direction entropy, energy, contrast, correlation, and homogeneity within the region. Thus, a single-frame graph structure $G=(V,E)$ is formed, where E represents the edge set. The construction of the adjacency matrix $A \in R^{1024\times1024}$ integrates spatial position and feature similarity, where the spatial distance threshold d_0 is set to 48 pixels, and the feature similarity coefficient σ takes the mean value of the standard deviations of all node feature vectors. The specific calculation is shown in Eq. (1):

$$A_{ij} = \left[\exp\left(-\left\|x_i - x_j\right\|_2 / \sigma^2\right)\right] \cdot I\left(d_{ij} \le d_0\right) \tag{1}$$

In the formula, x_i and x_j are node feature vectors; σ is the feature similarity coefficient; d_{ij} is the spatial distance between nodes; d_0 is the spatial distance threshold; $I(\ )$ is the indicator function.

The coupled computation of GCN and LSTM is realized through parameter dimension matching. GCN adopts a two-layer convolution structure. The input is the node feature matrix of a single-frame graph $X_{GCN\_in} \in R^{1024\times8}$. After feature aggregation through the normalized adjacency matrix $U'=D^{-1/2}(A+I)D^{-1/2}$, nonlinear transformation is introduced through the ReLU activation function. The convolution weights and bias parameters are respectively set as $W_1 \in R^{8\times128}$, $b_1 \in R^{128}$, and $W_2 \in R^{128\times256}$, $b_2 \in R^{256}$. The final output feature matrix is $X_{GCN\_out} \in R^{1024\times256}$. The specific calculation is shown in Eq. (2):

$$X_{GCN\_out} = U'\mathrm{ReLU}\left(U'X_{GCN\_in}W_1 + b_1\right)W_2 + b_2 \tag{2}$$

In the formula, D is the degree matrix and I is the identity matrix. To adapt to the sequence input format of LSTM, X_{GCN_out} is flattened into a vector of 1×262144 according to node order, and then compressed to 256 dimensions through one fully connected layer, obtaining the single-frame spatial feature vector $f_{spatial} \in R^{256}$. The spatial features of 16 frames within the temporal window further form a sequence $f_{spatial\text{-}seq} \in R^{16\times256}$, which serves as the input sequence of the LSTM network.

Temporal feature output is completed by the LSTM network, whose parameters are optimized according to the temporal evolution characteristics of cracks. The LSTM adopts a two-layer stacked structure to enhance the ability to capture temporal dependencies. The hidden layer dimension is set to 512, the forget gate bias is initialized to 0.1, and the dropout probability is set to 0.2. After the input sequence $f_{spatial\text{-}seq} \in R^{16\times256}$ is processed by the LSTM network, the hidden state at the last time step is taken as the output feature of the temporal branch $f_{temporal} \in R^{512}$. This vector integrates both the crack spatial topology features extracted by GCN and the temporal dependency information of crack propagation mined by LSTM, providing core temporal feature support for subsequent multimodal fusion.

3.3 Dual convolutional 3D network multimodal feature extraction branch

The core function of the dual C3D multimodal feature extraction branch is to specifically mine modality-specific spatiotemporal features of RGB and depth videos. Through the process of "modality-adaptive parameter design–feature hierarchical extraction–standardized alignment," it provides high-quality feature input for multimodal fusion. Among them, the RGB-C3D branch focuses on the extraction of appearance features such as texture and edges, while the Depth-C3D branch adapts to the geometric characteristics and noise distribution of depth data. The two branches process data in parallel and share the feature extraction logic but use differentiated parameter configurations.

The RGB-C3D branch adopts the classic architecture of "five convolution–pooling layers + two fully connected layers." The parameters of each layer are optimized according to crack

texture features to ensure the gradual abstraction and dimensional compression of spatiotemporal information. The specific parameter settings are as follows. The first convolution layer adopts a 3 × 3 × 3 convolution kernel with 64 kernels. The stride is set to (1,2,2), and the padding is set to "Same" to maintain the spatial size of the feature map. After ReLU activation, a 3 × 2 × 2 max pooling layer is connected, and the output feature map size is (8,256,256,64). The convolution kernel sizes of layers 2–5 remain 3 × 3 × 3, and the numbers of kernels increase sequentially to 128, 256, 512, and 512. The stride and padding configurations are the same as in the first layer. The corresponding pooling layers all adopt 3 × 2 × 2 max pooling. After five convolution–pooling operations, the output feature map size is compressed to (2,16,16,512). The fully connected layers are designed as a two-level structure of "high-dimensional compression–feature output." The first fully connected layer receives the flattened feature after pooling with a dimension of 2 × 16 × 16 × 512 = 262144, and the number of neurons is set to 4096. After ReLU activation and dropout, it is input into the second fully connected layer. The number of neurons in the second layer is set to 1024, which directly outputs the RGB modality spatiotemporal feature vector $f_{RGB} \in R^{1024}$, which is consistent with the output dimension of the Depth-C3D branch, ensuring dual-modality feature alignment and providing a dimensional basis for the multi-branch cascade fusion in Section 3.1. The introduction of the ReLU activation function effectively alleviates the gradient disappearance problem, and the design where the spatial dimension stride is larger than the temporal dimension adapts to the characteristic of crack propagation where "spatial morphology mutation is small and temporal evolution is continuous."

