Smart Management and Dissemination System for Cultural Heritage Based on Knowledge Graphs and Digital Twins

Chenling Xia
Jiangsu Normal University, Xuzhou 221116, Jiangsu, China
Corresponding author: Chenling Xia (e-mail: 3020240408@jsnu.edu.cn)

ABSTRACT Cross-modal information fusion and digital twin technologies have become essential for intelligent sensing, data integration, and real-time decision support in complex engineering systems. To address the semantic inconsistency between geometric models and heterogeneous knowledge representations, this study proposes a Digital Twin– Knowledge Graph (DT-KG) deep fusion framework based on cross-modal ontology mapping for smart management and digital dissemination. The framework combines joint entity–relationship extraction, 3D point cloud semantic segmentation, IoT sensing, and spatial–semantic dynamic alignment to establish bidirectional associations between physical objects and semantic knowledge. A contrastive learning-based alignment algorithm projects heterogeneous spatial and semantic features into a unified latent space, enabling accurate cross-modal mapping and intelligent reasoning. Based on the integrated architecture, an intelligent management and interaction platform is developed to support structural anomaly analysis, semantic retrieval, and immersive visualization. Experimental results demonstrate high mapping accuracy, efficient multimodal retrieval, and stable real-time rendering performance while maintaining robust operation under concurrent requests. Beyond cultural heritage applications, the proposed DT-KG framework provides an effective methodology for intelligent information fusion, digital twin communication, multimodal sensing, and adaptive spatial information processing, offering valuable engineering references for Electromagnetic Waves, Antennas and Propagation in areas such as wireless sensing, distributed perception, and smart infrastructure management.

INDEX TERMS digital twin communication, cross-modal information fusion, multimodal sensing, spatial-semantic alignment

I. INTRODUCTION

WITH the continuous advancement of digitalization of cultural heritage, massive historical assets are accelerating their migration to the digital space. This transition ensures that the technical evolution and material craftsmanship of the industry are preserved within high-fidelity virtual environments for both archival and analytical purposes. However, in the current digital transformation framework, there is a significant heterogeneity and physical isolation between the digital twin (DT) data that represents geometric and physical features and the knowledge graph (KG) data that carries humanistic semantics, and there is an insurmountable "semantic gap" between the two [1-2]. The fragmented state of this underlying data architecture directly leads to the long-term parallel and disconnected independent operation of intelligent protection management systems in professional fields and public cultural communication systems for the general public, seriously restricting the deep exploration and comprehensive collaborative utilization of the value of multimodal cultural heritage data [3-4].

In response to the problem of fragmented underlying data mentioned above, the specific analysis objects of this study focus on two types of core heterogeneous data sources: one is the geometric and physical layer data used to construct digital twin bases (including three-dimensional laser scanning high-density point clouds that characterize macroscopic topology and microscopic morphology, as well as IoT sensing continuous monitoring sequences that characterize dynamic physical states) [5-6]; The second is unstructured historical materials and archives that contain historical evolution, material composition, and topological associations (such as restoration logs, local chronicles, and archaeological reports) [7-8]. These two types of objects respectively constitute the geometric morphology and semantic attributes of cultural heritage. Based on the complete definition of digital twins, this study injects domain rules and behavioral models through knowledge graphs to compensate for the shortcomings of traditional 3D models in logical expression, thus reconstructing a cultural heritage digital twin system that integrates spatial data and humanistic knowledge at a

higher dimension. The efficient parsing and bidirectional binding of geometric components and discrete text objects is a key prerequisite for breaking information silos and building a global intelligent management system [9-10]. Regarding the problem of parsing and fusing heterogeneous objects from multiple sources, existing research often uses pipeline based natural language processing models for text entity mining [11-12], or relies on manually preset rule dictionaries and traditional string similarity algorithms to perform cross domain binding [13-14]. However, traditional pipeline extraction models are prone to local errors cascading layer by layer when processing low resource, highly specialized historical unstructured text, making it difficult to accurately extract highly fragmented nested relationships [15-16]; At the same time, matching algorithms based on hard coded rules and statistical features lack the ability to deeply perceive the topological laws of three-dimensional space, and cannot capture continuous semantic associations in non overlapping high-dimensional feature spaces, which can easily lead to false positive mismatches and low recall rate traps in cross modal matching processes [17-18].

