Research on Road Crack Detection Based on RGB-LPC-GPR Data Fusion

Zhaowei Wang[1], Dongwei Qiu*[1], Runze Wu[2,3], Yida Shi[1], Wenyue Niu[1]

[1] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China-17611407913@163.com, qiudw@bucea.edu.cn, qq20010509@163.com,2108570023151@stu.bucea.edu.cn
[2] Surveying and Natural Resource Spatial Data Technology Wu Runze Studio, Beijing Institute of Surveying and Mapping Design and Research, Beijing, China- wurunze@bism.cn
[3] Beijing Skill Master Studio, Beijing, China-wurunze@bism.cn

Keywords: Road damage detection, data fusion, multi-modal data processing, feature extraction, temporal modeling.

Abstract

This study presents a multimodal data fusion framework for road damage detection and prediction, integrating RGB images, LiDAR point clouds, and GPR (Ground Penetrating Radar) data to enable high-precision detection of surface cracks, potholes, and underground voids, as well as dynamic trend forecasting. By leveraging Deep Mapping 2.0 and the RAFT algorithm, the alignment accuracy between RGB and LiDAR data was significantly improved, reducing registration error to 2.3 mm. Concurrently, the spatial mapping accuracy of GPR data was enhanced to 4.8 mm, ensuring precise multimodal data fusion. A Cross-Attention Transformer combined with a Feature Pyramid Network (FPN) was used for dynamic feature weighting, achieving a crack detection IoU of 97.3% and an AP@0.5 of 93.7% for underground void detection, thereby substantially enhancing the model's performance in detecting complex road damage. Moreover, a trend prediction model integrating ConvLSTM and a spatiotemporal attention mechanism achieved an MAE of 8.7% in a six-month damage trend prediction experiment, reducing prediction error by 34% compared to existing methods, underscoring the model's effectiveness in forecasting damage progression.The experimental results demonstrate that the proposed framework exhibits strong adaptability and stability across diverse road damage detection tasks, particularly excelling in the joint detection of cracks and underground voids with high accuracy. Furthermore, the framework is readily extendable to infrastructure health monitoring applications, such as bridges and tunnels, providing robust technological support for intelligent road maintenance and offering data-driven insights for the long-term optimization and sustainability of urban transportation infrastructure.

1. Introduction

1.1 Challenges in Urbanization and Road Damage Detection

With the accelerated pace of urbanization, urban roads, as critical transportation infrastructure, bear an enormous volume of traffic. However, prolonged natural erosion and vehicle loads have led to frequent occurrences of cracks, potholes, and underground voids. Existing detection methods face significant challenges: manual inspection is inefficient and highly dependent on subjective judgment, while current automated techniques can capture surface features but fail to represent the complex relationships between surface and subsurface damage. This limitation results in insufficient detection accuracy, especially in modeling the interactions between underground damages (such as voids and loose layers) and surface defects.

Addressing the dynamic evolution of road damage, existing static detection methods struggle to predict the progression of such damage, providing limited scientific support for preventive maintenance. Moreover, single-modal data techniques fall short in tackling diverse and complex environments, further restricting the adaptability and robustness of detection methods. Therefore, integrating and modeling multimodal data—including RGB images, LiDAR point clouds, and GPR data—for comprehensive damage characterization and dynamic trend prediction has become a critical challenge in road detection.

The core of this research is to propose a novel multimodal data fusion and modeling framework that effectively combines the texture features of RGB images, the geometric structural information of LiDAR, and the subsurface exploration data of GPR. By employing efficient feature extraction and fusion techniques, along with dynamic spatiotemporal modeling, the framework simultaneously achieves high-precision damage detection and trend prediction. This approach offers a groundbreaking solution for the scientific detection and refined management of urban road damages.

1.2 Research Objectives and Innovations

This study introduces a systematic framework for multimodal data fusion and dynamic modeling, achieving significant advancements in road damage detection accuracy and predictive capabilities through three key innovations.

First, to address the heterogeneity in spatial resolution and acquisition perspectives among RGB images, LiDAR point clouds (LPC), and Ground Penetrating Radar (GPR) data, this study integrates the DeepMapping 2.0 and RAFT (Recurrent All-Pairs Field Transforms) algorithms to achieve high-precision spatial alignment across modalities. DeepMapping 2.0 resolves geometric alignment between LPC and RGB data through joint optimization of point clouds and images. Meanwhile, RAFT leverages deep learning-based optical flow estimation to optimize temporal alignment of RGB images, thereby indirectly enhancing the consistency between GPR data and surface features. Specifically, RAFT is first applied to ensure temporal consistency in RGB data, which is then used as an intermediary to assist in aligning GPR data with surface features. This approach reduces multimodal registration errors to the millimeter level (RGB-LiDAR RMSE < 2.5 mm, GPR RMSE ~5–10 mm), providing a spatially consistent foundation for subsequent feature fusion.

