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

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1077-1092
Journal homepage: http://iieta.org/journals/ts

Intelligent Management of Public–Private Partnership Infrastructure Integrating Digital Twins and Three-Dimensional Image Reconstruction

Jinglei Meng

School of Economics and Management, Harbin University, Harbin 150086, China

Corresponding Author Email: mjl@hrbu.edu.cn

https://doi.org/10.18280/ts.430301

Received: 2 January 2026
Revised: 28 May 2026
Accepted: 12 June 2026
Available online: 30 June 2026

Keywords:
digital twin, three-dimensional image reconstruction, neural radiance field, scene coordinate regression, vision-language model, public–private partnership infrastructure

ABSTRACT

Infrastructure operated under public–private partnership modes typically features prolonged service lives and complex operational environments. Structural performance degradation is frequently induced by long-term loading and natural erosion. Conventional manual inspection and experience-driven maintenance decision-making are characterized by low efficiency and high subjectivity, rendering them inadequate for the refined management demands of large-scale infrastructure assets. When current digital twin and three-dimensional reconstruction techniques are applied to infrastructure maintenance, several technical deficiencies persist, including insufficient reconstruction accuracy under weakly textured scenes, weak robustness of cross-temporal virtual–real mapping, and a lack of deep semantic reasoning in defect detection. To address these challenges, an integrated intelligent maintenance technical framework was constructed, in which a digital twin management architecture for public–private partnership infrastructure was established, encompassing image acquisition, three-dimensional reconstruction, virtual–real mapping, and intelligent diagnosis. Feature representation was optimized via a frequency-domain adaptive high-frequency enhancement method, and reconstruction accuracy for weakly textured structures under sparse views was improved by incorporating a geometry-constrained neural radiance field algorithm. A scene coordinate regression network, built upon a dual-path frequency-domain and spatial-domain feature fusion mechanism, was developed to enable high-precision pixel-level three-dimensional localization across temporally disparate images. Lightweight dynamic updating of the digital twin model was achieved through adaptive threshold-based change detection combined with a local incremental update strategy. Furthermore, two-dimensional defect perception results were elevated to three-dimensional space, and an intelligent semantic diagnosis framework for structural defects was formulated by integrating retrieval-augmented generation techniques with vision-language models, thereby producing standardized and actionable maintenance analysis reports. This study overcomes multiple technical bottlenecks in image-driven digital twin-based infrastructure maintenance, and provides a comprehensive and feasible technical solution for the intelligent and refined lifecycle management of public–private partnership infrastructure.

1. INTRODUCTION

The public–private partnership model has become the predominant framework for the development and operation of major infrastructure projects in transportation, water resources, and municipal engineering, occupying a strategically significant position within the global infrastructure landscape [1, 2]. Large-scale infrastructure undertakings are generally subject to long-term operational mechanisms, with service lives extending from twenty to thirty years. Prolonged exposure to complex natural environments and sustained traffic loading frequently induces material aging, localized damage, and performance deterioration in structural systems. The operational stability of such infrastructure is directly tied to public safety and the orderly progression of socioeconomic activities [3, 4]. Conventional infrastructure maintenance relies on periodic manual inspections and experience-based decision-making systems, which are characterized by limited detection efficiency and considerable subjectivity in evaluation criteria, rendering them inadequate for the refined and routine maintenance management demands of large-scale infrastructure clusters [5]. Digital twin technology enables the establishment of dynamic mapping relationships between physical entities and their virtual counterparts, thereby offering a novel paradigm for intelligent lifecycle control of infrastructure. Three-dimensional image reconstruction techniques, which recover geometric structures and textural features from two-dimensional visual data, constitute the core enabling technology underpinning the construction and dynamic updating of digital twin models [6, 7]. Although the deep integration of these two technological domains provides a promising pathway for intelligent infrastructure maintenance, significant deficiencies persist within this technical framework during practical engineering deployment. These limitations

hinder its adaptability to the weakly textured, large-scale, and long-temporal-sequence operational scenarios characteristic of public–private partnership infrastructure, thereby substantially constraining the large-scale implementation and engineering application of intelligent management technologies [8, 9].

Substantial research progress has been achieved in the fields of three-dimensional reconstruction, digital twin updating, and intelligent defect diagnosis. Nevertheless, existing methodologies remain inadequate for meeting the high-precision, high-robustness, and automated maintenance requirements of public–private partnership infrastructure [10, 11]. In the domain of three-dimensional reconstruction, mainstream neural radiance field algorithms exhibit excellent continuous scene modeling capabilities; however, under sparse-view sampling conditions, overfitting phenomena are readily induced, preventing the precise recovery of fine-scale geometric features in weakly textured regions such as concrete and steel structures within infrastructure scenes. Conventional three-dimensional reconstruction methods, by contrast, are hampered by feature matching failures and are consequently unable to accomplish complete and accurate modeling of large-scale, open infrastructure environments [12, 13]. In the domain of digital twin model updating, existing model iteration approaches rely predominantly on light detection and ranging resampling and manual data collection, resulting in overall low automation levels and prohibitive maintenance costs. The few lightweight updating methods based on visual registration perform image matching using only single-domain spatial features, exhibiting poor adaptability to temporal environmental variations such as seasonal changes and illumination fluctuations. Consequently, cross-temporal registration accuracy undergoes substantial volatility, rendering these methods insufficient to satisfy the precision standards required for engineering applications [14, 15]. In the domain of intelligent structural diagnosis, current deep learning-based visual detection techniques enable precise localization and classification of infrastructure defects. However, only shallow visual recognition outputs are produced, without the capability to quantify three-dimensional geometric parameters of defects or to integrate with technical codes for condition rating and maintenance strategy derivation. This creates a technical gap between visual perception outcomes and engineering decision-making applications [16, 17]. Furthermore, existing research on visual infrastructure maintenance is largely confined to single visible-light data sources, with insufficient exploitation of multi-source visual information fusion. An adaptive feature fusion strategy suitable for all-weather complex maintenance scenarios has yet to be formulated, substantially limiting the environmental adaptability and scene generalization performance of intelligent maintenance systems [18, 19].

To address the aforementioned industry pain points and technical deficiencies, an integrated intelligent management framework that fuses digital twin and three-dimensional reconstruction technologies is constructed for public–private partnership infrastructure. A three-dimensional reconstruction strategy synergizing frequency-domain enhancement and geometric constraints is proposed, through which the insufficient modeling accuracy under sparse-view and weakly textured scenarios is effectively overcome. A coordinate regression network integrating dual-path frequency-domain and spatial-domain features is established, and, in conjunction with adaptive change detection and local incremental updating mechanisms, high-robustness and lightweight dynamic iteration of the digital twin model are realized. By combining three-dimensional geometric mapping with retrieval-augmented vision-language models, the complete technical chain from two-dimensional visual perception to three-dimensional semantic diagnosis is established, thereby producing standardized engineering maintenance decision outputs. Furthermore, a multi-scenario, cross-temporal annotated dataset for public–private partnership infrastructure is constructed, providing reliable data support for algorithmic iteration and performance evaluation within the research community.

The study is organized into five chapters, with a progressive and coherent logical structure throughout. In Chapter 1, the research background, existing technical limitations, and core innovations are systematically delineated, and the research value and overall framework are clarified. In Chapter 2, frontier research achievements in the field are reviewed, and the applicable boundaries and principal shortcomings of existing methodologies are defined. In Chapter 3, the overall architecture of the proposed intelligent management approach and the technical principles of each constituent module are elaborated in detail. In Chapter 4, multiple sets of comparative and ablation experiments are designed, through which the effectiveness and superiority of the proposed method are quantitatively validated. In Chapter 5, the research findings are summarized, current technical constraints are analyzed, and future research directions are discussed.

2. RELATED WORK

2.1 Image-based three-dimensional reconstruction methods

Image-based three-dimensional reconstruction constitutes the core enabling technology for digital modeling of infrastructure, with its algorithmic framework having undergone iterative evolution from traditional geometric modeling to data-driven implicit modeling [20, 21]. Conventional modeling methods are centered on structure-from-motion and multi-view stereo matching, relying on hand-crafted visual features for image matching and scene reconstruction. Good applicability is demonstrated in conventional scenes characterized by rich textures and dense viewpoints. However, infrastructure surfaces are predominantly composed of homogeneous concrete and steel materials, where effective visual features are scarce. Consequently, feature matching failures, incomplete three-dimensional point clouds, and structural distortions are readily induced, rendering these methods insufficient for meeting engineering modeling accuracy requirements [22, 23]. Deep learning-driven multi-view stereo matching algorithms have enhanced adaptability to weakly textured scenes through learned feature extraction; nevertheless, dense viewpoint image inputs remain a prerequisite for these approaches, and modeling performance deteriorates substantially under sparse sampling conditions [24, 25]. Neural radiance fields, leveraging the advantages of continuous implicit scene representation, have transcended the inherent limitations of traditional discrete modeling and have become the predominant technique in the current three-dimensional reconstruction landscape [12, 26]. Although existing variants of neural radiance field algorithms can optimize rendering and

reconstruction outcomes, severe overfitting issues are commonly encountered under sparse-view constraints, preventing accurate inference of geometric structures in unseen scene regions. Moreover, the structural prior information inherent to regularized infrastructure geometries has not been incorporated, rendering these methods poorly suited to the large-scale, weakly textured, and sparse-sampling inspection conditions characteristic of public–private partnership infrastructure [13, 27]. In the proposed approach, the deficiency of weakly textured features is compensated through a frequency-domain feature enhancement mechanism, while engineering geometric regularization constraints are introduced to optimize the reconstruction logic, thereby effectively overcoming the scene adaptability bottlenecks of existing reconstruction algorithms.