To address the limitations of cascading extraction errors and spatial perception deficiencies in representing complex industrial structures, this study proposes a cultural heritage DT-KG deep fusion model based on cross-modal ontology mapping. Firstly, to address the issue of error accumulation in text processing, a joint deep learning model is utilized to synchronously perform entity recognition and relationship extraction in a shared feature space. A highly reliable domain knowledge graph is constructed from unstructured historical materials, and multi-source physical data is synchronously fused to construct the geometric physical layer of the digital twin base. Secondly, in order to overcome the shortcomings of traditional algorithms in spatial semantic perception, a cross modal spatial semantic dynamic alignment algorithm is used based on point cloud semantic segmentation. This algorithm maps the high-dimensional geometric features of twins to the semantic vectors of the knowledge graph through nonlinear projection constraints and contrastive learning mechanisms in continuous latent space, thereby resolving the spatial discreteness caused by modal heterogeneity at the feature manifold level and ensuring the accuracy of entity binding. Finally, based on the underlying architecture of the "fusion of form and spirit", a comprehensive system integrating disease knowledge reasoning and immersive interaction was developed.

II. CROSS MODAL TWIN GRAPH CONSTRUCTION AND FUSION MECHANISM

A. CONSTRUCTION OF KNOWLEDGE GRAPH IN THE FIELD OF CULTURAL HERITAGE BASED ON JOINT EXTRACTION

As shown in Figure 1, the cultural heritage intelligent management and dissemination system constructed in this study adopts a multi-source heterogeneous data fusion architecture, in which the knowledge graph module and the digital twin module are bidirectional data-driven through a cross modal alignment layer [19]. In the overall system process, the historical archive text is preprocessed and enters the knowledge graph construction module, where it is processed in parallel with 3D laser scanning and IoT sensing data. Finally, degradation analysis and immersive semantic retrieval functions are implemented at the application layer. In response to the low resource and strong professional characteristics exhibited by unstructured text of cultural heritage, the feature extraction process selects cleaned cultural heritage text corpus (such as restoration logs and historical documents) as the input.

Note: IoT stands for Internet of Things. Firstly, define the input unstructured text sequence as:

$$S = \{w_1, w_2, \ldots, w_n\} \tag{1}$$

w_i represents the i-th character in the sequence, and n is the total length of the text. To capture the implicit associations between ancient books and archives at a deep level, a domain-specific BERT-based

encoder (e.g., SikuBERT or GuwenBERT) is introduced to extract contextual features. To bridge the linguistic divergence between classical Chinese (prevalent in heritage archives) and modern technical Chinese, the model underwent secondary pre-training on a multi-epoch corpus of 50,000 classical texts, utilizing a character-level masking strategy to capture archaic syntactic structures and specialized industrial terminology. The underlying semantic encoding process is represented as:

$$H = \text{Encoder}(S) \in \mathbb{R}^{n \times d_{\text{embed}}} \tag{2}$$

In the formula, H is the hidden state feature matrix output after encoding, and d_{embed} represents the word vector dimension. By finely controlling the d_{embed}, domain independent noise can be effectively filtered out and the representation ability of text semantic features can be directly optimized.