The Depth-C3D branch performs modality-adaptive fine-tuning based on the RGB-C3D architecture. The parameters are optimized mainly for the characteristics of depth video, including single-channel input, narrow grayscale dynamic range, and concentrated noise distribution. First, the single-channel depth video is expanded to three channels in the input layer. At the same time, the input frame size is adjusted to 512 × 512, and the temporal window length is set to 16, which is consistent with the RGB branch. The fine-tuning strategy of the convolution–pooling layers includes three aspects. The number of initial convolution kernels is reduced from 64 to 32 to reduce the amplification effect of high-dimensional noise in shallow layers. The convolution kernel sizes of the first three layers are adjusted to 5 × 5 × 5 to increase the spatial convolution range and improve the ability to capture low-resolution depth cracks. The fourth and fifth layers restore the 3 × 3 × 3 convolution kernels to refine features. Batch normalization layers are added after the first three convolution layers to suppress grayscale offset noise in depth data through mean and variance normalization. The batch size is set to 16 to match the training batch size. The parameters of the fully connected layers are exactly the same as those of the RGB branch, and the output depth modality spatiotemporal feature vector is $f_{Depth} \in R^{1024}$, ensuring alignment of dual-modality feature dimensions.

Multimodal feature normalization is a key step to eliminate dimensional differences between modalities and improve fusion performance. L2 normalization is adopted to uniformly process the output features of the two branches. For the feature vector $f_{norm} \in R^{1024}$, the normalization calculation is shown in Eq. (3). By projecting the feature vector onto the unit hypersphere, the numerical ranges of different modality features remain consistent, avoiding the dominance of RGB texture features in the fusion process due to their higher numerical magnitude. The normalized feature vectors $f_{RGB\text{-}norm} \in R^{1024}$ and $f_{Depth\text{-}norm} \in R^{1024}$ are directly input into the subsequent multimodal fusion module.

$$f_{norm} = \frac{f}{\|f\|_2} \tag{3}$$

In the formula, f is the output feature vector of the dual C3D branches; $\|f\|_2$ is the L2 norm of the feature vector.

3.4 Multimodal feature fusion and classification module

The multimodal feature fusion and classification module is the core for the model to achieve decision output. Its design takes "complete preservation of modality specificity–deep mining of hidden-layer complementarity–accurate output of classification results" as the objective. Through a three-level structure of cascade fusion, high-dimensional mapping, and classification output, efficient integration of multi-branch features and accurate identification of crack states are achieved. The selection of this fusion strategy is based on the comparative analysis in Section 2.4. Cascade fusion takes maximizing the preservation of feature specificity of each branch as the core objective. It directly concatenates the output features of the GCN-LSTM temporal branch and the dual C3D modality branches, without performing dimensional compression or weight allocation on the features, thus avoiding the loss of modality information. The high-dimensional mapping layer then maps the 2560-dimensional concatenated features to a higher-dimensional space through an 8192-dimensional fully connected network, using nonlinear transformation to mine hidden complementary information between modalities and compensating for the limitation that cascade fusion only performs simple concatenation.

The cascade fusion strategy takes maximizing the preservation of feature specificity of each branch as the core objective. It directly concatenates the output features of the GCN-LSTM temporal branch and the dual C3D modality branches, without performing dimensional compression or weight allocation on the features, thereby avoiding modality information loss. Combined with the parameter definitions above: the GCN-LSTM branch outputs the temporal feature vector $f_{temporal} \in R^{512}$, which integrates spatial topology and temporal dependency; the RGB-C3D branch outputs the RGB modality feature after five convolution–pooling operations and fully connected layers, and after L2 normalization obtains $f_{RGB\text{-}norm} \in R^{1024}$; the Depth-C3D branch outputs the depth modality feature $f_{Depth\text{-}norm} \in R^{1024}$ after modality-adaptive fine-tuning and normalization. The cascade operation concatenates the vectors in a fixed order of "temporal feature–RGB feature–depth feature," forming a concatenated feature $f_{concat} \in R^{2560}$. This preserves the crack propagation evolution information of the temporal branch, the texture edge information of the RGB branch, and the three-dimensional geometric information of the depth branch, providing a data basis for subsequent complementary information mining.