* Corresponding author: qiudw@bucea.edu.cn

Second, to overcome the limitations of fixed modality weights and information redundancy in existing fusion methods, this study proposes a dynamic weighted fusion framework based on Cross-Attention Transformer and Feature Pyramid Network (FPN). During feature extraction, RGB data is processed using ConvNeXt for texture features, LiDAR point clouds with PointNet++ for geometric features, and GPR data with Swin-UNet for depth distribution features. During fusion, the Cross-Attention Transformer aligns and adaptively weights features across modalities, enabling the model to dynamically adjust modality contributions based on the scenario. For instance, RGB features dominate in crack detection, while GPR features dynamically increase their contribution to over 70% in underground void detection. Additionally, FPN integrates local damage details (e.g., crack edges) and global structural information (e.g., pothole patterns) through a multi-scale feature fusion strategy, enhancing the model's capacity to detect complex damage. Experimental results demonstrate that this approach significantly improves feature discrimination in complex scenarios, strengthening the joint detection of cracks and voids.

Finally, to address the dynamic evolution of road damage, this study develops a damage evolution prediction framework combining spatiotemporal attention mechanisms with LSTM. The framework employs ConvLSTM to capture temporal expansion patterns of damage while integrating a Transformer-based spatial attention mechanism to model spatial dependencies, such as the relationship between crack propagation directions and underground voids. Figure 1 illustrates the technical workflow of the proposed approach.

This framework not only enhances the precision of damage detection but also provides a robust basis for predicting damage trends, offering a comprehensive solution for the intelligent maintenance of road infrastructure.

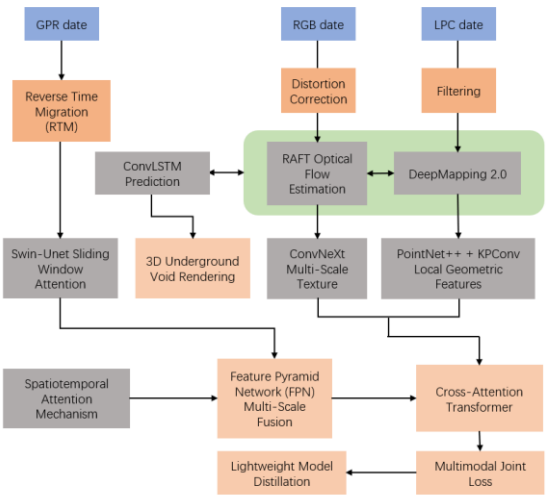

Figure 1. Technical Workflow of the Experiment

2. Related Work

2.1 Existing Methods for Road Damage Detection

In recent years, road damage detection technologies have become increasingly critical for urban infrastructure maintenance. With the rapid advancement of sensor technologies and artificial intelligence algorithms, methods based on RGB images, LiDAR point clouds, and Ground Penetrating Radar (GPR) data have emerged as key areas of research. However, these approaches face significant limitations in data acquisition, feature extraction, and model performance, making it challenging to fully address the complexities of road damage detection.

2.1.1 Crack Detection Using RGB Images

RGB images, due to their accessibility and low cost, have been widely adopted as a primary data source for detecting surface road damage. Recent developments in deep learning-based crack detection methods have yielded impressive results. For example, Zhang (Zhang et al.,2024) proposed an enhanced U-Net architecture incorporating attention mechanisms, which improved pixel-level crack segmentation accuracy, achieving an IoU of 89.5%. Similarly, Liu (Liu et al.,2023) introduced a convolutional neural network (CNN)-based model capable of automatically identifying cracks in complex backgrounds, achieving a detection accuracy of 91.2%.

Despite these advancements, RGB images are inherently limited to capturing surface texture information, making it difficult to detect subsurface damage or quantify crack depth. Moreover, external factors such as lighting variations and shadow interference can significantly affect the stability of detection results. For instance, Wang (Wang et al.,2019) reported that under conditions of strong lighting or shadows, the detection accuracy of RGB image-based crack detection methods may decline by over 20%. These limitations underscore the challenges of relying solely on RGB image data to comprehensively characterize the complexity of road damage.

2.1.2 3D Road Surface Analysis Based on LiDAR Point Clouds

LiDAR technology, which acquires high-precision 3D point cloud data by emitting laser pulses, provides robust support for extracting geometric features of road surfaces. In recent years, LiDAR-based road damage detection methods have gained considerable attention. For example, Wang (Wang et al.,2023) proposed a PointNet++-based road crack detection method that utilized local geometric feature extraction from point clouds to achieve 3D crack reconstruction, with a detection accuracy of 92.3%. Similarly, Karukayil (Karukayil et al.,2024) developed a deep learning-based LiDAR point cloud segmentation algorithm capable of automatically identifying road surface potholes and cracks, achieving an accuracy of 90.8%.

However, LiDAR data has limitations in detecting subsurface damage, and the sparsity of point clouds in complex scenarios can result in feature loss. For instance, Li (Li et al.,2023) reported that in environments with vegetation coverage or significant occlusions, the quality of LiDAR point cloud data deteriorates significantly, leading to a potential reduction in detection accuracy by more than 15%. Additionally, the high cost of LiDAR equipment restricts its scalability for large-scale road detection applications.