In the decoding stage of entity and relationship features, construct a joint information extraction model architecture for parameter sharing. To eliminate the error cascade caused by traditional pipeline operations, this model synchronously performs entity boundary definition and relationship classification within the same shared feature space H. This joint architecture prevents error propagation because it allows the gradients of entity recognition and relationship extraction tasks to interact during backpropagation, enabling the model to utilize relationship constraints as a prior to correct entity boundaries (and vice versa) rather than treating the noisy output of one stage as the absolute input for the next. The entity boundary recognition mechanism calculates the confidence level of each character as the starting and ending position of the entity by mapping it to a fully connected layer and a nonlinear activation function.

$$P^{\text{ent}} = \sigma(W_{\text{ent}} H + b_{\text{ent}}) \tag{3}$$

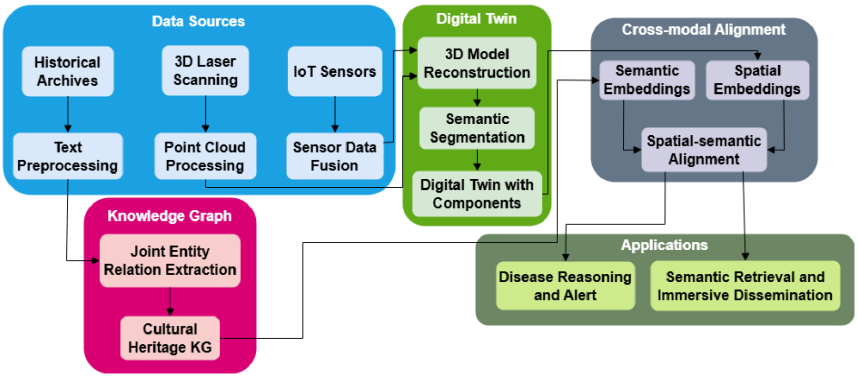

FIGURE 1. Overall system architecture diagram

In the formula, P^{ent} represents the probability distribution matrix of the entity boundary, W_{ent} and b_{ent} are the learnable weight matrix and bias vector of the entity discrimination layer, and σ represents the Sigmoid activation function.

Subsequently, the model extracts the currently determined hidden state slices of the head entity and aggregates them into the head entity feature vector h_{head} through pooling operations. The relationship extraction module uses the concatenation result of h_{head} and global feature H as a prior condition to determine the topological relationship between the current entity pair:

$$P^{\text{el}} = \text{Softmax}(W_{\text{rel}}[H; h_{\text{head}}] + b_{\text{rel}}) \tag{4}$$

In the formula, P^{el} represents the probability distribution on a predefined set of relationships, W_{rel} is the relationship classification weight matrix, b_{rel} is the corresponding bias term, $[\cdot;\cdot]$ represents vector concatenation operation.

The training optimization objective of this joint model is based on the multi-task joint loss function:

$$L_{\text{joint}} = L_{\text{ent}} + \alpha L_{\text{rel}} \tag{5}$$

In the formula, L_{joint} represents the total loss, L_{ent} and L_{rel} represent the relationship between entity recognition loss calculated using binary cross entropy and multi-class cross entropy calculation to extract loss, and α is used as a penalty coefficient to dynamically balance the gradient feedback amplitude of the two sub tasks. After model inference and confidence filtering, the system finally outputs structured knowledge triplets (such as <Changxin Palace Lantern, Excavated from Mancheng Han Tomb>), and stores them in the native graph database to support subsequent cross modal alignment and semantic retrieval applications. The semantic entity vector output by the knowledge graph module can serve as one of the input terminals for the cross modal alignment algorithm, and can be fused and mapped with the component feature vectors generated by the 3D point cloud semantic segmentation module described in Section 2.2 in the alignment layer of Section 2.3. This parallel processing architecture ensures that physical geometric information and human semantic information remain independently represented in the feature extraction stage, and deep binding is achieved in the subsequent alignment stage, laying the underlying data foundation for the final disease inference warning and immersive propagation application of the system.