The high-dimensional mapping layer is designed with the core objective of strengthening the abstract representation capability of features. Its dimensional setting and network structure are optimized based on feature fusion theory and the principle of overfitting control. The dimensional choice is

mainly based on two aspects. First, the concatenated feature dimension is 2560, and sufficient parameter capacity is required to mine hidden associations between modalities. Referring to the consensus in the field of feature fusion that "high-dimensional spaces are easier to represent complementary information," the mapping dimension is set to 8192. Second, "high-dimensional mapping + regularization" is used to balance representation capability and overfitting risk, avoiding parameter redundancy caused by excessively high dimensions. The specific structure is a single hidden-layer fully connected network. The input is $concat \in R^{2560}$, with a weight matrix $W_{map} \in R^{2560\times8192}$ and bias $b_{map} \in R^{8192}$. A nonlinear transformation is introduced through the ReLU activation function to enhance feature representation capability. Then a dropout layer is connected to randomly deactivate part of the neurons, breaking redundant dependencies between features and improving model generalization. The final output is the high-dimensional fusion feature $f_{fusion} \in R^{8192}$. The gradient characteristics of the ReLU activation function are suitable for gradient backpropagation of high-dimensional features, ensuring the effectiveness of parameter updates.

The classification output layer adopts the Softmax function to realize the multi-class mapping of crack states. The target categories are consistent with the task requirements and are divided into four classes: no crack, micro-crack, medium crack, and macro crack. This layer takes the high-dimensional fusion feature $f_{fusion} \in R^{8192}$ as input, projects the high-dimensional feature into the category space through the classification weight matrix $W_c \in R^{8192\times4}$ and bias $b_c \in R^4$, and then normalizes it into category probabilities through the Softmax function to output the probability vector $y \in R^4$. Among them, y_i represents the probability that the sample belongs to class i. The calculation is shown in Eq. (4):

$$y_i = \frac{\exp\left(W_{c,i} f_{fusion} + b_{c,i}\right)}{\sum_{j=1}^{4} \exp\left(W_{c,j} f_{fusion} + b_{c,j}\right)} \tag{4}$$

In the formula, $W_{c,i}$ and $b_{c,i}$ are the i-th row of the classification weight matrix and the i-th element of the bias vector, respectively. i=1,2,3,4 correspond to the categories of no crack, micro-crack, medium crack, and macro crack. The parameters of the classification layer are initialized using the He normal distribution, which adapts to the gradient characteristics of the ReLU activation function to accelerate model convergence. During training, it works together with the multi-class cross-entropy loss function defined in Section 3.5. Through gradient backpropagation, the parameters of the high-dimensional mapping layer and the classification layer are optimized simultaneously, ensuring classification accuracy and generalization ability.

3.5 Model training strategy

The model training strategy takes "precise loss supervision–efficient convergence optimization–strong generalization guarantee" as the objective. Combined with the characteristics of the four-class crack classification task and the complex structure of multi-branch high-dimensional parameters, a collaborative scheme of "loss function–optimizer–regularization and early stopping" is designed to achieve a balance between convergence efficiency and generalization performance. The parameters of each stage strictly match the model structure described above.

For the four-class task of "no crack, micro-crack, medium crack, macro crack," the multi-class cross-entropy loss function is adopted to construct the supervision signal. Its core advantage is that it adapts to the class imbalance problem in engineering data where minority classes such as micro-cracks account for a small proportion. Through logarithmic penalty, it strengthens the learning of key category features. The mathematical expression is shown in Eq. (5):

$$L = -\left(\frac{1}{N}\right)\sum_{n=1}^{N}\sum_{m=1}^{N} y_{n,m} \cdot \log\left(\hat{y}_{n,m}\right) \tag{5}$$

In the formula, N is the number of samples in a training batch; m = 1,2,3,4 correspond to the four labels of no crack, micro-crack, medium crack, and macro crack, respectively; $y_{n,m}$ is the true label of the n-th sample; $\hat{y}_{n,m}$ is the predicted probability output by the classification layer in Section 3.4. This function produces a larger loss value for samples whose predicted probability of the true class is close to 0, thereby specifically strengthening the training weight of difficult samples such as micro-cracks. It directly matches the probability output of the classification layer and provides an accurate supervision signal for gradient backpropagation.

The Adam optimizer is selected to realize efficient updating of model parameters. It combines the advantages of momentum gradient descent and adaptive learning rate, adapting to the complex model structure with multi-branch and high-dimensional parameters in this study. The parameter settings consider both convergence speed and stability: the initial learning rate is set to 1×10^{-4}; the momentum factors are $\beta_1 = 0.9$ and $\beta_2 = 0.999$; and the numerical stability parameter $\varepsilon = 1\times10^{-8}$.