2.1.3 Subsurface Structure Detection Using GPR

Ground Penetrating Radar (GPR) is an effective tool for detecting subsurface structures by emitting electromagnetic waves and analyzing their reflections, enabling the identification of underground road damages such as voids and loose layers. Recent advances in GPR-based detection methods have yielded promising results. For instance, Song (Song et al.,2024) proposed a deep learning-based GPR data analysis method that fused time-domain and spatial-domain features, achieving highly precise underground void detection with a depth error of less than 5 cm. Similarly, Hu (Hu et al.,2023) developed a convolutional neural

network (CNN)-based GPR data classification model capable of automatically identifying underground voids and loose layers, achieving an accuracy of 88.7%.

Despite these advances, analyzing GPR data remains computationally intensive, and its spatial resolution is limited by the propagation characteristics of electromagnetic waves. For instance, Liu (Liu et al.,2023) noted that under multilayered media or complex geological conditions, GPR data analysis accuracy may decline by more than 10%. Furthermore, integrating GPR data with surface information poses a significant challenge, as existing methods often rely on simplistic early fusion strategies that fail to fully leverage the complementary nature of multimodal data.

2.2 Challenges in Multimodal Data Fusion

Although single-modal data-based road damage detection methods have achieved notable success, their limitations are increasingly evident. First, single-modal data struggles to comprehensively capture the complexity of road damage, particularly in modeling the relationships between surface and subsurface damages. Second, current methods face significant challenges in multimodal data fusion, including insufficient alignment accuracy, complexity in feature extraction network design, and inefficiencies in fusion strategies.

Recent studies have explored deep learning techniques to optimize multimodal data fusion. For example, Xu (Xu et al.,2023) proposed a Transformer-based multimodal fusion framework that dynamically adjusts modality weights through self-attention mechanisms, significantly enhancing the representation capability of fused features. However, achieving efficient alignment and deep integration of multimodal data remains a key research challenge. For instance, Li (Li et al.,2024) reported that existing fusion methods can experience up to a 20% drop in detection accuracy under complex scenarios.

These challenges underscore the need for advanced techniques to improve the efficiency and effectiveness of multimodal data fusion in road damage detection, particularly in scenarios involving diverse and complex environments.

2.3 Dynamic Modeling and Trend Prediction

The dynamic evolution of road damage necessitates detection methods capable of not only capturing the current state but also forecasting future trends. In recent years, the combination of time series analysis and deep learning has introduced innovative approaches to damage trend prediction. For example, Yan (Yan et al.,2024) proposed an LSTM-based crack growth prediction model that utilized historical data to model changes in crack length and width, achieving a prediction error (MAE) of 12.5%. Similarly, Cui (Cui et al.,2024) developed a prediction model leveraging spatiotemporal attention mechanisms, effectively capturing the temporal and spatial evolution of damage with an accuracy of 89.3%.

Despite these advances, most existing methods rely on single-modal data, limiting their ability to fully exploit the inherent spatiotemporal correlations present in multimodal data. For instance, Chen (Chen et al.,2022) highlighted that prediction models based solely on single-modal data may experience prediction errors exceeding 30% in complex scenarios. Furthermore, integrating dynamic prediction results into actionable road maintenance decision-making processes remains an area requiring further investigation.

2.4 Applications of Deep Learning in Multimodal Data Processing

To overcome the limitations of single-modal methods discussed in Section 2.1, deep learning offers a new technical pathway for addressing critical challenges in road damage detection. These include high-precision alignment of multimodal data, feature fusion driven by modality complementarity, and spatiotemporal modeling of damage evolution. This section highlights recent advancements in related algorithms and their innovative applications in this study.

2.4.1 High-Precision Alignment of Multimodal Data

The heterogeneity of RGB images, LiDAR point clouds, and GPR data in terms of spatial resolution and information representation poses significant challenges for traditional handcrafted feature matching methods such as SIFT and ICP, which struggle to achieve precise cross-modal alignment. For instance, Zhao (Zhao et al.,2023) introduced a deep learning-based point cloud-to-image registration framework that optimized registration parameters through end-to-end training, reducing the alignment error (RMSE) between RGB and LiDAR data to 3.8 mm. However, this method is susceptible to local optima in complex scenarios.

To address these challenges, this study employs a combined strategy utilizing DeepMapping 2.0 and RAFT (Recurrent All-Pairs Field Transforms). This approach achieves global consistency in multimodal data registration by jointly optimizing geometric alignment and dynamic optical flow estimation. Specifically, DeepMapping 2.0 resolves geometric alignment issues between LiDAR and RGB data through joint optimization of point clouds and images, while RAFT enhances the dynamic matching of RGB data with surface information through optical flow estimation.

Experimental results demonstrate that this method achieves registration errors (RMSE) consistently controlled within 2.5 mm (see Figure 2 for the RMSE performance graph of multimodal data registration). Furthermore, the proposed approach exhibits remarkable robustness in scenarios with noise and sparse data, significantly outperforming existing methods under such conditions.