B. SEMANTIC SEGMENTATION AND COMPONENT EXTRACTION OF 3D POINT CLOUDS FOR DIGITAL TWINS

To parse the complete digital twin grid or point cloud model into a semantically meaningful set of independent components, the underlying algorithm takes the original 3D point cloud data of cultural heritage, which includes XYZ coordinates and RGB (Red, Green, Blue) color information, as the input matrix [20-21]. Let the input raw point cloud be:

$$P = \{p_1, p_2, \ldots, p_N\} \in \mathbb{R}^{N\times 6} \tag{6}$$

N represents the total number of sampling points. A single data point is defined as $p_i = [c_i \| f_i]$, where $c_i \in \mathbb{R}^3$ is the spatial geometric coordinate vector (XYZ), $f_i \in \mathbb{R}^3$ is the surface color feature vector (RGB), and the symbol $\|$ represents the concatenation operation of multimodal feature dimensions.

In the stage of local geometric topology perception, key parameters such as the radius of the search sphere and the number of segmentation categories used for feature extraction directly determine the analytical precision of spatial geometric details. Establish a spherical neighborhood search domain centered on coordinate c_i:

$$\mathcal{N}(p_i) = \{p_j \in P \mid \|c_i - c_j\|_2 \le R_{\text{search}}\} \tag{7}$$

In the formula, p_j is the coordinate point within the neighborhood, and R_{search} is the preset radius of the local search sphere. By adjusting the scale of R_{search}, the dynamic control of the feature receptive field can be realized: reducing the radius value can preserve the micro carving details such as arch of wooden architecture, tenon and mortise; Increasing the radius value is used to capture macroscopic topological structures such as column foundations and

beam frames. After aggregating neighborhood information through a symmetric pooling function, a high-dimensional point level feature matrix $F \in \mathbb{R}^{N\times d}$ is generated, where d is the number of feature channels after dimensionality enhancement [22-23].

Subsequently, high-dimensional features are mapped to a predefined K_{num} component category space, and the semantic segmentation probability matrix $P \in \mathbb{R}^{N\times K_{\text{num}}}$ is calculated using a multi-layer perceptron. The classification operation formula is expressed as:

$$P = \text{Softmax}(W_{\text{seg}}F^T + B_{\text{seg}}) \tag{8}$$

In the equation, $W_{\text{seg}} \in \mathbb{R}^{K_{\text{num}}\times d}$ is the learnable weight parameter of the fully connected layer at the end of the network, $B_{\text{seg}} \in \mathbb{R}^{K_{\text{num}}}$ is the corresponding bias vector, and T represents the matrix transposition operation. The discrete semantic label l_i of the current sampling point p_i is determined by the maximum a posteriori probability, that is:

$$l_i = \arg\max_{k\in\{1,\ldots,K_{\text{num}}\}} P_{i,k} \tag{9}$$

In the formula, k is the category index variable. The set size of the category number K_{num} directly maps the depth of the hierarchy at which the twin is deconstructed. Based on the above discrete labels, the point cloud is stripped and reorganized into independent semantic clusters $C_k = \{p_i \mid l_i = k\}$. For each semantic cluster, extract its orthogonal spatial extremum points to generate a three-dimensional bounding box, and calculate the vertex set V_{bbox} of its bounding box as follows:

$$V_{\text{bbox}} = \left[\min_{p_i\in C_k}(c_i), \max_{p_i\in C_k}(c_i)\right] \tag{10}$$

In the formula, min and max respectively represent the operation of finding the lower and upper bounds of three-dimensional coordinates in each coordinate axis direction within the cluster.

As shown in Figure 2, the original point cloud data is accurately stripped into independent component sets with clear semantic labels after being processed by semantic segmentation algorithms. The left subfigure (a) shows the distribution state of the raw point cloud without processing, and the point cloud data presents an undifferentiated spatial scatter pattern; The sub image on the right (b) presents the component classification results after segmentation is completed. The blue dot cluster corresponds to the column component, the green dot cluster corresponds to the platform component, and the red dot cluster corresponds to the arch of wooden architecture component. The colored wireframe around each semantic cluster is the three-dimensional space bounding box generated by the above bounding box formula. This visualization result intuitively verifies that the point cloud semantic segmentation module can deconstruct continuous spatial geometric data into discrete semantic component units, providing standardized physical anchor points for subsequent cross modal alignment.