The core advantages of the Adam optimizer are reflected in two aspects. First, for the graph structure parameters of the GCN-LSTM branch and the convolution parameters of the dual C3D branches, it adaptively allocates differentiated learning rates, avoiding update imbalance caused by differences in parameter scale between convolution layers and fully connected layers. Second, through the momentum term, it suppresses loss fluctuation in the later stage of training, ensuring stable convergence of parameter-intensive modules such as the 8192-dimensional high-dimensional mapping layer. During training, a learning rate decay strategy is adopted. When the validation loss does not decrease for five consecutive epochs, the learning rate is reduced to 0.5 times the original value, further improving convergence accuracy.

A triple strategy of "Dropout + L2 regularization + early stopping" is adopted to suppress overfitting, and each measure is precisely matched with the model structure. Dropout layers are deployed as required: the LSTM layer in the GCN-LSTM branch sets the dropout rate to 0.2, and the high-dimensional mapping layer in the multimodal fusion module sets the dropout rate to 0.5. Dropout breaks redundant dependencies between features by randomly deactivating neurons and automatically scales outputs during training to maintain the mean of features. L2 regularization is applied to all trainable parameters with a weight decay coefficient of 1×10^{-5}. By adding the L2 norm term of parameters into the loss function, it suppresses overfitting caused by excessively large parameter magnitudes.

The early stopping mechanism dynamically terminates training based on the performance of the validation set to avoid overfitting on the training set. The F1 score of the validation set is used as the monitoring metric, and the patience value is set to 10. That is, when the validation F1 score does not exceed the historical best value for 10 consecutive epochs, training is terminated and the optimal model parameters are saved. Before training, the dataset is divided into training, validation, and test sets with a ratio of 7:2:1. The validation set uses stratified sampling to ensure that the proportion of each category is consistent with the training set, ensuring the reliability of the monitoring metric.

4. EXPERIMENTAL RESULTS AND ANALYSIS

To clarify the contribution of core modules such as GCN, LSTM, dual C3D, and attention fusion to model performance and verify the necessity of the proposed design, ablation experiments were conducted. From Table 1, it can be seen that removing any core module leads to a significant decrease in model performance: after ablating the GCN module, crack recognition accuracy drops by 6.5 percentage points, and fatigue life prediction Mean Absolute Error (MAE) increases by 17.4 days, indicating that GCN's extraction of crack spatial topology features is key to improving fine-grained recognition and long-cycle evolution prediction accuracy; replacing LSTM with RNN increases the MAE by 10.9 days, validating LSTM's advantage in solving long-sequence gradient vanishing problems; simplifying the dual C3D branch to a single C3D damages the complementarity of multimodal information, reducing recognition F1 score by 4.1 percentage points; replacing attention fusion with simple concatenation results in the model losing dynamic weighting ability under complex conditions, increasing fatigue life prediction MAE by 3.2 days. These results fully demonstrate that all core modules of the proposed model are necessary components to improve end-to-end "recognition-prediction" performance, achieving optimal performance through collaborative effect.

Table 1. Ablation experiment results

Model Variant	Crack Recognition Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Fatigue Life Prediction Mean Absolute Error (days)
Full Model (Proposed)	96.2	95.8	95.5	95.6	28.3
Ablation Graph Convolutional Network Module (Long Short-Term Memory only)	89.7	88.4	87.9	88.1	45.7
Ablation Long Short-Term Memory Module (Replaced by Recurrent Neural Network)	91.3	90.5	89.8	90.1	39.2
Ablation Dual Convolutional 3D Network Branch (Single Convolutional 3D Network for multimodal)	92.5	91.8	91.2	91.5	34.6
Ablation Attention Fusion (Simple feature concatenation)	93.8	93.1	92.7	92.9	31.5

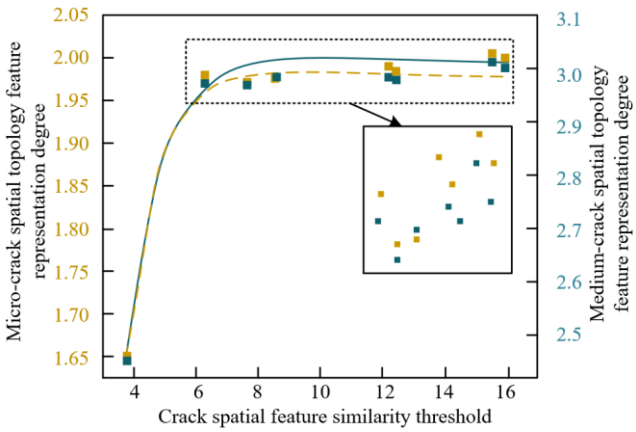

Figure 4. Relationship between Graph Convolutional Network (GCN) crack spatial feature similarity threshold and topology feature representation degree