2.2 Scene coordinate regression and digital twin updating

Scene coordinate regression technology has reshaped the updating logic of visual localization and digital twin models, whereby the cumbersome pipeline of sequential matching and optimization is obviated. An end-to-end mapping relationship between two-dimensional image pixels and three-dimensional scene coordinates is directly established, providing a novel technical paradigm for the automated iteration of infrastructure digital twin models [28, 29]. Current mainstream scene coordinate regression models rely on spatial-domain visual features as the core representational basis, enabling high-precision pixel-level localization in scenes with stable illumination and temporally consistent conditions. Preliminary applications have been demonstrated in three-dimensional registration tasks for indoor and outdoor scenes [30]. However, the operational lifecycle of public–private partnership infrastructure spans several decades, during which inspection data are subject to significant cross-seasonal and cross-temporal illumination variations as well as environmental interferences. Reliance solely on single-domain spatial features proves insufficient for resisting environmental noise, resulting in inadequate registration robustness and substantially escalated cross-temporal localization errors. Consequently, the long-term virtual–real consistency of digital twin models is difficult to sustain [14, 15]. Furthermore, existing digital twin updating approaches predominantly adopt global reconstruction strategies, incurring prohibitively high computational costs and hardware overhead. The few lightweight updating schemes lack precise change region awareness, preventing targeted model iteration [31]. In the proposed approach, a dual-path adaptively fused frequency-domain and spatial-domain coordinate regression network is constructed, through which environmental noise resistance is reinforced from the underlying signal dimension. In conjunction with adaptive change detection and local incremental updating mechanisms, high-precision and low-cost dynamic maintenance of digital twin models is achieved.

2.3 Vision-language models and intelligent infrastructure diagnosis

The rapid advancement of multi-modal vision-language models has propelled infrastructure defect analysis from shallow visual recognition toward deep semantic intelligent diagnosis, offering novel avenues for addressing the intelligence challenges in engineering maintenance decision-making [32, 33]. Current research has applied general-purpose vision-language models to infrastructure defect detection scenarios, whereby shallow tasks such as defect classification and basic attribute description are accomplished using image-derived visual information, thereby replacing traditional manual identification and simple algorithmic detection [33, 34]. Nevertheless, significant application limitations persist in existing technical solutions. Most models perform semantic reasoning solely on the basis of two-dimensional planar images, divorced from the authentic three-dimensional spatial geometric context of infrastructure. Consequently, the quantification of actual spatial dimensions, deformation extents, and positional characteristics of defects is precluded, and the diagnostic outputs lack physical authenticity and engineering reference value [16, 17]. Furthermore, general-purpose multi-modal models are not imbued with infrastructure industry codes or expert maintenance knowledge, resulting in semantic outputs that are predominantly vague natural language descriptions. Condition rating, code compliance verification, and specialized maintenance strategy derivation are consequently unattainable, preventing the translation of visual perception outcomes into actionable decision-making bases [35]. In the proposed approach, three-dimensional twin geometric context is deeply integrated with vision-language models, and domain-specific professional knowledge is dynamically injected through retrieval-augmented generation techniques. A complete diagnostic chain integrating perception, understanding, and decision-making is thereby constructed, overcoming the deficiencies of existing intelligent diagnosis technologies in terms of engineering semantic deficiency and insufficient practical deployability [36, 37].

3. METHODOLOGY

Fragmentation of technical chains is a pervasive issue in existing infrastructure digital twin operation and maintenance systems, wherein three-dimensional modeling, spatiotemporal registration, model updating, and defect diagnosis modules operate independently, and a unified and coherent automated processing workflow has yet to be established. Simultaneous satisfaction of weakly textured scene modeling accuracy, long-temporal-sequence environmental robustness, and engineering-grade semantic decision-making capabilities is unattainable through any single technical approach. Conventional solutions either rely on dense-view data and prohibitive computational costs for modeling and updating, or remain confined to the shallow perceptual level of two-dimensional visual recognition, rendering them inadequate for the long-cycle, high-precision, and unmanned refined maintenance requirements of public–private partnership infrastructure. To address the aforementioned deficiencies in the technical framework, an end-to-end four-layer progressive intelligent management architecture is constructed in this study. Driven by visual data, the construction, updating, and intelligent analysis of digital twin models are realized, thereby forming an integrated operation and maintenance technical system adapted to complex engineering scenarios.

3.1 Frequency-domain enhancement and geometry-constrained neural radiance field three-dimensional reconstruction for sparse-view weakly textured scenes

Given the significantly homogeneous textures and highly

regular overall structures characteristic of public–private partnership infrastructure, combined with the engineering constraint that unmanned aerial vehicle inspection can only acquire sparse-view images, conventional neural radiance field algorithms are prone to insufficient feature representation, view overfitting, and geometric structural distortions, thereby precluding high-precision digital modeling. To overcome the aforementioned limitations, a neural radiance field reconstruction framework synergistically driven by frequency-domain optimization and geometric priors is established in this section, through which systematic algorithmic improvements are implemented across three dimensions: low-level image feature enhancement, intelligent spatial ray sampling, and model training regularization constraints. In this approach, the strong dependence of conventional spatial-domain feature modeling on rich texture information is circumvented, whereby subtle structural features of weakly textured components are extracted through frequency-domain signal analysis. Simultaneously, regularization constraints are introduced in accordance with the inherent planar and linear geometric properties of infrastructure, fundamentally improving reconstruction accuracy and structural integrity for large-scale infrastructure scenes under sparse-view conditions. Figure 1 illustrates the frequency-domain enhanced and geometry-constrained neural radiance field three-dimensional reconstruction architecture.

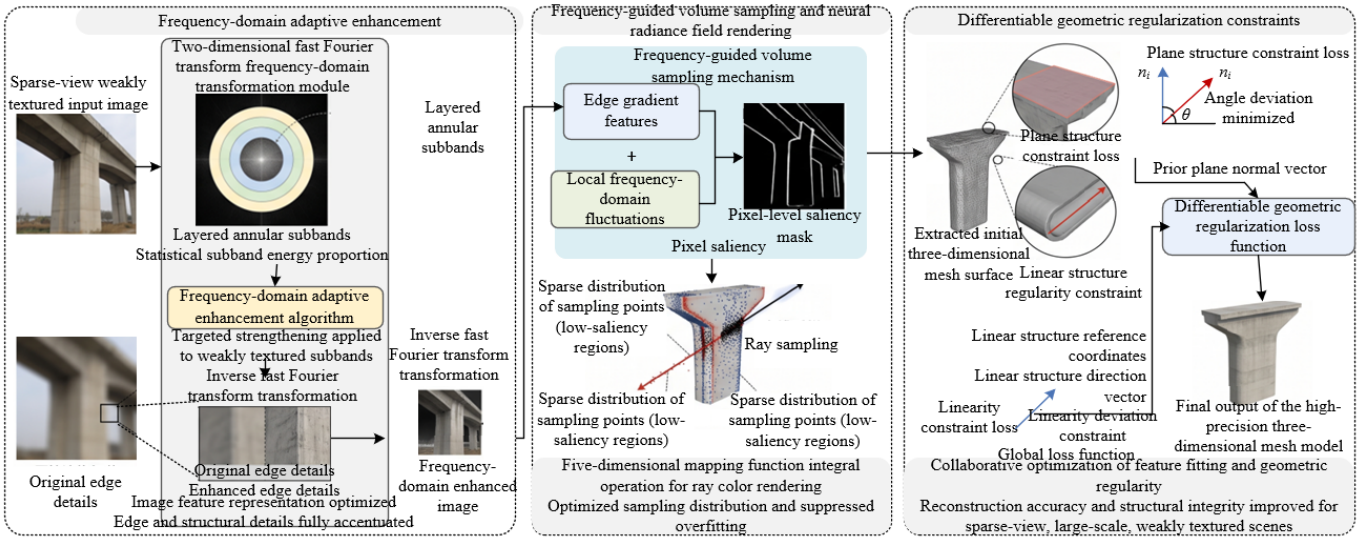


Figure 1. Architecture of frequency-domain enhanced and geometry-constrained neural radiance field three-dimensional reconstruction

A multi-scale frequency-domain adaptive enhancement strategy is first employed, through which global feature optimization is performed on the original inspection images, thereby effectively addressing the issues of sparse surface features and missing structural details in concrete and steel structures. The input spatial-domain image is assumed to have dimensions of H rows and W columns, with the corresponding image matrix denoted as I. Through two-dimensional discrete Fourier transform, the spatial-domain signal is converted into the frequency-domain signal, enabling global decoupling of structural features. The specific computational form is given as:

$$F(u,v)=\sum_{x=0}^{H-1}\sum_{y=0}^{W-1}I(x,y)e^{-j2\pi(\frac{ux}{H}+vy/W)} \tag{1}$$

where, u and v denote the horizontal and vertical coordinates in the frequency domain, respectively, x and y denote the pixel coordinates in the spatial domain, and $F(u,v)$ represents the corresponding complex frequency-domain eigenvalue of the image. To achieve layered differential enhancement, the frequency-domain space is partitioned into multiple layered annular subbands, and the energy proportion of each subband is statistically computed to quantify the richness of regional features. The subband energy proportion is calculated as:

$$E_s=\frac{\sum_{(u,v)\in\Omega_s}|F(u,v)|^2}{\sum_{(u,v)}|F(u,v)|^2} \tag{2}$$

where, E_s denotes the energy weight of the s-th frequency-domain subband, and Ω_s denotes the corresponding annular frequency-domain interval. Based on the energy distribution characteristics, an adaptive gain coefficient is constructed to achieve targeted enhancement of weakly textured regions:

$$\lambda_s=\lambda_{base}\cdot exp(-E_s/\gamma) \tag{3}$$

where, λ_{base} denotes the baseline gain magnitude, and γ denotes the gain attenuation adjustment coefficient. Higher enhancement weights are assigned to weakly textured regions with low energy proportions. After feature correction is performed through an adaptive frequency-domain filtering kernel, the optimized image is reconstructed via inverse Fourier transform, through which the concealed edge and structural details of infrastructure components are fully accentuated while the original illumination consistency is maintained.

On the basis of image feature enhancement, a frequency-guided volume sampling mechanism is designed, through which the spatial ray sampling distribution of the neural radiance field is dynamically optimized, addressing the issues of insufficient sampling of critical structures and redundant sampling in non-informative regions under sparse-view conditions. A pixel-level saliency evaluation metric is constructed by integrating edge gradient responses and local frequency-domain fluctuation characteristics, enabling precise localization of high-value structural regions. The computation

formula is given as:

$$Q(x,y)=\|\nabla^2 I'(x,y)\|+\eta\cdot Var\{F_{window}(x,y)\} \tag{4}$$

where, I' denotes the frequency-domain enhanced image, ∇^2 denotes the Laplacian operator, which is used to extract second-order edge features of the image, $Var\{\ \}$ denotes the local window frequency-domain variance, and η denotes the dual-feature weight balancing coefficient. The pixel-level saliency value directly determines the ray sampling probability, whereby the model sampling resources are automatically concentrated on critical regions such as component edges and structural joints. The neural radiance field characterizes the three-dimensional scene through a five-dimensional mapping function, and ray color rendering is accomplished via integral operations. The core rendering formula is given as:

$$C(r) = \int_{t_n}^{tf} T(t)\sigma\big(r(t)\big)c(r(t),d)dt, T(t) = \exp\Big(-\int_{t_n}^{t}\sigma\big(r(s)\big)ds\Big) \tag{5}$$

where, $r(t)$ denotes the spatial sampling ray, t_n and t_f denote the near and far bounds of the ray sampling interval, respectively, σ denotes the spatial volume density, c denotes the corresponding color features of the ray, and $T(t)$ denotes the ray transmittance, which is used to characterize the occlusion attenuation effect of the spatial medium. Through this sampling strategy, the overfitting problem induced by sparse inputs is effectively suppressed, and the capability of the model to infer geometric structures in unseen regions is enhanced.