The set of semantic components with bounding boxes output by this module can be used as the spatial input X_{3D} for the cross modal space semantic alignment algorithm in Section 2.3, and mapped and matched with the entity node vectors of the knowledge graph constructed in Section 2.1 in the shared latent space. This progressive processing flow from raw point clouds to semantic components and then to cross modal binding ensures that physical geometric information obtains humanistic semantic annotations while maintaining spatial accuracy, ultimately supporting system level disease inference warning and immersive semantic retrieval applications.

C. CROSS MODAL SPACE SEMANTIC DYNAMIC ALIGNMENT ALGORITHM

To bridge the "semantic gap" between the geometric and physical data of the digital twin model and the humanistic semantic data of the knowledge graph, and achieve bidirectional binding between physical states and historical knowledge, the system maps heterogeneous data to a unified feature space through a cross modal space semantic dynamic alignment algorithm [24-25]. The input of this algorithm contains two independent feature representations: one is the feature vector X_{3D} of the 3D component (covering spatial position, geometric shape, and label category); The second is the feature vector Y_{KG} of entity nodes in the knowledge graph (KG), which includes entity names, attribute features, and topological relationships. Due to the fact that the two types of features are in a non overlapping high-dimensional space, they need to be transformed into a shared latent space dimension through nonlinear projection. The feature projection process is defined as:

$$Z_{\text{spa}} = \tanh(W_\alpha X_{3D} + B_\alpha) \tag{11}$$

$$Z_{\text{sem}} = \tanh(W_\beta Y_{KG} + B_\beta) \tag{12}$$

In the formula, Z_{spa} and Z_{sem} represent the mapped continuous spatial feature vector and semantic feature vector, respectively; W_α and W_β are the projection matrices responsible for the corresponding modal transformations of the two groups; B_α and B_β are the bias terms of their respective network layers; Tanh represents the hyperbolic tangent nonlinear activation function.

In the shared latent space, the model adopts contrastive learning mechanism to optimize the distribution structure of heterogeneous vectors [26-27]. This optimization strategy forcibly narrows the geometric distance between positive sample pairs (representing 3D components of the same entity and KG nodes) in the feature space by comparing the loss function, while pushing away negative sample pairs. The calculation formula for the loss function is as follows:

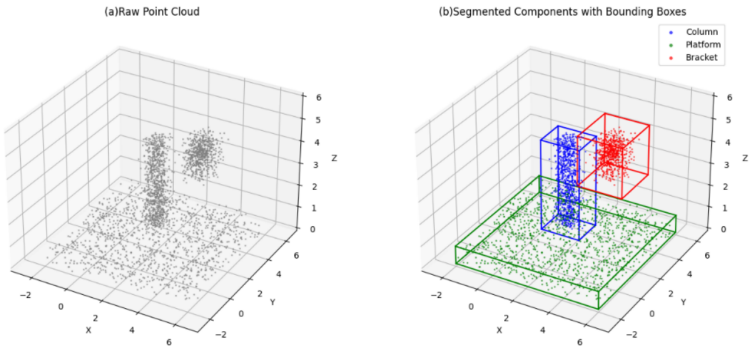

FIGURE 2. Spatial segmentation results of 3D point cloud components: (a) Original point cloud; (b) Segmented component with bounding box

$$L_{\text{align}} = y_{\text{ind}}(D_{\text{dist}})^2 + (1 - y_{\text{ind}})\max(0, m_{\text{margin}} - D_{\text{dist}})^2 \tag{13}$$

In the formula, L_{align} is the calculated cross modal alignment loss value; y_{ind} is the pairing indicator variable (when the input is a positive sample pair with a mapping relationship, the value is 1, and when there are unrelated negative sample pairs, the value is 0); D_{dist} represents the Euclidean distance between vectors Z_{spa} and Z_{sem}; m_{margin} is a margin parameter used to limit the minimum safe distance threshold at which unrelated negative sample pairs can be pushed away.