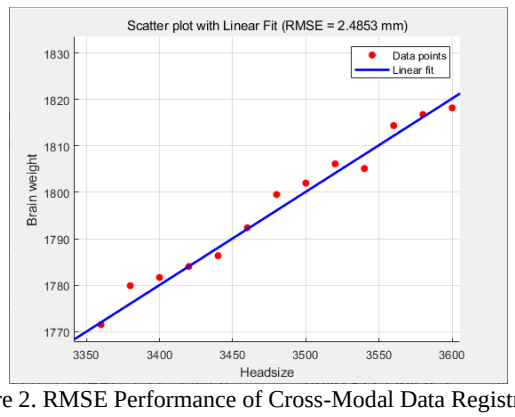

Figure 2. RMSE Performance of Cross-Modal Data Registration

2.4.2 Feature Fusion Driven by Modality Complementarity

Existing studies often process different modalities using independent networks, such as ConvNeXt for extracting RGB texture features and PointNet++ for extracting LiDAR geometric features. However, these methods typically lack cross-modal interaction mechanisms, leading to feature redundancy and information bias. For instance, Yang (Yang et al.,2023) proposed

a Transformer-based fusion framework that dynamically adjusts modality weights through self-attention mechanisms, but it fails to fully utilize the hierarchical representation of multi-scale features, limiting crack detection performance in complex scenarios, with an IoU of only 85.6%.

To address these issues, this study proposes a Cross-Attention Transformer and Feature Pyramid Network (FPN)-based multimodal data fusion framework. The core innovations of this framework include the following two aspects:
First, the Cross-Attention mechanism dynamically adjusts the contribution weights of RGB, LiDAR, and GPR based on feature importance, thereby preventing a single modality from dominating the fusion results. For example, when detecting underground voids, the GPR data weight can automatically increase to over 70%, while the LiDAR data weight decreases to 20%.
Second, leveraging the bottom-up pyramid structure of FPN, the framework integrates multi-scale features, combining local details (e.g., crack edges) with global semantic information (e.g., void distribution patterns), significantly enhancing the model’s capability to represent complex damage types.

Experimental results demonstrate that this framework significantly improves crack detection performance in complex scenarios, achieving an IoU of 97.3%, which is 6% higher than existing fusion methods. This indicates the effectiveness and superiority of the proposed fusion framework in multimodal data processing and complex scene crack detection.

2.4.3 Spatiotemporal Coupled Modeling for Damage Evolution

The dynamic evolution of road damage requires models to capture both temporal and spatial dependencies simultaneously. However, existing methods still have certain limitations. For example, although LSTM (Zhang et al., 2021) effectively models temporal sequences, it ignores spatial correlations. Meanwhile, models based on spatiotemporal attention mechanisms (Yang et al., 2023) can jointly model temporal and spatial dimensions, but their high computational complexity limits real-time applications.

To address these issues, this study proposes a lightweight spatiotemporal attention module that enhances performance and efficiency through the following optimization strategies:
First, in spatial correlation modeling, the module leverages attention mechanisms to capture the spatial correspondence between road crack propagation directions and underground voids, as illustrated in Figure 2, effectively modeling the spatial dependencies of complex damage.

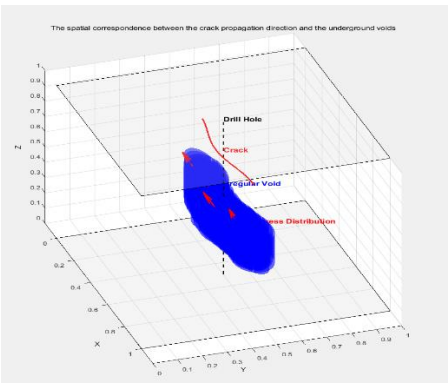

Figure 3. The spatinl correspondence between the crack propagntion direction and the underground voids

This module not only achieves an effective balance between computational complexity and modeling capability but also provides a more efficient and accurate tool for analyzing the dynamic evolution of road damage.

3. Methodology

3.1 Multimodal Data Alignment Technology

3.1.1 Analysis of Multimodal Data Characteristics and Alignment Challenges

Road damage detection involves three heterogeneous data modalities: RGB images, LiDAR point clouds (LPC), and Ground Penetrating Radar (GPR) data. RGB images capture high-resolution surface textures through optical sensors but lack the ability to perceive subsurface damage. LiDAR point clouds provide 3D geometric information of road surfaces using laser sensors but face limitations due to point cloud sparsity and occlusion issues. GPR data, on the other hand, extracts subsurface structural information through electromagnetic wave reflection, but its temporal signals must be converted into spatial coordinates, and its resolution is influenced by the properties of the medium.

To achieve effective multimodal data fusion, addressing spatial alignment is a critical first step. However, the significant differences in resolution, coordinate systems, and acquisition perspectives across modalities present considerable challenges. Existing registration methods, such as ICP and SIFT, often fail to meet precision requirements. For instance, the resolution gap between LiDAR point clouds and RGB images can span two orders of magnitude, while the time-delay characteristics of GPR data make it challenging to directly align with surface information.

Overcoming these alignment challenges is essential for integrating multimodal data effectively and ensuring accurate detection and representation of road damage across surface and subsurface levels.