To constrain the geometric plausibility of implicit modeling in accordance with the regular structural morphology characteristic of public–private partnership infrastructure, a differentiable geometric regularization loss function is introduced, through which precise constraints are imposed on the model training process. The normal vector n_{est} of the fitted surface is solved via spatial density gradient computation, and the scene prior normal vector n_{prior} is obtained through local plane fitting of sparse point clouds. A plane structure constraint loss is constructed to regulate the modeling accuracy of planar components such as bridge decks and utility tunnel walls:

$$L_{plane}=\frac{1}{|S|}\sum_{x\in S} arccos^2\left(n_{est}(x)\cdot n_{prior}(x)\right) \tag{6}$$

where, S denotes the set of spatially uniformly sampled points. Through this loss function, the angular deviation between the predicted normal vectors and the prior normal vectors is quantified, whereby convex and concave distortions of planar structures are suppressed. For linear structural components such as main girders and guardrails, a linearity constraint loss is further constructed to ensure the regularity of linear structures:

$$L_{line}=\frac{1}{|L|}\sum_{x\in L}\|(x\text{-}x_0)\text{-}vv^{\top}(x\text{-}x_0)\|^2 \tag{7}$$

where, L denotes the set of sampling points on linear structures, x_0 denotes the structural reference coordinate point, and v denotes the direction vector of the linear structure. Multiple constraints are integrated to construct a global loss function, through which the collaborative optimization of feature fitting and geometric regularity is achieved. The total loss expression is given as:

$$L_{total}=L_{rgb}+\gamma_{plane}L_{plane}+\gamma_{line}L_{line}+\gamma_{sparse}\|\sigma\|_1 \tag{8}$$

where, L_{rgb} denotes the image color reconstruction loss, and γ_{plane}, γ_{line}, and γ_{sparse} denote the weight coefficients of the respective constraints. A coarse-to-fine progressive iterative training strategy is adopted. After training convergence, a high-precision triangular mesh model is extracted using the marching cubes algorithm, through which the refined reconstruction of the initial digital twin scene for public–private partnership infrastructure is ultimately accomplished.

3.2 Hierarchical scene coordinate regression network with frequency-domain and spatial-domain dual-path fusion

In response to the temporal environmental variation challenges inherent to long-term public–private partnership infrastructure maintenance scenarios, scene coordinate regression methods driven solely by single-domain spatial features are highly susceptible to illumination fluctuations, seasonal changes, and shadow variations. Feature degradation and pixel-level registration drift are readily induced, making it difficult to stably achieve high-precision mapping between two-dimensional inspection images and three-dimensional digital twin models. To address this problem, a hierarchical scene coordinate regression network with frequency-domain and spatial-domain dual-path fusion is constructed in this section, through which an end-to-end pixel-level three-dimensional coordinate prediction mechanism with strong environmental robustness is established. Figure 2 illustrates the scene coordinate regression network with frequency-domain and spatial-domain dual-path fusion. The network takes temporally new three-channel inspection images as input, and spatial prior constraints are provided by the established high-precision three-dimensional point cloud twin model. Through dual-dimensional feature encoding, adaptive feature fusion, and residual refinement regression, stable three-dimensional spatial coordinates are output pixel by pixel, providing a precise spatiotemporal alignment basis for subsequent temporal change detection and incremental model updating.

Two mutually independent and functionally complementary feature encoding branches, namely the spatial-domain and frequency-domain branches, are constructed, through which effective scene information is mined from different signal dimensions. The spatial-domain branch is built upon an improved High-Resolution Network-W48 (HRNet-W48) backbone to establish a multi-scale feature extraction structure, with four parallel resolution branches configured to perform cross-scale feature interaction and fusion, through which component contours, edge textures, and local geometric details are completely preserved, and high-precision spatial-domain feature representations are output. The frequency-domain branch focuses on eliminating feature interference caused by illumination style variations. Patch-based Fourier transform is first applied to the input images, with a patch size of 16 × 16 pixels and a sliding step of 8 pixels, whereby both the completeness of local frequency-domain information and the continuity of global features are ensured. To suppress feature distortion induced by illumination intensity shifts, standardization correction is performed on the frequency-

domain magnitude spectrum. The calculation formula is given as:

$$A_{norm}=\frac{A-\mu_A}{\sigma_A+\epsilon} \tag{9}$$

where, A denotes the original frequency-domain magnitude spectrum, μ_A and σ_A denote the global mean and standard deviation of the magnitude spectrum, respectively, and ϵ denotes a small constant introduced to avoid division-by-zero issues. The correction process preserves the phase spectrum, which possesses structural invariance, while only the magnitude information, which is susceptible to illumination interference, is normalized. After inverse Fourier transform, the illumination-robust frequency-domain enhanced image is reconstructed, from which multi-scale frequency-domain features are ultimately extracted through a four-layer convolutional downsampling structure.

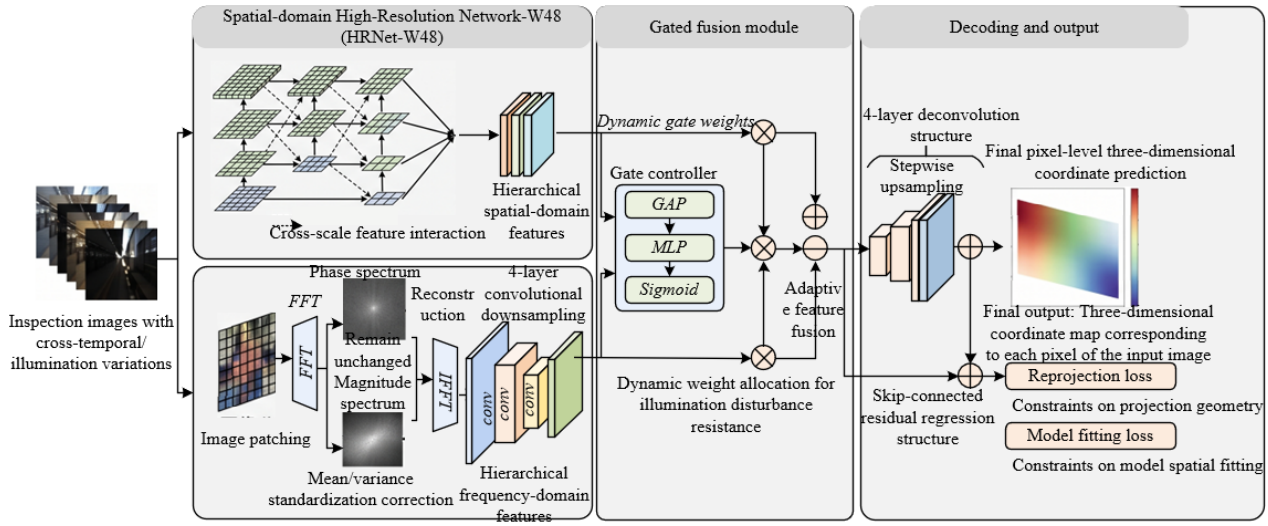


Figure 2. Scene coordinate regression network with frequency-domain and spatial-domain dual-path fusion
Note: FFT: Fast Fourier transform; IFFT: Inverse fast Fourier transform; GAP: Global average pooling; MLP: Multilayer perceptron.

To fully integrate the high-precision advantages of spatial-domain features with the robustness advantages of frequency-domain features, a gated adaptive fusion mechanism is introduced at the four-level multi-scale features of the encoder, through which dynamic adaptive allocation of feature weights is achieved. The hierarchical feature fusion formula is given as:

$$F^{(l)}_{fusion}=g^{(l)}\odot F^{(l)}_{spatial}+(1-g^{(l)})\odot F^{(l)}_{frequency} \tag{10}$$

where, l denotes the feature scale level index, $F^{(l)}{}_{spatial}$ and $F^{(l)}{}_{frequency}$ denote the spatial-domain and frequency-domain features at the corresponding level, respectively, $g^{(l)}$ denotes the adaptive gating weight matrix, and $\odot$ denotes the element-wise multiplication operation of features. The gating weights are solved by being driven by global contextual information. Global features from both branches are aggregated through global average pooling, and weight mapping is then accomplished via a multilayer perceptron and a Sigmoid activation function:

$$g^{(l)}=Sigmoid\left(MLP^{(l)}\left(\begin{bmatrix}GAP\left(F^{(l)}_{spatial}\right);\\ GAP\left(F^{(l)}_{frequency}\right)\end{bmatrix}\right)\right) \tag{11}$$

where, GAP denotes the global average pooling operation, which is used to compress global feature dimensions and capture the overall properties of the scene. Through this mechanism, the feature weights of the two branches are dynamically balanced according to the real-time scene state, whereby high-precision spatial-domain features are preferentially adopted in regions with stable illumination and rich textures, while frequency-domain feature weights are adaptively strengthened in regions with drastic illumination changes and weakened textures. The network decoder employs a four-layer deconvolution structure for feature upsampling, and a residual regression structure is introduced to optimize output accuracy. The final pixel-level three-dimensional coordinates are obtained by summing coarse-grained base coordinates and refined residual coordinates, through which fitting deviations in the deep network are effectively compensated.