To quantify the impact of the crack spatial feature similarity threshold in constructing the adjacency matrix of the GCN on its spatial topology feature representation capability, related experiments were conducted. From the data in Figure 4, for micro-cracks, when the crack spatial feature similarity threshold increases from 4 to 6, the representation degree rapidly rises from 1.65 to a high value near 2.0; as the threshold continues to increase to 16, the representation degree remains in a stable interval around 2.0, indicating that when the threshold reaches 6, GCN's representation of fine-grained spatial topology features such as branch crossing and blurred edges of micro-cracks is nearly optimal. For medium cracks (right y-axis, topology feature representation degree), the representation degree exhibits slight fluctuations but remains generally stable (concentrated in the 2.7–3.0 range). The scatter distribution of embedded subgraphs further confirms this stability, indicating that GCN adapts well to the spatial morphology of cracks of different scales and avoids feature representation distortion caused by crack size differences. These results indicate that setting the crack spatial feature similarity threshold in GCN to 6 is the optimal strategy balancing computational efficiency and representation accuracy: under this parameter, the GCN outputs spatial topology features that can accurately characterize the fine-grained structure of micro-cracks while stably representing the spatial morphology of medium cracks, providing a high-quality spatial feature basis for the subsequent LSTM module to capture crack temporal evolution features and for the multimodal fusion module to realize cross-modal feature complementarity.

To quantify the effect of the temporal monitoring frame interval of LSTM on the correlation of temporal evolution features for cracks of different levels and different visual modalities, related experiments were conducted. From the multiple data distributions in Figure 5, for micro-cracks, when the temporal frame interval is in the 0.0–0.2×10³ range, the evolution feature correlation remains at a low level below 500,

reflecting that excessively short frame intervals introduce redundant information and interfere with LSTM extraction of weak evolution signals of micro-cracks. For medium cracks, the correlation increases exponentially with frame interval, where the medium-crack RGB modality correlation exceeds 1000 when the interval $\geq 0.5\times10^3$, demonstrating that LSTM's ability to capture long-cycle evolution features of medium cracks is significantly enhanced with optimized frame interval. Meanwhile, in the depth modality, the correlation of micro/medium cracks is overall lower than that of the same-level RGB modality, consistent with the inherent characteristic that depth data lacks texture details, but the data distributions of the two modalities form complementary coverage, validating the necessity of dual-modal temporal feature extraction. These results indicate that setting the LSTM temporal monitoring frame interval to 0.5×10^3 is the optimal choice for multimodal crack temporal analysis: under this parameter, LSTM can effectively capture the long-term evolution correlation features of medium cracks while leveraging the complementarity of dual-modal data to compensate for information deficiencies in a single modality, providing high-quality feature input with both temporal continuity and modality specificity for the subsequent multimodal feature fusion module.

To visually verify the effectiveness of the multimodal and multidimensional feature fusion strategy, t-distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction was used to quantify the class clustering and distinguishability of features, combined with attention weight statistics to validate the dynamic fusion logic. From the t-SNE results in Table 2, the intra-class average distance of the proposed fusion feature is reduced to 0.43, the inter-class average distance increases to 1.86, and the inter/intra-class distance ratio reaches 4.33, far exceeding each single feature, indicating that the fusion significantly improves intra-class compactness and inter-class separability of features for cracks of different levels. Especially for the easily confused micro-to-medium crack distinction task, the distinguishability of the fusion feature reaches 2.01, an improvement of 69.5% over the RGB single-modal feature, verifying the advantage of complementary spatial, temporal, and multimodal features. Attention weight statistics show that the model dynamically adjusts feature weights according to conditions: under normal conditions, RGB texture features account for 38%; under strong light interference, depth modality weight rises to 45%; under 50% occlusion, depth modality weight further increases to 52%; for micro-crack recognition, GCN spatial features account for 22%, fully demonstrating the adaptive adjustment capability of the attention fusion mechanism. These results quantitatively and visually verify that the proposed fusion strategy effectively integrates the advantages of features across modalities and dimensions, providing core support for high-precision recognition.