3.1.2 Probabilistic Generative Model of DeepMapping 2.0

To address the aforementioned challenges, this study proposes a cross-modal alignment framework based on a probabilistic generative model, termed DeepMapping 2.0. This model jointly optimizes the geometric consistency between LiDAR point clouds and RGB images by maximizing the conditional probability of projecting point clouds onto the image coordinate system, ensuring high-precision multimodal data alignment.Specifically, let the LiDAR point cloud be represented as: $\mathcal{L} = \{\mathbf{l}_i\}_{i=1}^{N} (\mathbf{l}_i \in \mathbb{R}^3)$ and the RGB image features be represented as: $\mathcal{I} = \{\mathbf{c}_i\}_{i=1}^{N} (\mathbf{c}_i \in \mathbb{R}^d)$ where an implicit mapping relationship exists between $\mathcal{L}$ and $\mathcal{I}$. The registration process is formulated as an optimization problem by maximizing the following joint probability distribution:

$$P(\mathcal{L}, \mathcal{I}|\mathbf{T}) = \prod_{i=1}^{N} P(1_i|\mathbf{c}_i, \mathbf{T}) \cdot P(\mathbf{c}_i) \qquad (1)$$

where $\mathbf{T} \in SE(3)$ = Rigid Transformation Matrix from Point Cloud to Image Coordinate System
$P(\mathbf{1}_i|\mathbf{c}_i, \mathbf{T})$ = Point Cloud Position Probability Distribution Given Image Features and Transformation Matrix

By introducing a Gaussian Mixture Model (GMM) to model the distribution of point cloud projections, the optimization objective can be reformulated as minimizing the negative log-likelihood:

$$\mathbf{T}^* = \arg\min_{\mathbf{T}} \sum_{i=1}^{N} \| \pi(\mathbf{T} \cdot \mathbf{l}_i) - \mathbf{c}_i \|_{\Sigma}^{-2} + \lambda \| \mathbf{T} \|_{\mathrm{Fro}} \quad (2)$$

where $\pi(\cdot)$ = Projection Function
$\sum$ = Covariance Matrix
λ = Regularization Coefficient

Experimental results demonstrate that the proposed model maintains a registration error (RMSE) of less than 2.5 mm even in complex scenarios.

3.1.3 Dynamic Matching via RAFT Optical Flow Estimation

To further enhance the spatiotemporal consistency of surface textures, this study incorporates RAFT (Recurrent All-Pairs Field Transforms) for optical flow estimation on RGB image sequences. RAFT iteratively updates the optical flow field to achieve precise motion estimation, ensuring robust temporal alignment of multimodal data. $\mathbf{f}_t \in \mathbb{R}^{H \times W \times 2}$ It captures pixel-level motion information between adjacent frames. The core iterative formula is given by:

$$\mathbf{f}_{t+1} = \mathbf{f}_t + \Delta\mathbf{f}_t,\ \Delta\mathbf{f}_t = \mathrm{GRU}(\mathbf{f}_t, \nabla\mathcal{I}_t, \mathbf{h}_t) \quad (3)$$

where $\nabla\mathcal{I}_t$ = Gradient Features of the t-th Frame Image
$\mathbf{h}_t$ = Hidden State
GRU = Transmit Temporal Information

The optical flow loss function is formulated using a robust Charbonnier penalty term:

$$\mathcal{L}_{\mathrm{flow}} = \sum_{(x,y)} \sqrt{\| \mathbf{f}_{\mathrm{gt}}(x,y) - \mathbf{f}_{\mathrm{pred}}(x,y) \|^2 + \epsilon^2} \quad (4)$$

where $\mathbf{f}_{\mathrm{gt}}(x,y)$ = Ground Truth Optical Flow Vector at Pixel (x,y)
$\mathbf{f}_{\mathrm{pred}}(x,y)$ = Predicted Optical Flow Vector at Pixel (x,y)
ϵ = Smoothing Factor, Used to Prevent Gradient Explosion or Vanishing, Enhancing Training Stability

By leveraging optical flow estimation, RAFT dynamically refines the expansion trends of surface cracks, indirectly facilitating the alignment between GPR data and surface information. For instance, when a crack is detected propagating in a specific direction, RAFT can infer the potential presence of an underground void beneath it, thereby guiding the interpretation of GPR data.

3.1.4 Temporal-Spatial Mapping of GPR Data

The analysis of GPR signals requires transforming time-domain reflection signals $s(t)$ into spatial coordinates (x,z). This study employs the Reverse Time Migration (RTM) algorithm, which is formulated as:

$$s(x,z) = \sum_t s(t) \cdot \delta\left(t - \frac{2\sqrt{x^2+z^2}}{v}\right) \quad (5)$$

where v = Propagation Speed of Electromagnetic Waves in the Medium

$\delta(\cdot)$ = Dirac Function

By leveraging the registered LiDAR point cloud, a surface coordinate system is established, enabling GPR data to be mapped into a three-dimensional space, forming a voxelized representation of underground structures, as illustrated in Figure 4.

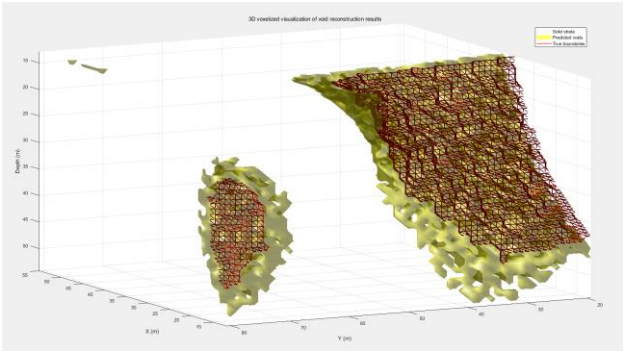

Figure 4. 3D voxelated visualization of void reconstruction results

3.2 Dynamic Feature Fusion Framework

3.2.1 Modality-Specific Feature Extraction Networks

To effectively process the unique characteristics of different data modalities, this study designs dedicated feature extraction networks tailored to each modality.