To regulate the network training process and ensure the geometric validity and spatial fitting accuracy of the two-dimensional-to-three-dimensional mapping, a dual-consistency joint loss function is constructed. The projection consistency loss is used to constrain the predicted coordinates to satisfy the camera imaging geometry and suppress reprojection deviations. The calculation formula is given as:

$$L_{proj}=\sum_{(u,v)}\left\|K[R|t]\widehat{X}(u,v)-[u,v,1]^T\right\|^2 \tag{12}$$

where, K denotes the camera intrinsic matrix, R and t denote the camera rotation matrix and translation vector, respectively, and $\widehat{X}(u,v)$ denotes the predicted three-dimensional coordinate corresponding to pixel (u,v). The model consistency loss is introduced with a Huber robust kernel function, through which the spatial distance between predicted coordinates and the reference twin point cloud is constrained, and the interference of outlier noise points is reduced:

$$L_{model}=\sum_{(u,v)}\rho\left(dist\left(\widehat{X}(u,v),P\right)\right) \tag{13}$$

where, *dist*( ) denotes the three-dimensional Euclidean distance computation, ρ( ) denotes the Huber robust loss function, and P denotes the reference digital twin point cloud set. The total network loss is composed of the two consistency losses jointly, expressed as:

$$L_{regress}=L_{proj}+L_{model} \tag{14}$$

Through this dual-constraint formulation, network convergence is constrained from both the perspective of imaging projection plausibility and three-dimensional model fitting accuracy, whereby the stability and engineering precision of cross-temporal and cross-scene pixel-level localization are significantly enhanced.

3.3 Adaptive threshold change detection and topology-preserving incremental updating

Based on the preceding pixel-level three-dimensional coordinate mapping results, a spatial deviation characterization between temporal images and the reference digital twin model can be established, providing a quantitative basis for structural deformation and local damage detection. Based on the temporal differences of registered coordinates, a three-dimensional spatial deviation map is constructed pixel by pixel. The calculation is given as:

$$\Delta(u,v)=\left\|\widehat{X}(u,v)-X_{ref}(u,v)\right\| \tag{15}$$

where, $\widehat{X}(u,v)$ denotes the predicted three-dimensional coordinate of the pixel in the current temporal image, $X_{ref}(u,v)$ denotes the corresponding reference three-dimensional coordinate from the reference model, and $\Delta(u,v)$ characterizes the spatial geometric offset at the pixel location. Traditional fixed-threshold discrimination methods are unable to adapt to the texture differences and structural deformation characteristics across different regions of infrastructure, readily leading to missed detections in weakly deformed regions and false detections in noisy regions. To address this issue, a local adaptive threshold discrimination mechanism is introduced, through which the discrimination criterion is dynamically updated based on the local statistical characteristics of the deviation map. Figure 3 illustrates the workflow of adaptive change detection and local topology-preserving updating. The adaptive threshold calculation formula is given as:

$$\delta(u,v)=\mu_{\Delta}(u,v)+\kappa\cdot\sigma_{\Delta}(u,v) \tag{16}$$

where, $\mu_{\Delta}(u,v)$ and $\sigma_{\Delta}(u,v)$ denote the mean and standard deviation of the deviation values within the local sliding window, respectively, which are used to characterize the deviation distribution properties of the local region, and κ denotes the confidence adjustment coefficient, which is fixed at a value of 2.5. Based on the dynamic threshold, the pixel-level structural change probability is computed as:

$$p_{change}(u,v)=Sigmoid\left(\frac{\Delta(u,v)-\delta(u,v)}{\tau}\right) \tag{17}$$

where, τ denotes the probability scaling coefficient, which is used to smooth the gradient of the probability output. To eliminate isolated noise points and ensure the spatial connectivity of change regions, a fully connected conditional random field is employed to perform global spatial smoothing on the probability map. Finally, through a binarization operation, a precise structural change mask is generated, through which the precise localization of temporal structural change regions is accomplished.

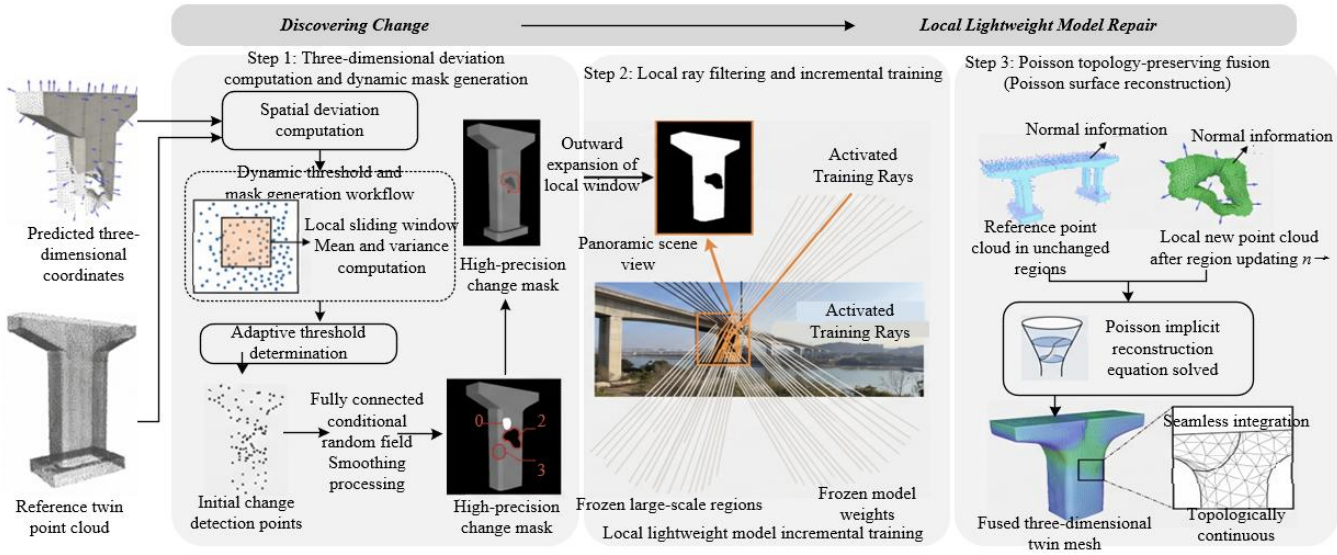


Figure 3. Workflow of adaptive change detection and local topology-preserving updating

On the basis of precise localization of structural change regions, a local incremental updating strategy is adopted to replace the traditional global reconstruction mode, whereby computational costs are substantially reduced while model updating accuracy is ensured. The algorithm takes the detected change mask region as the core, and a local reconstruction window is constructed by outward expansion of a fixed pixel range. Only the imaging rays corresponding to the window are selected to participate in model optimization training, effectively eliminating redundant computations in non-informative regions. During the training process, the base network weights of the initial neural radiance field model are frozen, and gradient updates are performed only for the network parameters corresponding to the local change regions,

whereby the original modeling accuracy of unchanged regions is preserved to the greatest extent. To constrain the fitting quality and structural smoothness of the local update, a local joint loss function is constructed:

$$L_{local}=\sum_{r\in R_{window}}\|C(r)-C_{new}(r)\|^2+\lambda_{smooth}\left\|\nabla M_{update}\right\|^2 \quad (18)$$

where, R_{window} denotes the set of rays within the local reconstruction window, $C(r)$ and $C_{new}(r)$ denote the rendered color from the original model and the actual color from the new image, respectively. The first loss term is used to constrain the color fitting accuracy of the local region, and λ_{smooth} denotes the smoothness regularization weight coefficient. The second gradient regularization term is used to constrain the spatial continuity of the updated mesh, thereby suppressing structural artifacts and geometric distortions induced by local updating.

To resolve the issues of topological discontinuity and seam distortions at the boundary between the new and old models after local updating, the Poisson surface reconstruction algorithm is introduced to achieve topology-preserving fusion of the models. Oriented point cloud data are extracted from both the locally updated mesh and the effective regions of the original reference model, and global stitching and fusion of the two point cloud sets are performed, whereby structural discontinuities and normal vector abrupt changes caused by direct stitching are avoided. By solving the Poisson implicit reconstruction equation, a globally continuous surface structure is uniformly fitted using the spatial positions and normal information of the point clouds, achieving seamless integration between the updated region and the original model region, and ensuring that the fused overall model satisfies first-order geometric continuity. Through this updating mechanism, the high-precision modeling results of non-degraded regions are completely preserved, while the model parameters of structurally deformed and damaged regions are precisely corrected. Low-cost, high-precision, and high-topological-consistency long-term dynamic iteration of the digital twin model is thereby realized, meeting the long-cycle, high-frequency, and automated maintenance updating requirements of public–private partnership infrastructure scenarios.

3.4 Multi-modal prompting and retrieval-augmented generation-enhanced defect semantic diagnosis

Instance segmentation of infrastructure surface defects is accomplished using a pre-trained DINOv2 visual encoder, through which defect instance masks, category attributes, and prediction confidences are obtained at the image level. Conventional visual detection methods can only output two-dimensional planar defect regions, without the capability to characterize the scale features of defects in physical space, rendering them insufficient for supporting structural safety assessment. Based on the pixel-level three-dimensional coordinate mapping relationship established in the preceding sections, the two-dimensional defect masks can be precisely mapped into the dynamically updated digital twin three-dimensional space, through which precise quantification of defect geometric features is achieved. Figure 4 illustrates the retrieval-augmented generation-enhanced multi-modal defect semantic diagnosis framework. By computing the three-dimensional coordinate deviations corresponding to the pixels in the defect region, the maximum extension dimension of the defect can be solved. The specific expression is given as:

$$L_k=\max_{(u,v)\in R_k}\left\|\hat{X}(u,v)-X_{ref}(u,v)\right\| \quad (19)$$

where, R_k denotes the pixel set corresponding to the k-th defect instance, $\hat{X}$ denotes the three-dimensional spatial coordinates mapped from the current image pixels, and X_{ref} denotes the reference coordinates from the reference twin model. Based on the topological relationships of the matched three-dimensional point sets, the effective surface area and convex hull bounding volume of the defect region are further solved, through which a refined three-dimensional defect feature system is comprehensively constructed, encompassing spatial location, deformation scale, and damage extent, thereby providing quantifiable physical prior support for subsequent engineering semantic diagnosis.