As shown in Figure 3, after cross modal alignment algorithm mapping, the 3D component feature vectors (circular markers) and KG semantic feature vectors (triangular markers) exhibit significant clustering alignment characteristics in the shared latent space. Different colors represent different categories of entities, and each set of circles and triangles of the same color corresponds to the physical components and semantic nodes of the same cultural heritage entity. From the scatter distribution, it can be seen that under the optimization of the contrastive learning mechanism, positive sample pairs have achieved aggregation in the two-dimensional latent space projection, while maintaining clear boundary spacing between entity clusters of different categories. This spatial distribution feature intuitively verifies the effectiveness of the synergistic effect of nonlinear projection and contrastive loss, proving that the algorithm can successfully resolve the spatial discreteness caused by modal heterogeneity in continuous latent space, laying a solid feature foundation for subsequent accurate matching.

Entering the inference matching stage, the system calculates the cosine similarity $S_{\cos}$ between spatial vectors and semantic vectors to perform dynamic alignment of entities. The criteria for matching strictly depend on the preset similarity threshold parameter τ_{sim}. When the judgment condition $S_{\cos} \geq \tau_{\text{sim}}$ is met, the algorithm confirms that the spatial entity and semantic entity are the same object. After global matching operation, the system finally outputs a cross modal mapping dictionary (Key-Value alignment list) containing all confirmed pairings, establishing a robust bidirectional indexing mechanism in the underlying architecture. The cross modal mapping dictionary can be the core object of experimental evaluation in Chapter 4, and its mapping accuracy directly determines the effectiveness of disease inference warning and immersive semantic retrieval applications at the system level. At the same time, the performance of this alignment algorithm can be quantitatively validated in section 4.3 through multidimensional indicators such as precision, recall, and F1 score, and compared horizontally with traditional rule matching and similarity matching baseline algorithms.

III. EXPERIMENTAL ENVIRONMENT AND EVALUATION STRATEGY

A. CONSTRUCTION AND PREPROCESSING OF MULTIMODAL DATASET

To support the training and validation of cross modal deep fusion models, the construction of the experimental dataset covers two heterogeneous modalities: semantic text corpus and physical 3D sensing data. In the construction of the semantic corpus, the collection of raw texts comprehensively covers unstructured historical archives such as archaeological reports, local chronicles, and maintenance records over the years. In response to the complex layout of paper files and ancient books, optical character recognition (OCR) technology is used to perform batch digital conversion [28-29]. To ensure the input quality of the subsequent joint extraction model and eliminate data ambiguity, the cleaning specification is strictly set to automatically remove garbled characters, meaningless watermarks, and non-standard tabs generated during scanning and transcription using regular matching algorithms. The process of text standardization can be formally defined as:

$$S_{\text{clean}} = F_{\text{regex}}(S_{\text{raw}}) \tag{14}$$

In the formula, S_{raw} represents the original OCR recognized text sequence, F_{regex} represents the filtering function based on regular expressions, and S_{clean} is the standardized clean text sequence. The cleaned continuous corpus is truncated into fixed length short sentence sequences based on specific punctuation and semantic boundaries, and then