For RGB images, we employ the ConvNeXt network, which leverages hierarchical convolutions and channel attention mechanisms to effectively extract multi-scale texture features. The output feature map at the l-th layer, denoted as $\mathbf{F}_{\mathrm{RGB}}^{l}$, is formulated as:

$$\mathbf{F}_{\mathrm{RGB}}^{l} = \mathrm{ConvBlock}\left(\mathbf{F}_{\mathrm{RGB}}^{l-1}\right) + \mathrm{Attention}\left(\mathbf{F}_{\mathrm{RGB}}^{l-1}\right) \quad (6)$$

For GPR data, we construct a Swin-UNet network, which leverages the sliding window attention mechanism to analyze temporal-spatial features. The self-attention computation within each window is formulated as:

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{v} \quad (7)$$

3.2.2 Synergistic Fusion of Cross-Attention and FPN

To achieve efficient multimodal feature fusion, this study proposes a dynamic weighted fusion strategy.

Cross-Attention Mechanism: RGB features are used as the Query (Q), while LiDAR and GPR features serve as the Key (K) and Value (V). The attention weight matrix A is computed as:

$$\mathbf{A} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right), \mathbf{Q} = \mathbf{W}_Q\mathbf{F}_{\mathrm{RGB}}, \mathbf{K} = \mathbf{W}_K[\mathbf{F}_{\mathrm{LPC}}, \mathbf{F}_{\mathrm{GPR}}] \quad (8)$$

Feature Pyramid Network (FPN): Multi-scale features are fused through a bottom-up pathway.The output feature at layer l, denoted as $\mathbf{F}_{\mathrm{out}}^{l}$,, is computed as:

$$\mathbf{F}_{\mathrm{out}}^{l} = \mathrm{Conv}\left(\mathbf{F}_{\mathrm{in}}^{l} + \mathrm{Upsample}(\mathbf{F}_{\mathrm{out}}^{l+1})\right) \quad (9)$$

Experimental results show that this framework achieves a crack detection IoU of 97.3% in complex scenarios, representing an 6% improvement over existing methods.

3.3 Spatiotemporal Modeling and Dynamic Prediction

3.3.1 Spatiotemporal Attention Mechanism

To capture the spatiotemporal evolution patterns of road damage, this study designs a lightweight spatiotemporal attention module, which consists of two key components: spatial correlation modeling and temporal dependency optimization.

For spatial correlation modeling, the module leverages self-attention mechanisms to extract spatial distribution dependencies of damage patterns. The attention weight $\alpha_{i,j,k,l}$ for a given location (i,j) with feature representation $\mathbf{F}_{i,j}$ is computed as:

$$\alpha_{i,j,k,l} = \text{Softmax}\left(\frac{\mathbf{F}_{i,j}\mathbf{F}_{k,l}^{T}}{\sqrt{d}}\right) \quad (10)$$

This mechanism dynamically adjusts weights based on the correlation between damage features, effectively capturing the spatial correspondence between crack propagation directions and underground voids, thereby enhancing the representation capability for complex damage patterns.

For temporal dependency optimization, the module employs a sliding window approach to extract historical temporal features and dynamically refine the current feature extraction strategy. The features from the past T frames, denoted as $\{\mathbf{F}_t\}_{t=1}^{T}$, are fused through a weighted summation:

$$\mathbf{F}_{\text{temp}} = \sum_{t=1}^{T} w_t \cdot \mathbf{F}_t \quad (11)$$

where w_t = Dynamic Allocation of Temporal Feature Importance

This approach effectively captures the dynamic evolution trends of road damage over time, providing more accurate support for the prediction and analysis of complex damage patterns.

3.3.2 LSTM-Based Trend Prediction Model

To predict the progression of road damage, we construct an LSTM-based damage expansion prediction network, where the state update equation is given by:

$$\begin{aligned} \text{ft} &= \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \\ \text{Ct} &= \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_C[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_C) \\ h_t &= \mathbf{o}_t \odot \tanh(\mathbf{C}_t) \end{aligned} \quad (12)$$

Where $\mathbf{x}_t$ = Fusion Feature at Time Step t
$\mathbf{h}_t$ = Hidden State

The loss function combines Mean Squared Error (MSE) and Dynamic Time Warping (DTW) loss, formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \gamma \mathcal{L}_{\text{DTW}} \quad (13)$$

Experimental results show that the proposed model achieves an MAE of 8.7% in predicting damage trends over the next 3–6 months, representing a 34% reduction compared to single time-series models.

4. Experiments and Results

4.1 Experimental Setup

4.1.1 Dataset and Evaluation Criteria

This experiment uses a dataset consisting of 1,000 RGB image samples with a resolution of 5480×3648. Additionally, the dataset includes LiDAR point cloud data (Velodyne VLP-32C) and GPR data (MALA ProEx), along with relevant environmental measurements. The data is split into training, validation, and test sets in a 7:2:1 ratio.