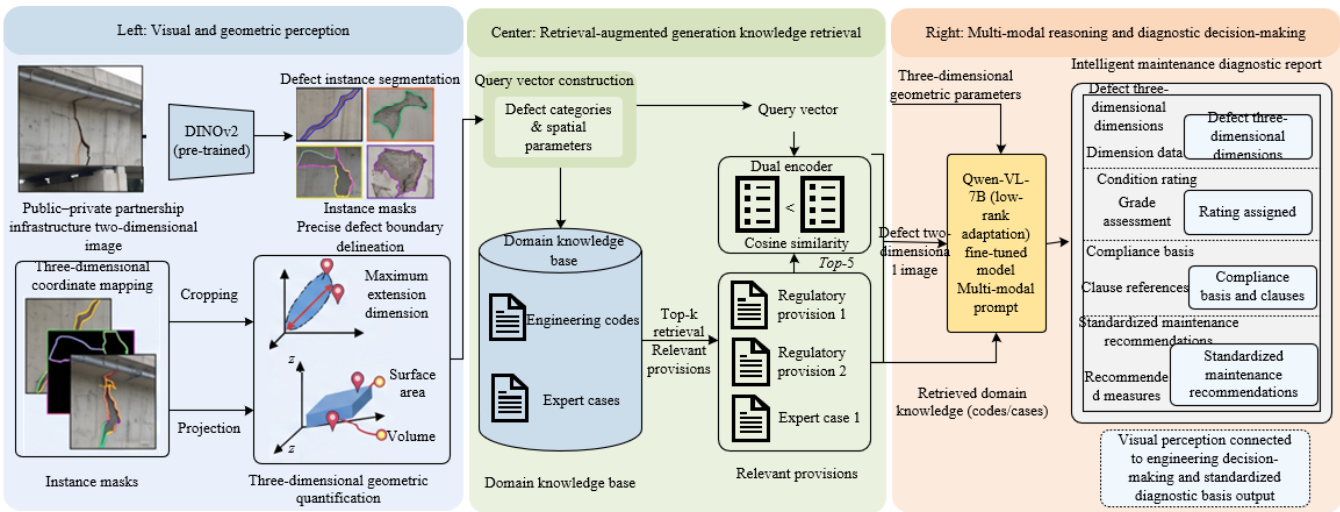


Figure 4. Retrieval-augmented generation-enhanced multi-modal defect semantic diagnosis framework

To achieve deep coupling between visual perception results and engineering domain knowledge, a hierarchical multi-modal prompting mechanism is constructed, through which a standardized input paradigm is established by integrating facility scene attributes, component types, and three-dimensional defect parameters, guiding the multi-modal

model to perform specialized structural diagnosis. General-purpose vision-language models rely on parametric knowledge solidified during the pre-training phase, suffering from industry code memorization biases and insufficient adaptability to engineering scenarios, rendering them unable to produce decision content that conforms to infrastructure maintenance standards. A retrieval-augmented generation mechanism is introduced to establish a dynamic knowledge invocation pipeline. Through a dual-encoder architecture, unified vector encoding is performed on infrastructure maintenance code provisions and engineering expert cases, and a structured domain knowledge vector database is constructed. A retrieval query vector is constructed by fusing defect categories, spatial locations, and three-dimensional geometric parameters, and the matching relationship between the sample and knowledge base entries is quantified through cosine similarity. The calculation formula is given as:

$$Sim(q_k,e_j)=\frac{q_k \cdot e_j}{\|q_k\| \cdot \|e_j\|} \tag{20}$$

where, q_k denotes the defect feature query vector, and e_j denotes the encoded vector of the j-th entry in the knowledge base. The top five domain knowledge entries with the highest similarity scores are selected and embedded into the prompt context, through which real-time injection of industry standards and expert experience is achieved. The limitations of generic model parameter memorization are thereby overcome, and the compliance and professionalism of defect diagnostic outputs are significantly enhanced.

A low-rank adaptation fine-tuning strategy is employed to achieve domain-specific optimization of the vision-language model. Using Qwen-VL-7B as the base model, lightweight fine-tuning is performed with a rank parameter set to 16, through which precise adaptation to the exclusive task scenario of public–private partnership infrastructure defect diagnosis is accomplished while the general cross-modal understanding capabilities of the model are preserved. Through multi-modal prompt inputs that integrate three-dimensional geometric information and retrieved knowledge, joint reasoning of visual features, spatial parameters, and domain knowledge is performed by the model, and structured diagnostic text outputs are generated. To eliminate stochastic errors in model-generated content, a post-processing verification mechanism is designed, through which automated verification and correction are performed on the physical dimension units, validity of code provisions, and condition rating criteria in the output content, ensuring the rigor and engineering usability of diagnostic results. Through this technical pipeline, the barrier between three-dimensional defect perception and intelligent engineering decision-making is effectively bridged. The conventional shallow visual recognition results are upgraded to standardized diagnostic reports that encompass three-dimensional defect parameters, damage grades, code compliance bases, and targeted maintenance recommendations, achieving an integrated closed-loop analysis of infrastructure defect perception, semantic reasoning, and intelligent decision-making.

4. EXPERIMENTAL DESIGN AND RESULT ANALYSIS

In this chapter, the effectiveness and superiority of the proposed three-dimensional reconstruction, spatiotemporal registration, model incremental updating, and intelligent semantic diagnosis modules were systematically validated through multiple sets of comparative experiments, parametric analyses, and ablation studies. Based on a self-constructed large-scale public–private partnership infrastructure dataset and publicly available general scene datasets, quantitative evaluations were conducted across multiple dimensions, including modeling accuracy, temporal robustness, computational efficiency, and diagnostic accuracy. Furthermore, the synergistic enhancement mechanisms among the technical components were verified through module-wise ablation experiments, through which the engineering applicability and algorithmic superiority of the complete methodology were comprehensively demonstrated.

4.1 Datasets and evaluation metrics

Three dedicated datasets were constructed in this study, serving model training, performance testing, and generalization validation, respectively, with comprehensive coverage of real engineering scenes, controllable simulated scenes, and temporal evolution scenes. The PPP-Real dataset was collected from five in-service bridges and three urban comprehensive utility tunnel facilities, encompassing cross-seasonal environments (spring and summer) and multi-temporal illumination conditions (morning, noon, and evening), with a total of 6,842 sets of real inspection images. High-precision ground-truth light detection and ranging point clouds were synchronously matched as reference data, through which the characteristics of texture deficiency, illumination fluctuations, and structural variability in real maintenance scenarios were effectively represented. The PPP-Syn dataset was generated through simulation rendering of images, depth maps, and camera poses based on infrastructure building information modeling, with a total of 5,000 fully annotated samples, providing noise-free, high-precision prior supervision information for network training. The PPP-Change dataset was constructed with multi-temporal image sequences, simulating structural deformation and defect evolution processes at zero, three, six, and twelve months, through which the long-term stability of cross-temporal updating and change detection is validated.

To enable standardized quantitative evaluation of multi-task performance, a multi-dimensional evaluation metric system was adopted. For the three-dimensional reconstruction task, chamfer distance, F1-score at a 10 cm threshold, and peak signal-to-noise ratio were selected as core evaluation metrics, through which geometric fitting accuracy, structural completeness, and texture rendering quality were characterized, respectively. For the spatiotemporal localization task, mean absolute error and localization recall at a 5 cm threshold were employed to quantify pixel-level three-dimensional registration accuracy and effective localization coverage. For the scene change detection task, the F1-score and intersection over union metrics were adopted to comprehensively measure the precision and completeness of change region identification. For the intelligent diagnosis task, the Recall-Oriented Understudy for Gisting Evaluation–Longest Common Subsequence (ROUGE-L) and Bilingual Evaluation Understudy (BLEU-4) metrics from the text generation domain were used to evaluate diagnostic statement fluency and content matching, while semantic matching accuracy and industry code provision citation accuracy were

additionally introduced to quantify the engineering professionalism and compliance of diagnostic outcomes.

4.2 Sparse-view three-dimensional reconstruction accuracy evaluation

To validate the reconstruction capability of the proposed method in sparse-view weakly textured infrastructure scenarios, traditional geometric reconstruction algorithms and mainstream neural radiance field variants were selected for multi-view quantitative comparative experiments. Tests were conducted under two-view, three-view, and five-view sparse input conditions, respectively. The overall reconstruction performance results are presented in Table 1.

Table 1. Comparison of three-dimensional reconstruction quality under different sparse-view conditions

Method	Two-view			Three-view			Five-view		
	Chamfer Distance↓	F1↑	Peak Signal-to-Noise Ratio↑	Chamfer distance↓	F1↑	Peak Signal-to-Noise Ratio↑	Chamfer Distance↓	F1↑	Peak Signal-to-Noise Ratio↑
COLMAP	0.152	0.341	—	0.098	0.512	—	0.054	0.723	—
Neural radiance field (original)	0.186	0.287	18.2	0.112	0.468	21.5	0.061	0.698	24.3
Voxel-based neural radiance field	0.134	0.402	20.1	0.085	0.567	22.9	0.048	0.751	25.7
Sparse geometry-constrained neural radiance field	0.121	0.439	21.3	0.072	0.612	23.7	0.042	0.782	26.1
Proposed method	0.073	0.568	24.7	0.041	0.724	27.2	0.025	0.846	29.8

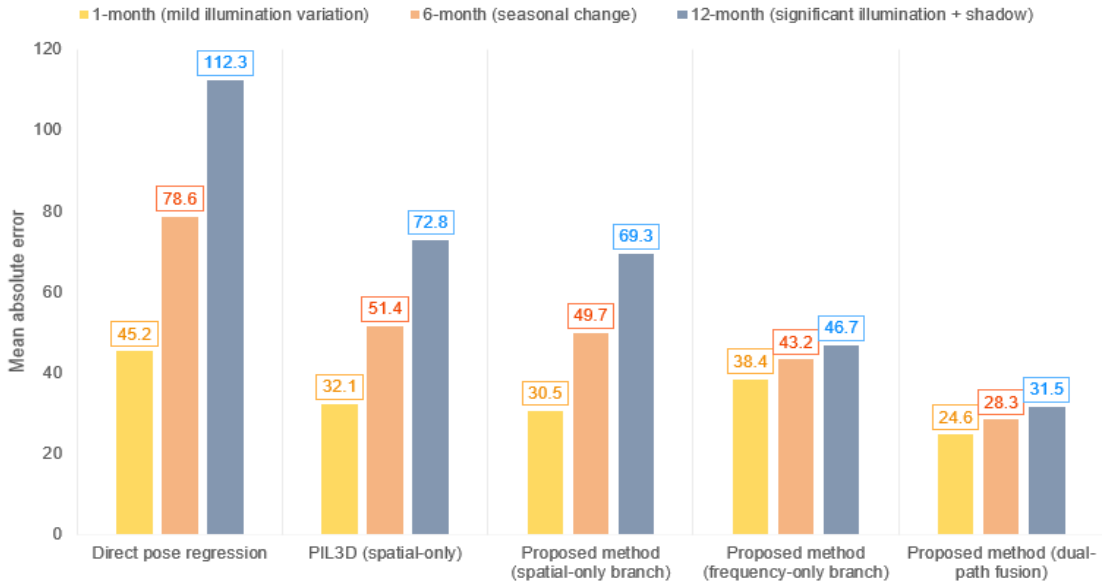


Figure 5. Comparison of localization mean absolute error under different cross-temporal conditions (mm)