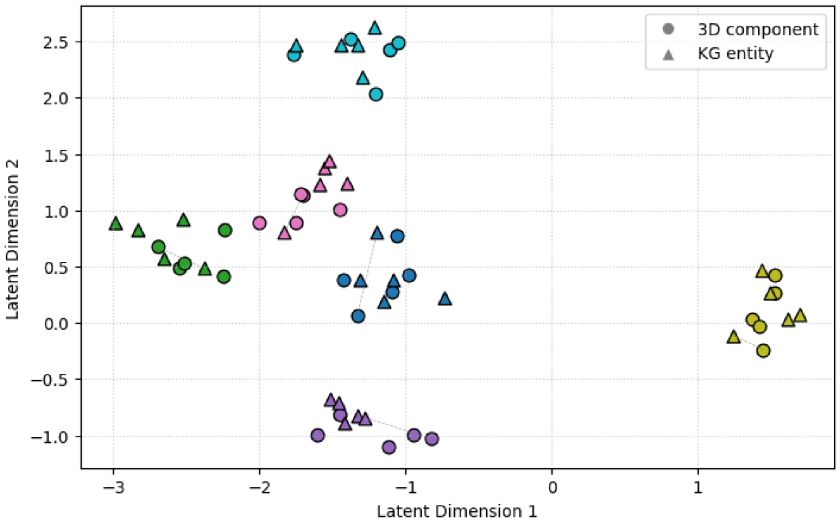

FIGURE 3. Distribution of 3D entity and knowledge node features

the BIO (Begin, Inside, Outside) sequence annotation mechanism is introduced to manually and rule-based pre-annotate the boundaries and topological relationships of specific entities, thereby transforming unstructured characters into standardized training sets that can be directly ingested by deep neural networks.

The physical data acquisition mechanism aims to obtain high fidelity geometric modes and dynamic physical features. The acquisition of spatial 3D information integrates unmanned aerial vehicle oblique photography and high-precision 3D laser scanning technology on the ground to capture the macroscopic topological structure and surface microscopic point clouds of cultural heritage without loss [30-31]. To solve the problem of viewpoint offset and local coordinate system fragmentation caused by multi-source heterogeneous device scanning, the Iterative Closest Point (ICP) algorithm is used in the preprocessing stage for rigid registration and fusion of multi-station data [32-33]. The objective function of registration optimization is defined as:

$$E(R_t) = \sum_{i=1}^{N} \|R_t p_i - q_i\|_2^2 \tag{15}$$

In the formula, p_i and q_i are the corresponding point pairs in the source point cloud and the target point cloud, R_t is the rigid transformation matrix to be solved, and $\|\cdot\|_2$represents the Euclidean distance norm. By minimizing $E(R_t)$, output a high-density seamless point cloud model in a globally unified coordinate system. In addition, for the data acquisition standards of IoT sensors deployed at key load-bearing components and environmental nodes, Kalman filtering algorithm is used to smooth the original sensing signals to filter out high-frequency environmental noise interference [34-35]. The state estimation update equation is expressed as:

$$\hat{x}_t = \hat{A}x_{t-1} + K_t(z_t - O_m\hat{x}_{t-1}) \tag{16}$$

In the equation, $\hat{x}_t$ is the estimated state value at time t, A is the state transition matrix, K_t is the Kalman gain, z_t is the observation vector, and O_m is the observation matrix. The one-dimensional dynamic monitoring sequence after filtering processing needs to be forcibly aligned in time and space based on a unified absolute timestamp and three-dimensional static component coordinates. The time synchronization judgment criteria are set as follows:

$$|t_{\text{sensor}} - t_{3D}| \leq \delta_t \tag{17}$$

In the formula, t_{sensor} and t_{3D} are the timestamps of sensor data and 3D scanning data, respectively, and δ_t is the maximum allowed time synchronization error threshold. This spatiotemporal alignment mechanism lays the foundation for cross modal data fusion of twin bases, ensuring strict consistency and temporal correlation of the data input into the hyperparameter configuration model described in section 3.2, and providing a standardized data base for the evaluation index calculation in section 3.3.