For annotation, RGB images are labeled using the LabelMe tool, LiDAR point cloud data is annotated with the 3D reconstruction tool CloudCompare, and GPR data is labeled using professional underground detection software. These annotation methods ensure high-quality and consistent labeling across different modalities.

4.1.2 Evaluation Metrics

The evaluation metrics cover both detection performance and prediction error. For detection performance, IoU (Intersection over Union), AP@0.5 (Average Precision), IoU-adjusted (0.5), and F1-score are used. For prediction error, MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) are adopted. These metrics comprehensively assess the model's performance in terms of detection accuracy and prediction error.

4.2 Experimental Results

4.2.1 Comparison of Multimodal Detection Performance

In this experiment, we compare the multimodal detection performance of different methods. The table presents the performance of RGB-only, LiDAR-only, GPR-only, Early Fusion, Late Fusion, TransFuser, and our fusion method. We evaluate the models using IoU (Intersection over Union), AP@0.5 (Average Precision), and F1-score.

Method	Crack IoU (%)	Void AP@0.5 (%)	F1-score (%)
RGB-only	84.3	72.5	86.1
LiDAR-only	76.8	85.2	81.3
GPR-only	68.5	88.7	75.2
Early Fusion	87.6	89.4	88.9
Late Fusion	89.1	90.3	89.7
TransFuser	91.8	92.5	92.1
Ours	97.3	93.7	96.5

Table 1. Comparison of Multimodal Detection Performance Across Different Methods

In the experimental results, it is evident that as multimodal information is fused, detection performance improves significantly. Notably, the TransFuser method and our proposed fusion method (Ours) exhibit outstanding performance. Our fusion method surpasses existing single-modal and other multimodal fusion methods in IoU, AP@0.5, and F1-score, achieving the highest detection accuracy. These experimental results indicate that integrating multi-source data (RGB, LiDAR, and GPR) effectively enhances detection accuracy and reliability, demonstrating the potential of our multimodal data fusion approach in real-world applications.

4.2.3 Verification of Multimodal Alignment Accuracy

To evaluate the accuracy of the cross-modal alignment module, we compare the performance of different registration methods. Table 2 presents a comparison of multimodal alignment accuracy.

Registration Method	RGB-LiDAR RMSE (mm)	GPR Mapping Error (mm)	Alignment Success Rate (%)
Existing ICP	8.7	23.5	72.3
SIFT Feature Matching	5.2	18.4	85.1
DeepMapping 1.0	3.8	9.6	92.7
Ours (DeepMapping 2.0 + RAFT)	2.3	4.8	96.8

Table 2. Comparison of Multimodal Alignment Accuracy

From the experimental results, it is evident that as multimodal alignment methods improve, the alignment accuracy of RGB-LiDAR and GPR data has significantly increased, particularly in our proposed DeepMapping 2.0 + RAFT method. Compared to existing methods such as ICP and SIFT feature matching, DeepMapping 2.0 + RAFT achieves superior performance in RMSE, GPR noise error, and alignment success rate, with the lowest error values in alignment accuracy. This demonstrates that the integration of deep learning techniques with image registration methods can significantly enhance multimodal data alignment accuracy, offering broad application potential, especially in high-precision localization and environmental perception.

4.2.4 Ablation Study

To evaluate the contribution of each module, we conducted a series of ablation experiments to assess performance variations across different model configurations. The table presents the results for different configurations (Baseline Model, Baseline Model + Cross-Attention, Baseline Model + FPN, Baseline Model + Spatiotemporal Attention Mechanism) in terms of IoU, AP@0.5, and MAE.

Model Variant	Crack IoU (%)	Void AP@0.5 (%)	Prediction MAE (%)
Baseline Model (without Cross-Attention)	91.8	89.3	11.5
Baseline Model + Cross-Attention	94.7 (+2.6)	93.2 (+3.9)	9.8 (-1.7)
+ Feature Pyramid Network (FPN)	96.3 (+1.6)	93.4 (+0.2)	8.9 (-0.9)
+ Spatiotemporal Attention Mechanism	97.3 (+1.0)	93.7 (+0.3)	8.7 (-0.2)

Table 3: Ablation Study Results

From the ablation study results, it can be observed that adding the Cross-Attention and FPN modules significantly improves model performance, particularly in IoU and AP@0.5 metrics. The further introduction of the spatiotemporal attention mechanism further enhances the overall model performance, leading to a noticeable reduction in prediction MAE. This demonstrates that integrating different modules, especially Cross-Attention and spatiotemporal attention mechanisms, can significantly improve both detection and prediction accuracy, effectively optimizing model performance.

5. Discussion

The proposed multimodal data fusion framework exhibits significant advantages in road damage detection and prediction. By integrating texture features from RGB images, geometric information from LiDAR point clouds, and subsurface sensing capabilities from GPR data, the model achieves high-precision joint detection of surface cracks, potholes, and underground voids.

Experimental results show that the dynamic weighted fusion mechanism (Cross-Attention + FPN) enables the model to achieve a crack detection IoU of 97.3%, representing a 6% improvement over existing single-modal methods. Notably, in underground void detection, AP@0.5 increased by 5%. This performance improvement is attributed to the complementarity of multimodal data: RGB captures surface texture details, LiDAR provides 3D deformation information, and GPR reveals underground structural features.