Table 2. Comparison of localization recall (%) at 5 cm threshold under different scenarios

Method	Bridge (Strong-Texture Area)	Bridge (Weak-Texture Area)	Utility Tunnel (Uniform Illumination)	Utility Tunnel (Low-Light)
PIL3D	78.2	52.3	81.5	44.6
Proposed method	89.5	76.8	92.1	71.3

From the experimental results, it can be observed that the reconstruction accuracy of all algorithms exhibits a steady improvement trend with an increasing number of input views. However, the proposed method achieves optimal performance under all sparse input conditions. Under the extreme two-view input condition, compared to the best-performing baseline sparse geometry-constrained neural radiance field, the proposed method reduces chamfer distance by 39.7% and improves the structural F1-score by 29.4%, fully demonstrating that the frequency-domain feature enhancement mechanism can effectively compensate for feature deficiency in weakly textured scenes and alleviate geometric reasoning distortions induced by sparse views. The traditional COLMAP algorithm suffers from severe feature matching failures under few-view conditions, with a structural F1-score of only 0.341 under two-view input, rendering complete scene reconstruction unattainable. The original neural radiance field algorithm exhibits severe overfitting issues, with both geometric accuracy and texture rendering quality remaining at relatively low levels. Under the five-view input condition, which is commonly employed in engineering practice, the proposed method reduces chamfer distance to

0.025, meeting high-precision engineering modeling standards. The peak signal-to-noise ratio is improved by 5.5 dB compared to the original neural radiance field algorithm, demonstrating that the frequency-guided sampling strategy enables efficient allocation of sampling resources, focusing on key structural regions to achieve refined modeling.

4.3 Cross-temporal pixel-level localization robustness

To validate the registration stability of the dual-path fused feature network under temporal environmental variations, cross-temporal localization experiments were conducted with different time spans, covering three typical maintenance scenarios: mild illumination fluctuations, seasonal changes, and strong illumination with shadow variations. The average localization errors of each method are shown in Figure 5. In addition, fine-grained scenario tests were performed on bridge scenes with strong versus weak textures, and utility tunnel scenes under normal illumination and low-light conditions. The localization recall performance is presented in Table 2.

The experimental results demonstrate that the single-domain spatial feature methods are highly sensitive to temporal environmental variations. The localization errors of both PIL3D and the spatial-only branch model of the proposed method increased by over 127% over the twelve-month temporal span, indicating significantly insufficient environmental adaptability. The frequency-only branch model effectively resists illumination and seasonal disturbances, with an error increase of only 21.6%, exhibiting excellent environmental robustness. The proposed gated dual-path fusion mechanism dynamically balances the weights of the two feature types according to the real-time scene state. In short-term temporal scenarios with stable illumination, ultra-high localization accuracy of 24.6 mm is achieved by relying on spatial-domain features, while in long-term temporal scenarios with drastic illumination changes, the frequency-domain feature weights are adaptively strengthened, and the localization error is stably controlled within 31.5 mm. The fine-grained scenario experimental results further validate the scene adaptability of the proposed method. Compared with the PIL3D algorithm, the localization recall of the proposed method is improved by 46.8% and 59.9% in bridge weak-texture regions and utility tunnel low-light scenarios, respectively, effectively addressing the registration failure problem caused by degradation of traditional spatial-domain features.

4.4 Change detection and incremental updating performance

To determine the optimal hyperparameters of the adaptive threshold and validate the engineering value of the incremental updating mechanism, parameter sensitivity experiments and computational efficiency comparison experiments were conducted. The change detection performance under different threshold coefficients is shown in Figure 6, and the efficiency and accuracy comparison between incremental updating and global reconstruction is presented in Table 3.

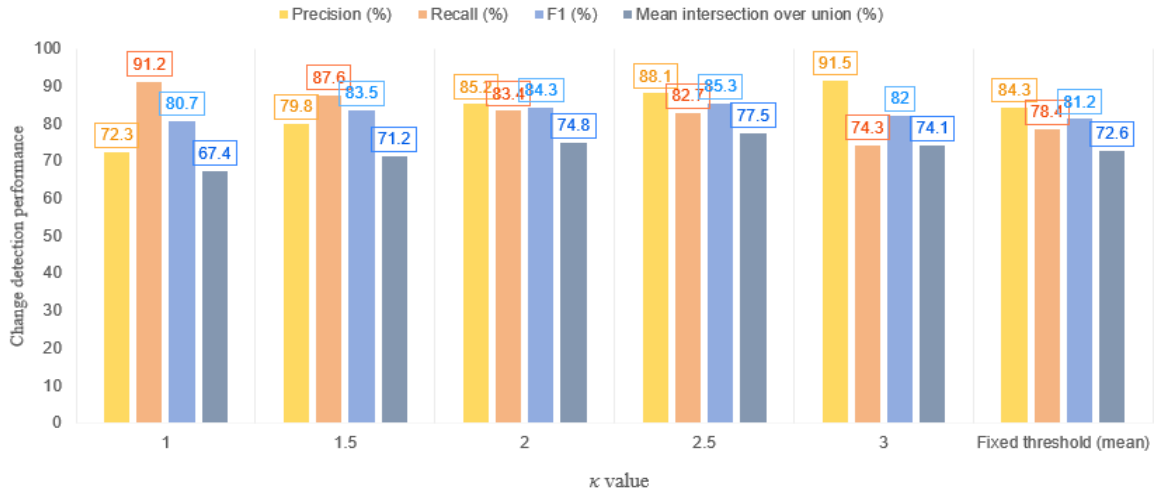


Figure 6. Influence of adaptive threshold parameter κ on change detection performance

Table 3. Computational efficiency comparison between incremental updating and full model reconstruction

Update type	Changed Region Ratio	Training Iterations	Time per Session (min)	Graphics Processing Unit Memory (GB)	Quality Retention (ΔPSNR)
Full model reconstruction	100%	200k	185	22.4	—
Incremental update (5% region)	5%	10k	12	8.6	-0.3 dB
Incremental update (15% region)	15%	30k	35	10.2	-0.5 dB
Incremental update (30% region)	30%	60k	68	13.7	-0.8 dB

From the parameter experimental results, it can be observed that the value of the threshold coefficient directly balances the precision and recall of change detection. A smaller coefficient introduces a substantial number of noise-induced false detections, leading to a significant decrease in precision, while a larger coefficient weakens the change characteristics of

weakly deformed regions, resulting in extensive missed detections. When the coefficient is set to 2.5, the model achieves the optimal F1-score and Intersection over Union, which are improved by 4.1% and 4.9%, respectively, compared to the conventional fixed-threshold scheme, demonstrating that the local adaptive statistical threshold can effectively adapt to structural differences across different regions and achieve precise segmentation of change regions. The computational efficiency experimental results demonstrate that the local incremental updating mechanism can substantially reduce the iteration cost of digital twin models. When the structural change region accounts for only 5% of the overall scene, the time required for incremental updating is merely 6.5% of that for global reconstruction, with graphics processing unit memory consumption reduced by 61.6%, while model rendering quality degrades by only 0.3 dB, a difference imperceptible to the human eye. Even when the changed region ratio increases to 30%, the updating time remains less than 40% of that for global reconstruction, and the rendering quality loss is controlled within 1 dB. This approach achieves a balance between model updating accuracy and maintenance efficiency, making it well-suited for the high-frequency, long-cycle maintenance iteration requirements of public–private partnership infrastructure.

4.5 Defect semantic diagnosis quality evaluation

To validate the improvement effect of the three-dimensional geometric mapping and retrieval-augmented mechanisms on intelligent diagnosis performance, comparisons were conducted between the proposed method and traditional template matching, general-purpose vision-language models, and existing industrial anomaly diagnosis algorithms across multiple dimensions, including text generation quality, semantic matching accuracy, and code citation accuracy. The experimental results are presented in Table 4.

Table 4. Output quality comparison of different diagnosis methods

Method	Recall-Oriented Understudy for Gisting Evaluation–Longest Common Subsequence (ROUGE-L) (%)	Bilingual Evaluation Understudy (BLEU-4) (%)	Semantic Matching Accuracy (%)	Code Provision Citation Accuracy (%)
YOLOv8 + rule template	38.2	12.4	52.1	34.6
Original Qwen-VL (zero-shot)	45.7	21.3	61.8	28.9
Original Qwen-VL (low-rank adaptation fine-tuned)	52.3	28.6	70.5	45.2
AnomalyGPT	47.1	24.8	65.4	31.7
Proposed method (vision-language model+low-rank adaptation+three-dimensional mapping, without retrieval-augmented generation)	56.8	32.1	76.3	56.4
Proposed method (full + retrieval-augmented generation)	64.2	39.7	84.6	78.9

Table 5. Influence of module-wise ablation on overall system performance

Configuration	Reconstruction Chamfer Distance↓	Localization Mean Absolute Error↓ (mm)	Change F1↑ (%)	Diagnostic Semantic Matching↑ (%)	Comprehensive Score*
Full system	0.025	28.1	85.3	84.6	92.4
Without frequency enhancement (replaced with histogram equalization)	0.048	—	—	—	74.6
Without geometric constraints ($\gamma = 0$)	0.042	—	—	—	78.2
Without dual-path fusion (spatial-only)	—	49.8	76.4	—	70.5
Without adaptive threshold (fixed δ)	—	—	81.2	—	81.3
Without retrieval-augmented generation retrieval	—	—	—	76.3	79.8
Without local incremental (full update)	—	—	—	—	74.1 (extremely low efficiency)

The experimental results clearly demonstrate a progressive gain pattern in diagnostic performance. The conventional two-dimensional detection combined with fixed templates can only achieve simple defect classification, lacking spatial geometric cognition and industry knowledge support, with all metrics remaining at the lowest levels. The zero-shot inference mode of general-purpose vision-language models is affected by the absence of domain-specific knowledge, with code provision citation accuracy below 30%, rendering them insufficient for engineering diagnosis requirements. The domain-fine-tuned model achieves modest improvement in semantic adaptation capability, though the deficiency in specialized knowledge reserves persists. After the introduction of the three-dimensional geometric mapping mechanism, defect reasoning can be performed by the model based on real spatial scale parameters, with semantic matching accuracy improved to 76.3%, demonstrating that three-dimensional spatial context is the core foundation for achieving accurate condition rating. Following the incorporation of the retrieval-augmented mechanism, standardized maintenance codes and expert cases

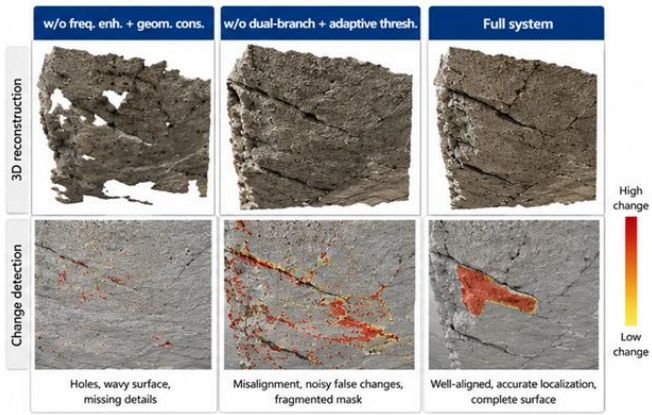
can be retrieved in real time by the model, completely overcoming the limitations of parametric memory. Ultimately, a semantic matching accuracy of 84.6% and a code citation accuracy of 78.9% are achieved, with text generation fluency and professionalism significantly superior to existing methods, enabling the output of structured diagnostic reports with engineering practical value.