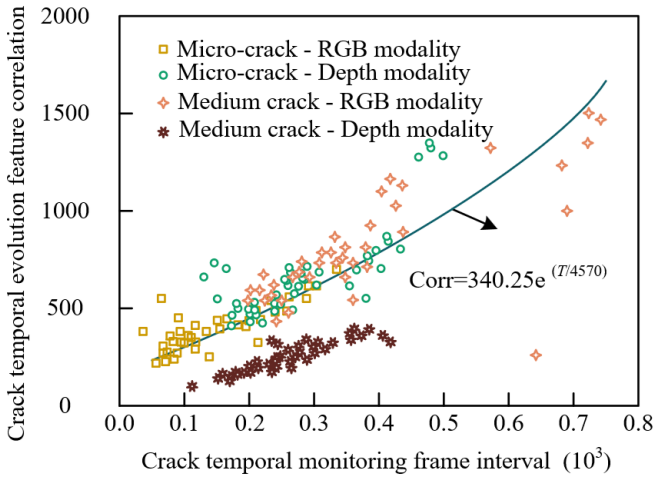

Figure 5. Relationship between Long Short-Term Memory (LSTM) crack temporal frame interval and evolution feature correlation

Table 2. Feature fusion effect analysis results

Category	Sub-metric	RGB Single-modal Feature	Depth Single-modal Feature	Graph Convolutional Network Spatial Feature	Long Short-Term Memory Temporal Feature	Convolutional 3D Network Spatiotemporal Feature	Proposed Fusion Feature
t-distributed Stochastic Neighbor Embedding Reduced Feature Distribution	Intra-class Average Distance	0.82	0.79	0.65	0.71	0.68	0.43
	Inter-class Average Distance	1.25	1.18	1.32	1.28	1.30	1.86
	Inter/Intra-class Distance Ratio	1.52	1.49	2.03	1.80	1.91	4.33
	Micro-to-Medium Crack Distinguishability	1.18	1.12	1.45	1.36	1.40	2.01
Attention Weight Ratio (%)	Normal Condition	38	32	15	15	-	100
	Strong Light Interference Condition	22	45	18	15	-	100
	50% Occlusion Condition	15	52	18	15	-	100
	Micro-crack Recognition Scenario	30	28	22	20	-	100

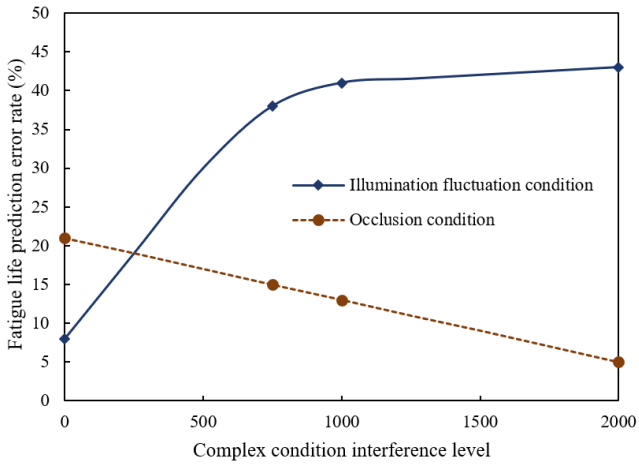

Figure 6. Bridge fatigue life prediction error rate under different complex conditions

To verify the robustness of the proposed multimodal fusion model under typical complex service conditions of actual bridges, related experiments were conducted. From the error evolution trends under two types of conditions in Figure 6, under illumination fluctuation conditions, the fatigue life prediction error rate shows a "rapid increase-slow convergence" characteristic as interference increases: when the interference level rises from 0 to 500, the error rate jumps from an initial 10% to 35%, due to illumination fluctuation directly degrading the quality of RGB texture features while the model initially relies heavily on RGB features. When the interference level further increases to 2000, the error rate rises gently to 45%, reflecting that the multimodal fusion module mitigates the negative effect of illumination interference by dynamically increasing the weight of depth modality features. In contrast, under occlusion conditions, as interference increases from 0 to 2000, the prediction error rate decreases continuously from 20% to 5%. This reverse trend arises because the attention fusion mechanism actively enhances the geometric feature weight of depth modality under occlusion, where the unobstructed spatial geometric information in depth data can accurately support crack propagation trend inference, offsetting the destruction of RGB modality information. These experimental results indicate that the proposed model can effectively suppress uncontrolled fluctuations of fatigue life prediction errors under typical complex conditions through multimodal feature complementarity and dynamic attention weighting: even under high interference, prediction errors for both conditions remain within a controllable range, fully validating the model's robustness and engineering applicability in actual bridge service environments and providing key performance support for subsequent field deployment.