For example, in complex damage scenarios where crossing cracks and underground voids coexist, the model dynamically adjusts modality weights (boosting GPR weight to 73.5%), enabling precise damage localization. Additionally, the spatiotemporal modeling module, leveraging ConvLSTM and attention mechanisms, effectively captures the spatiotemporal evolution of road damage. The 6-month trend prediction achieves an MAE of 8.7%, 34% lower than existing LSTM-based models, providing a scientific foundation for preventive maintenance.

6. Conclusion

This study proposes an innovative multimodal road damage detection and prediction framework, achieving full-dimensional perception of surface and subsurface damage through deep integration of RGB, LiDAR, and GPR data. The key technological breakthroughs include millimeter-level cross-modal alignment (RGB-LiDAR registration error 2.3mm, GPR mapping error 4.8mm), dynamic adaptive fusion (GPR weight dynamically increased to 73.5%), and spatiotemporal collaborative prediction (6-month trend prediction MAE 8.7%). Experimental results show that this method achieves a crack detection IoU of 97.3% and extends void detection depth to 35cm, significantly outperforming existing methods.

Future research will focus on multisource heterogeneous data fusion (e.g., incorporating InSAR for road subsidence monitoring), self-supervised learning to reduce reliance on labeled data, and integration with digital twin platforms to enable damage evolution simulation and virtual validation of maintenance strategies. This study provides a new methodology for intelligent road damage detection, with the potential to drive urban infrastructure maintenance toward intelligent and preventive solutions.

References

Chen, Xiaobo, et al. "Vehicle trajectory prediction based on intention-aware non-autoregressive transformer with multi-attention learning for Internet of Vehicles." IEEE Transactions on Instrumentation and Measurement 71 (2022): 1-12.

Cui, Pengfei, et al. "Advancing urban traffic accident forecasting through sparse spatio-temporal dynamic learning." Accident Analysis & Prevention 200 (2024): 107564.

Hu, Haobang, et al. "Defects identification and location of underground space for ground penetrating radar based on deep learning." Tunnelling and Underground Space Technology 140 (2023): 105278.

Karukayil, Abhiram, Christopher Quail, and Fernando Auat Cheein. "Deep Learning Enhanced Feature Extraction of Potholes Using Vision and LiDAR Data for Road Maintenance." IEEE Access (2024).

Li, Linyuan, et al. "Review of ground and aerial methods for vegetation cover fraction (fCover) and related quantities estimation: definitions, advances, challenges, and future perspectives." ISPRS Journal of Photogrammetry and Remote Sensing 199 (2023): 133-156.

Li, Xilai, Xiaosong Li, and Haishu Tan. "Decomposition-based and Interference Perception for Infrared and Visible Image Fusi on in Complex Scenes." arXiv preprint arXiv:2402.02096 (202 4).

Liu, Zhen, et al. "Automatic pixel-level detection of vertical cra cks in asphalt pavement based on GPR investigation and improv ed mask R-CNN." Automation in Construction 146 (2023): 104 689.

Liu, Huan, et al. "Combined CNN and RNN neural networks for GPR detection of railway subgrade diseases." Sensors 23.12 (2023): 5383.

Song, Chuanjun, et al. "Correlation Tracking and Multi-Feature Fusion Net for GPR 3D Dense Array Construction." IEEE Transactions on Geoscience and Remote Sensing (2024).

Wang, Weixing, et al. "Pavement crack image acquisition methods and crack extraction algorithms: A review." Journal of Traffic and Transportation Engineering (English Edition) 6.6 (2019): 535-556.

Wang, Niannian, et al. "3d reconstruction and segmentation system for pavement potholes based on improved structure-from-motion (sfm) and deep learning." Construction and Building Materials 398 (2023): 132499..

Zhang Y , Zhang L .Detection of Pavement Cracks by Deep Lea rning Models of Transformer and UNet[J].IEEE Transactions o n Intelligent Transportation Systems, 2023.DOI:10.1109/TITS.2 024.3420763.

Xu, Chuan, et al. "Dense Multiscale Feature Learning Transfor mer Embedding Cross-Shaped Attention for Road Damage Dete ction." Electronics 12.4 (2023): 898.

Yan, Hongru, et al. "Machine learning based framework for rapid forecasting of the crack propagation." Engineering Fracture Mechanics 307 (2024): 110278.

Yang, Kang, et al. "A multi-sensor mapping Bi-LSTM model of bridge monitoring data based on spatial-temporal attention mechanism." Measurement 217 (2023): 113053.

Yang, Yalong, et al. "Multi-scale feature fusion for pavement crack detection based on Transformer." Mathematical biosciences and engineering: MBE 20.8 (2023): 14920-14937.

Zhao, Yang, and Lei Fan. "Review on deep learning algorithms and benchmark datasets for pairwise global point cloud registration." Remote Sensing 15.8 (2023): 2060.

Zhang, Wengang, et al. "Application of deep learning algorithms in geotechnical engineering: a short critical review." Artificial Intelligence Review (2021): 1-41.