4.6 End-to-end integration and ablation analysis

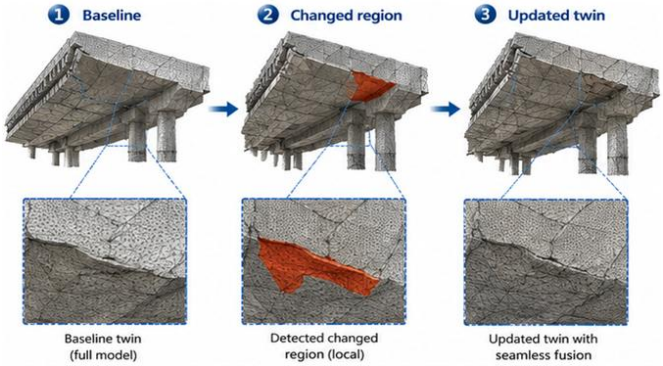
To quantify the independent contributions of each innovative module and the synergistic effects of the system, comprehensive module-wise ablation experiments were designed, in which the core technical units were sequentially removed and the corresponding changes in overall system performance were measured. A global quantitative evaluation was conducted through a normalized comprehensive score. The ablation results are presented in Table 5.

The ablation experimental results demonstrate that each core module provides an irreplaceable positive contribution to system performance, and the modules collectively form a complete synergistic technical chain. The dual-path feature fusion module exerts the most significant influence on the overall system; after its removal, the spatiotemporal localization error increases substantially, and the comprehensive score decreases by 21.9 points. This is because pixel-level spatiotemporal registration serves as the foundational prerequisite for model updating and defect diagnosis, and the degradation of registration accuracy directly leads to performance attenuation in subsequent tasks. In the three-dimensional reconstruction stage, frequency-domain enhancement and geometric regularization constraints address the issues of weak-texture feature deficiency and geometric structural distortion, respectively; the removal of either module results in substantial degradation of reconstruction accuracy. The adaptive threshold mechanism effectively ensures the precision of structural change detection, while the retrieval-augmented mechanism significantly enhances the professionalism and compliance of terminal diagnosis. Although the local incremental updating module does not directly affect quantitative accuracy metrics, it fundamentally resolves the high time consumption and computational demands of global reconstruction, serving as a critical enabler for practical engineering deployment. Through multi-module collaborative optimization, the complete system achieves optimal performance across the entire workflow of modeling, registration, updating, and diagnosis, fully validating the rationality and completeness of the proposed technical framework.

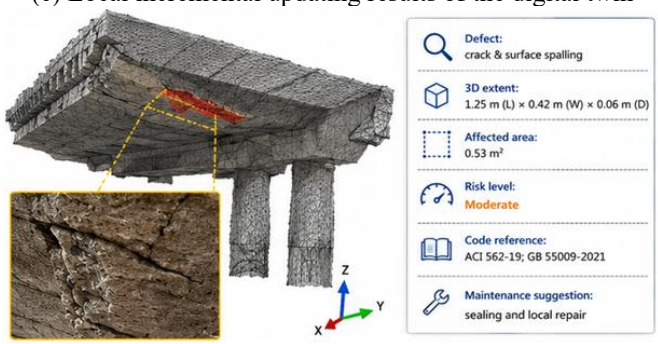
(a) Cross-temporal sparse-view inspection image inputs



(b) Comparison of key module ablation effects



(c) Local incremental updating results of the digital twin



(d) Three-dimensional defect mapping and intelligent diagnostic outputs

Figure 7. End-to-end digital twin intelligent management and ablation results of key modules

To validate the end-to-end synergistic effects of each core module in real bridge defect scenarios and to reveal the intrinsic mechanisms by which individual performance improvements are transformed into engineering diagnostic capabilities, integrated implementation and ablation visualization experiments were conducted, as shown in Figure 7. The results demonstrate that under complex conditions involving cross-temporal illumination variations, concrete weak textures, and localized spalling, the complete system can stably recover a continuous and geometrically regular three-dimensional undersurface of the bridge from sparse inspection images, while accurately constraining cross-temporal changes to the actual defect regions. After the removal of frequency-domain enhancement and geometric constraints, holes, surface undulations, and edge fractures appeared in the reconstruction results. After the removal of dual-path fusion and adaptive thresholding, significant registration drift, spurious change responses, and mask fragmentation were observed, indicating that the quality of underlying reconstruction and spatiotemporal localization accuracy directly affect the

reliability of subsequent change perception. The complete method achieves a reconstruction chamfer distance of 0.025, a localization error reduced to 28.1 mm, and a change detection F1-score of 85.3%. Through local incremental updating, only the damaged regions are corrected, whereby the geometric accuracy and topological continuity of unchanged regions are preserved while full reconstruction overhead is significantly reduced. Furthermore, by integrating three-dimensional defect mapping and retrieval-augmented semantic reasoning, the system can transform cracks and surface spalling from two-dimensional visual results into structured diagnostic information encompassing spatial location, geometric scale, influence range, risk level, and maintenance recommendations, with a semantic matching accuracy of 84.6%. These results demonstrate that frequency-domain enhancement, geometric constraints, dual-path feature fusion, adaptive change detection, local incremental updating, and domain knowledge augmentation are not mutually independent local optimizations, but collectively constitute a continuous technical chain spanning from visual acquisition, three-dimensional reconstruction, and twin updating to maintenance decision-making. Through this chain, the model credibility, change identification accuracy, and decision executability in long-term public–private partnership infrastructure maintenance are effectively enhanced.

4.7 Generalization performance and computational complexity

To validate the cross-domain generalization capability and practical inference efficiency of the proposed method, transfer experiments from synthetic datasets to real datasets, tests on publicly available general scene datasets, and end-to-end inference time statistics were conducted. The cross-domain generalization performance results are presented in Table 6.

Table 6. Cross-dataset and cross-sensor generalization performance

Training Set → Test Set	Reconstruction Chamfer Distance	Localization Mean Absolute Error (mm)	Diagnostic Semantic Matching (%)
PPP-Syn → PPP-Real	0.038	35.6	79.2
PPP-Real → PPP-Real (in-domain)	0.025	28.1	84.6
PPP-Real → BlendedMVS (public dataset)	0.042	—	—

From the experimental results, it can be observed that after the model is transferred from the synthetic dataset to the real engineering dataset, a modest degradation in performance metrics occurs, though the overall accuracy remains at a high level, demonstrating that the proposed method possesses good domain adaptation capability, through which the distribution discrepancy between simulated data and real scenes is effectively mitigated. In the test results on the publicly available BlendedMVS dataset, the reconstruction accuracy of the proposed method is comparable to state-of-the-art algorithms in the domain, indicating that the proposed reconstruction module is not limited to infrastructure scenarios but also exhibits good generalization adaptability to general weakly textured scenes. The inference efficiency test results demonstrate that the entire end-to-end intelligent management pipeline requires 2.8 seconds per image, of which 0.6 seconds are consumed by three-dimensional reconstruction and pixel-level coordinate localization, and 2.2 seconds by vision-language model semantic reasoning, satisfying the engineering application requirements of offline infrastructure inspection and periodic maintenance, with good practical deployability.

5. CONCLUSION AND FUTURE WORK

In this study, the core technical challenges in the digital twin-based intelligent operation and maintenance system for public–private partnership infrastructure—namely, insufficient accuracy in sparse-view three-dimensional reconstruction, excessive costs in cross-temporal model updating, and inadequate depth in defect semantic diagnosis—were systematically addressed, and an integrated, end-to-end image-driven intelligent management framework was constructed. A progressively layered technical system was established, in which the neural radiance field modeling process was optimized through frequency-domain adaptive enhancement and geometric regularization strategies, through which the reconstruction bottleneck in weakly textured sparse-view scenes was effectively overcome, and the chamfer distance under five-view input conditions was reduced to 0.025. Through a frequency-domain and spatial-domain dual-path adaptive fusion mechanism, the robustness of pixel-level spatiotemporal registration was enhanced, and high-precision localization of 31.5 mm was achieved under complex cross-temporal conditions over a twelve-month period. By integrating adaptive change detection with local incremental updating mechanisms, maintenance computational costs were substantially reduced while model accuracy was preserved, requiring only 6.5% of the computational cost of full reconstruction. Through three-dimensional geometric upscaling representation combined with retrieval-augmented vision-language models, a standardized engineering semantic diagnosis pipeline was established, achieving a defect semantic matching accuracy of 84.6% and an industry code citation accuracy of 78.9%. Multi-dimensional comparative experiments and module-wise ablation experiments fully validated the effectiveness and synergistic enhancement relationships among the technical components, providing comprehensive technical support for long-cycle, high-precision, and automated public–private partnership infrastructure operation and maintenance management.