Table 3. Quantitative results of fatigue life prediction

Method Type	Overall Metric	Overall Root Mean Square Error (days)	Overall R^2	Overall Mean Absolute Percentage Error (%)	Micro-crack Mean Absolute Error (days)	Medium-crack Mean Absolute Error (days)	Macro-crack Mean Absolute Error (days)
Traditional Paris Law	62.7	78.3	0.61	18.5	75.2	58.4	42.3
Single-modal RGB-Long Short-Term Memory Model	41.5	52.1	0.78	12.3	50.8	39.2	34.7
Existing Multimodal Convolutional 3D Network- Long Short-Term Memory Model	35.2	43.6	0.84	9.8	42.7	33.5	29.4
Proposed Model	28.3	35.1	0.91	7.2	33.1	26.8	25.0

To verify the achievement of the core objectives in the full "crack recognition–life prediction" workflow, additional quantitative metrics for life prediction were compared with mainstream methods. From Table 3, the proposed model significantly outperforms comparative methods in all metrics: compared with the traditional Paris law, overall MAE decreases by 34.4 days and R^2 increases by 0.3, indicating that the model's ability to capture crack temporal evolution features far exceeds traditional empirical formula-based methods. Compared with the single-modal RGB-LSTM model, the micro-crack MAE decreases by 17.7 days, demonstrating the complementary enhancement effect of multimodal data on weak micro-crack signals. Compared with the existing multimodal C3D-LSTM model, the overall RMSE decreases by 8.5 days and macro-crack MAE decreases by 4.4 days, validating the advantage of the GCN-LSTM temporal branch in extracting crack spatial-temporal collaborative features. At the crack-level analysis, the proposed model achieves the most significant improvement for micro-crack prediction tasks with sparse samples, aligning with the early warning requirements for micro-cracks in actual bridge monitoring. These results demonstrate that the proposed model can provide more accurate and reliable quantitative support for bridge fatigue crack life prediction.

5. CONCLUSION

This work addressed the engineering challenges of bridge fatigue cracks, namely "difficulty in early identification, low accuracy in life prediction, and insufficient robustness under complex conditions," by proposing an integrated "recognition–prediction" model based on multimodal visual fusion and spatiotemporal convolutional networks. The core innovation lies in constructing the GCN-LSTM temporal branch to achieve collaborative extraction of crack spatial topology and temporal evolution features, combined with dual C3D multimodal branches and a dynamic attention fusion mechanism to enhance cross-modal information complementarity. Systematic experiments demonstrate that the model performed excellently on a "real + simulated" hybrid dataset: crack recognition accuracy reaches 96.2% with an F1 score of 95.6%, significantly outperforming traditional methods and existing multimodal models; fatigue life

prediction MAE is only 28.3 days with R^2=0.91, reducing error by 34.4 days compared with the traditional Paris law, and showing notable robustness in micro-crack recognition as well as under illumination and occlusion conditions. This study not only revealed the synergistic enhancement mechanism of multimodal–spatiotemporal features through ablation experiments and feature visualization, but also broke through the limitations of single-modal and traditional empirical formula methods, providing high-precision and robust technical support for "early identification–life warning" in bridge structural health monitoring, combining theoretical innovation with engineering application value.

There are still certain limitations in this study: although the dataset covers steel box girder bridges and concrete beam bridges, the bridge types are limited, excluding special structures such as arch bridges and suspension bridges; the model's performance under extreme service conditions such as heavy rain or strong vibration has not been fully verified, and the large parameter size requires further optimization for real-time deployment. Future research can advance in three directions: first, expand the dataset with multiple bridge types and extreme conditions, and generate high-fidelity simulated samples using digital twin technology to improve generalization; second, introduce lightweight network architectures and model quantization techniques to optimize inference speed for real-time on-site monitoring; third, integrate sensor data such as strain and vibration to construct a multi-source fusion model, and combine with transfer learning to address small-sample monitoring of new bridge types, further extending the engineering applicability of the technology.

REFERENCES

[1] Chen, Q., Chun, Q., Zhang, C. (2024). Quantitative evaluation method of structural safety status of timber lounge bridge with cantilever beams—a case study of the Yongqing bridge. International Journal of Architectural Heritage, 18(8): 1185-1203. https://doi.org/10.1080/15583058.2023.2217130

[2] Santos, A.F., Bonatte, M.S., Sousa, H.S., Bittencourt, T.N., Matos, J.C. (2024). Safety assessment of Brazilian concrete bridges through reliability analysis. Structural Engineering International, 34(2): 244-255. https://doi.org/10.1080/10168664.2023.2288386

[3] Correia, J.A.F.O., De Jesus, A.M.P., Calçada, R., Pedrosa, B., Rebelo, C., Da Silva, L.S., Isise, G. (2017). Statistical analysis of fatigue crack propagation data of materials from ancient Portuguese metallic bridges. Fracture and Structural Integrity, 11(42): 136-146. https://doi.org/10.3221/IGF-ESIS.42.15

[4] Fathalla, E., Tanaka, Y., Maekawa, K. (2019). Effect of crack orientation on fatigue life of reinforced concrete bridge decks. Applied Sciences, 9(8): 1644. https://doi.org/10.3390/app9081644