Nevertheless, certain performance constraints and technical limitations persist in the proposed methodology under extremely complex operating conditions and real-time application scenarios. The initialization phase relies on reliable camera pose prior information, limiting adaptability in complex operational environments where satellite positioning is denied. The inference latency of the vision-language model remains insufficient for meeting the requirements of real-time

inspection applications. Furthermore, a comprehensive theoretical quantification and constraint system has yet to be established for the cumulative modeling errors induced by long-term iterative updating. In response to the future development trends of unmanned, intelligent, and real-time infrastructure operation and maintenance, subsequent research will introduce a four-dimensional spatiotemporal reconstruction mechanism, through which temporal evolution characteristics will be incorporated into the scene modeling pipeline, enabling continuous characterization and trend prediction of infrastructure structural deformation. Model predictive control strategies will be integrated to optimize unmanned aerial vehicle inspection paths and sampling viewpoints, achieving active adaptive optimization of data acquisition and model updating. In addition, reinforcement learning algorithms will be employed to enable the intelligent transformation from diagnostic results to graded maintenance strategies, thereby completing the full closed-loop technical system encompassing scene perception, model iteration, defect diagnosis, and maintenance decision-making, and further expanding the application boundaries of digital twin technology in the field of intelligent lifecycle control for large-scale infrastructure.

REFERENCES

[1] Selim, A.M., ElGohary, A.S. (2020). Public-private partnerships (PPPs) in smart infrastructure projects: The role of stakeholders. HBRC Journal, 16(1): 317-333. https://doi.org/10.1080/16874048.2020.1825038

[2] Verweij, S., van Meerkerk, I. (2021). Do public-private partnerships achieve better time and cost performance than regular contracts? Public Money & Management, 41(4): 286-295. https://doi.org/10.1080/09540962.2020.1752011

[3] Mahmoodian, M., Shahrivar, F., Setunge, S., Mazaheri, S. (2022). Development of digital twin for intelligent maintenance of civil infrastructure. Sustainability, 14(14): Article8664. https://doi.org/10.3390/su14148664

[4] Lu, Q., Xie, X., Parlikad, A.K., Schooling, J.M. (2020). Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance. Automation in Construction, 118: 103277. https://doi.org/10.1016/j.autcon.2020.103277

[5] Amirkhani, D., Allili, M.S., Hebbache, L., Hammouche, N., Lapointe, J.-F. (2024). Visual concrete bridge defect classification and detection using deep learning: A systematic review. IEEE Transactions on Intelligent Transportation Systems, 25(9): 10483-10505. https://doi.org/10.1109/TITS.2024.3365296

[6] Boje, C., Guerriero, A., Kubicki, S., Rezgui, Y. (2020). Towards a semantic construction digital twin: Directions for future research. Automation in Construction, 114: 103179. https://doi.org/10.1016/j.autcon.2020.103179

[7] Opoku, D.-G.J., Perera, S., Osei-Kyei, R., Rashidi, M. (2021). Digital twin application in the construction industry: A literature review. Journal of Building Engineering, 40: 102726. https://doi.org/10.1016/j.jobe.2021.102726

[8] Song, H., Yang, G., Li, H., Zhang, T., Jiang, A. (2023). Digital twin enhanced BIM to shape full life cycle digital transformation for bridge engineering. Automation in Construction, 147: 104736. https://doi.org/10.1016/j.autcon.2022.104736

[9] Torzoni, M., Tezzele, M., Mariani, S., Manzoni, A., Willcox, K.E. (2024). A digital twin framework for civil engineering structures. Computer Methods in Applied Mechanics and Engineering, 418: 116584. https://doi.org/10.1016/j.cma.2023.116584

[10] Franciosi, M., Kasser, M., Viviani, M. (2024). Digital twins in bridge engineering for streamlined maintenance and enhanced sustainability. Automation in Construction, 168: 105834. https://doi.org/10.1016/j.autcon.2024.105834

[11] Omrany, H., Al-Obaidi, K.M., Husain, A., Ghaffarianhoseini, A. (2023). Digital twins in the construction industry: A comprehensive review of current implementations, enabling technologies, and future directions. Sustainability, 15(14): 10908. https://doi.org/10.3390/su151410908

[12] Xie, Y., Takikawa, T., Saito, S., et al. (2022). Neural fields in visual computing and beyond. Computer Graphics Forum, 41(2): 641-676. https://doi.org/10.48550/arXiv.2111.11426

[13] Yuan, Y.-J., Lai, Y.-K., Huang, Y.-H., Kobbelt, L., Gao, L. (2023). Neural radiance fields from sparse RGB-D images for high-quality view synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(7): 8713-8728. https://doi.org/10.1109/TPAMI.2022.3232502

[14] Zhang, Z., Sattler, T., Scaramuzza, D. (2021). Reference pose generation for long-term visual localization via learned features and view synthesis. International Journal of Computer Vision, 129(4): 821-844. https://doi.org/10.1007/s11263-020-01399-8

[15] Khelifi, L., Mignotte, M. (2020). Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis. IEEE Access, 8: 126385-126400. https://doi.org/10.1109/ACCESS.2020.3008036

[16] Jeon, C.-H., Shim, C.-S., Lee, Y.-H., Schooling, J. (2024). Prescriptive maintenance of prestressed concrete bridges considering digital twin and key performance indicator. Engineering Structures, 302: 117383. https://doi.org/10.1016/j.engstruct.2023.117383

[17] Ouyang, A., Di Murro, V., Daakir, M., Osborne, J.A., Li, Z. (2025). From pixel to infrastructure: Photogrammetry-based tunnel crack digitalization and documentation method using deep learning. Tunnelling and Underground Space Technology, 155: 106179. https://doi.org/10.1016/j.tust.2024.106179

[18] Sun, C., Zhang, C., Xiong, N. (2020). Infrared and visible image fusion techniques based on deep learning: A review. Electronics, 9(12): 2162. https://doi.org/10.3390/electronics9122162

[19] Zhang, X., Demiris, Y. (2023). Visible and infrared image fusion using deep learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(8): 10535-10554. https://doi.org/10.1109/TPAMI.2023.3261282

[20] Wang, X., Wang, C., Liu, B., et al. (2021). Multi-view stereo in the deep learning era: A comprehensive review. Displays, 70: 102102. https://doi.org/10.1016/j.displa.2021.102102

[21] Wu, J., Wyman, O., Tang, Y., Pasini, D., Wang, W. (2024). Multi-view 3D reconstruction based on deep

learning: A survey and comparison of methods. Neurocomputing, 582: 127553. https://doi.org/10.1016/j.neucom.2024.127553
[22] Hu, Z., Hou, Y., Tao, P., Shan, J. (2021). IMGTR: Image-triangle based multi-view 3D reconstruction for urban scenes. ISPRS Journal of Photogrammetry and Remote Sensing, 181: 191-204. https://doi.org/10.1016/j.isprsjprs.2021.09.009
[23] Wang, T., Gan, V.J.L. (2024). Multi-view stereo for weakly textured indoor 3D reconstruction. Computer-Aided Civil and Infrastructure Engineering, 39(10): 1469-1489. https://doi.org/10.1111/mice.13149
[24] Liao, Y., Zhang, X., Huang, N., et al. (2024). High completeness multi-view stereo for dense reconstruction of large-scale urban scenes. ISPRS Journal of Photogrammetry and Remote Sensing, 209: 173-196. https://doi.org/10.1016/j.isprsjprs.2024.01.018
[25] Li, G., Li, K., Zhang, G., et al. (2024). Enhanced multi view 3D reconstruction with improved MVSNet. Scientific Reports, 14: 14106. https://doi.org/10.1038/s41598-024-64805-y
[26] Wei, P., Yan, L., Xie, H., et al. (2024). LiDeNeRF: Neural radiance field reconstruction with depth prior provided by LiDAR point cloud. ISPRS Journal of Photogrammetry and Remote Sensing, 208: 296-307. https://doi.org/10.1016/j.isprsjprs.2024.01.017
[27] Han, X., Liu, Z., Nan, H., Zhao, K., Zhao, D., Jin, X. (2024). PW-NeRF: Progressive wavelet-mask guided neural radiance fields view synthesis. Image and Vision Computing, 147: https://doi.org/105073. 10.1016/j.imavis.2024.105073
[28] Guan, P., Cao, Z., Yu, J., Zhou, C., Tan, M. (2021). Scene coordinate regression network with global context-guided spatial feature transformation for visual relocalization. IEEE Robotics and Automation Letters, 6(3): 5737-5744. https://doi.org/10.1109/LRA.2021.3082473
[29] Wang, S., Laskar, Z., Melekhov, I., et al. (2024). HSCNet++: Hierarchical scene coordinate classification and regression for visual localization with transformer. International Journal of Computer Vision, 132(7): 2530-2550. https://doi.org/10.1007/s11263-023-01982-9
[30] Gao, H., Dai, K., Wang, K., Li, R., Zhao, L., Wu, M. (2024). ALNet: An adaptive channel attention network with local discrepancy perception for accurate indoor visual localization. Expert Systems with Applications, 250: 123792. https://doi.org/10.1016/j.eswa.2024.123792
[31] Shi, W., Zhang, M., Zhang, R., Chen, S., Zhan, Z. (2020). Change detection based on artificial intelligence: State-of-the-art and challenges. Remote Sensing, 12(10): 1688. 10.3390/rs12101688
[32] Jung, Y., Cho, I., Hsu, S.-H., Golparvar-Fard, M. (2024). VisualSiteDiary: A detector-free vision-language transformer model for captioning photologs for daily construction reporting and image retrievals. Automation in Construction, 165: 105483. https://doi.org/10.1016/j.autcon.2024.105483
[33] Kunlamai, T., Yamane, T., Suganuma, M., Chun, P.-J., Okatani, T. (2024). Improving visual question answering for bridge inspection by pre-training with external data of image-text pairs. Computer-Aided Civil and Infrastructure Engineering, 39(3): 345-361. https://doi.org/10.1111/mice.13086
[34] Zhang, Z., Yu, Y., Pan, Z., Antwi-Afari, M.F. (2025). Training-free few-shot construction tool and material detection using pre-trained vision-language model. Computer-Aided Civil and Infrastructure Engineering, 40(30): 6004-6023. https://doi.org/10.1111/mice.70129
[35] Yin, S., Fu, C., Zhao, S., et al. (2024). A survey on multimodal large language models. National Science Review, 11(12): nwae403. https://doi.org/10.1093/nsr/nwae403
[36] Lee, J., Ahn, S., Kim, D., Kim, D. (2024). Performance comparison of retrieval-augmented generation and fine-tuned large language models for construction safety management knowledge retrieval. Automation in Construction, 168: 105846. https://doi.org/10.1016/j.autcon.2024.105846
[37] Uhm, M., Kim, J., Ahn, S., Jeong, H., Kim, H. (2025). Effectiveness of retrieval augmented generation-based large language models for generating construction safety information. Automation in Construction, 170: 105926. https://doi.org/10.1016/j.autcon.2024.105926