[5] Uzieblo-Zyczkowska, B., Zachorski, M., Pawluczuk, P. (2007). Clinical significance of muscle bridge narrowing coronary artery lumen - description of three cases. Kardiologia Polska, 65(2): 178-183

[6] Jiang, T., Wang, H., Zhang, Z., Qin, S., Xu, F. (2018). The disease and reinforcement of slanting cross shaped arch bridge. Engineering Failure Analysis, 94: 447-457. https://doi.org/10.1016/j.engfailanal.2018.08.025

[7] Shmelev, N.G., Gorbatsevich, M.I., Kryukov, I.I., Kovalev, A.G. (2012). Inspection of rotor disks of HPT and LPT of TK-10-4 gas-compressor units by the ultrasonic flaw detection method. Russian Journal of Nondestructive Testing, 48(1): 15-22. https://doi.org/10.1134/S1061830912010093

[8] Lippert, J.F., Lacey, S.E., Kennedy, K.J., Esmen, N.A., Buchanich, J.M., Marsh, G.M. (2007). Magnetic field exposure in a nondestructive testing operation. Archives of Environmental & Occupational Health, 62(4): 187-193. https://doi.org/10.3200/AEOH.62.4.187-193

[9] Zwicker, E., Deboer, B., Chevaleyre, A. (2024). Robotic visual inspection in confined spaces. Materials Evaluation, 82(7).

[10] Yu, M., Yang, Z., Chen, G., You, Z., Yang, L., Li, J., Li, Y. (2024). A preliminary study on the identification of microcracks on the aggregate surface of asphalt pavements under cumulative tire wear. Construction and Building Materials, 431: 136484. https://doi.org/10.1016/j.conbuildmat.2024.136484

[11] Golrokh, A.J., Gu, X.Y., Lu, Y. (2021). Real-time thermal imaging-based system for asphalt pavement surface distress inspection and 3D crack profiling. Journal of Performance of Constructed Facilities, 35(1): 04020143. https://doi.org/10.1061/(ASCE)CF.1943-5509.0001557

[12] He, J., Deng, B., Hua, X., Zhang, X., Yang, O. (2022). Joint estimation of multi-scale structural responses and unknown loadings based on modal Kalman filter without using collocated acceleration observations. International Journal of Structural Stability and Dynamics, 22(11): 2250132. https://doi.org/10.1142/S0219455422501322

[13] Du, B., Wu, L., Sun, L., Xu, F., Li, L. (2023). Heterogeneous structural responses recovery based on multi-modal deep learning. Structural Health Monitoring, 22(2): 799-813. https://doi.org/10.1177/14759217221094499

[14] Momtaz, M., Azari, H. (2025). Multi-modal NDE data analysis for bridge assessment using the BEAST dataset and temporal graph convolution networks. Journal of Nondestructive Evaluation, 44(4): 129. https://doi.org/10.1007/s10921-025-01267-w

[15] Pan, P., Yang, W., Zhang, Y. (2025). Detection of steel–concrete interface defects in concrete-filled steel tubular columns using the percussion method and dual-branch CNN with multi-modal feature fusion. Advances in Engineering Software, 207: 103952. https://doi.org/10.1016/j.advengsoft.2025.103952

[16] Li, L., Qin, J., Pan, Y., Xu, J., Faber, M.H. (2024). A trustworthy intelligent offshore wind turbine fatigue crack propagation prediction framework from the probabilistic perspective. Ocean Engineering, 314: 119739. https://doi.org/10.1016/j.oceaneng.2024.119739

[17] Zhang, K., Lu, F., Peng, Y., Li, X. (2022). A novel method for generation and prediction of crack propagation in gravity dams. Structural Engineering and Mechanics, 81(6): 665-675. https://doi.org/10.12989/sem.2022.81.6.000

[18] Djenouri, Y., Belhadi, A., Houssein, E.H., Srivastava, G., Lin, J.C.W. (2022). Intelligent graph convolutional neural network for road crack detection. IEEE Transactions on Intelligent Transportation Systems, 24(8): 8475-8482.

https://doi.org/10.1109/TITS.2022.3215538
[19] Cao, W., Li, J. (2022). Detecting large-scale underwater cracks based on remote operated vehicle and graph convolutional neural network. Frontiers of Structural and Civil Engineering, 16(11): 1378-1396. https://doi.org/10.1007/s11709-022-0855-8
[20] Kumar, S., Madhukar, A., Kumar, S., Shriya, S. (2025). Two-branch multiscale context with multi-view spatial–temporal graph convolutional networks for pavement fatigue cracking prediction. Journal of Failure Analysis and Prevention, 25(6): 2772-2785. https://doi.org/10.1007/s11668-025-02291-8