B. EXPERIMENTAL ENVIRONMENT CONFIGURATION AND HYPERPARAMETER SETTING

To ensure efficient training and high concurrency retrieval response of cross modal deep fusion models, the underlying hardware relies on high-performance graphics processor computing clusters for unified deployment to solve the bottleneck of large-scale matrix operations in point cloud segmentation and high-dimensional feature mapping processes. At the data persistence layer, in response to the storage requirements of semantic nodes and their complex topological relationships, a native graph database is selected as the graph base, and its index free adjacency mechanism is utilized to significantly reduce the computational latency of multi-hop relationship inference and semantic retrieval. For the front-end presentation layer, the real-time visualization of digital twin scenes is driven by a lightweight

3D rendering engine that supports hardware acceleration. Through dynamic hierarchical detail loading, visual cone removal algorithms, and a multi-scale level-of-detail (LOD) data management strategy, it effectively avoids the risks of terminal rendering lag and memory overflow caused by massive grid fragment parsing. To mitigate the computational and storage burden of high-fidelity models for large-scale sites, the system employs an out-of-core rendering mechanism and octree-based spatial indexing, ensuring that only the relevant geometric subsets are loaded into GPU memory, thus maintaining a stable rendering performance of over 60 FPS.

In terms of model optimization and feature space convergence, the performance of each sub network module strictly depends on the precise optimization and control of hyperparameters. For the joint information extraction model, the word vector dimension parameter is set to $d_{\text{embed}} = 768$, which is consistent with the hidden layer dimension of mainstream pre-trained language models to ensure the adequacy of semantic feature extraction. The loss balance coefficient $\alpha = 0.6$ is used to adjust the gradient contribution ratio of entity recognition and relationship extraction subtasks, preventing a single task from dominating the optimization direction of the model.

In the processing flow for 3D twin bases, the radius of the search sphere for local geometric topology perception is set to $R_{\text{search}} = 0.15\,\text{m}$. This scale can effectively capture the macro topological structures such as column base and beam frame while retaining the micro carving details such as arch of wooden architecture, tenon and mortise. The point cloud sampling scale $N_{\text{points}} = 32768$ has been established as the core control variable for balancing geometric fidelity and computational complexity, and the feature dimension of the input matrix is forcibly aligned through the farthest point sampling mechanism.

In the cross modal alignment network, the projection dimension of the shared latent space is set to $d_{\text{proj}} = 256$, which provides sufficient feature representation capability to accommodate the fusion representation of spatial geometry and semantic information. The margin parameter $m_{\text{margin}} = 1.2$ in the contrastive loss function is used to limit the minimum safe distance threshold for unrelated negative sample pairs to be pushed away in the feature space. The similarity determination threshold $\tau_{\text{sim}} = 0.75$ is used as the final decision boundary for cross modal entity matching, and the setting of this value directly determines the trade-off between the accuracy and recall of the mapping task.

At the optimizer configuration level, the adaptive learning rate $\eta = 1.0 \times 10^{-4}$ is combined with a decay strategy to ensure the initial gradient descent efficiency while avoiding oscillations near the global optimal solution. Batch processing scale $B = 32$ is set in combination with video memory capacity constraints to maintain the generalization ability and gradient smoothness of the training process. The configuration logic and specific set values of the above core hyperparameters are detailed in Table 1.

C. EVALUATION INDICATORS AND COMPARISON BENCHMARKS

For the dual application scenarios of disease warning and public interaction, the construction of system stress testing indicators focuses on spatiotemporal response and rendering stability under high concurrency conditions. On the smart management end, the core assessment indicator is set as the end-to-end physical delay of multi-source sensor data access and graph anomaly inference, in order to quantify the dynamic diagnosis rate of the warning module. On the public interaction end, random burst traffic simulating Poisson distribution is injected to monitor the real-time rendering frame rate (FPS) of 3D scenes and the average response time of cross modal semantic retrieval under concurrent requests from multiple terminals, in order to verify the system's anti-crash robustness and usability boundary after massive knowledge ontology fusion. To rigorously evaluate the true mapping performance of cross modal spatial semantic dynamic alignment algorithms, a core quantitative evaluation matrix is constructed using precision P_{val}, recall R_{val}, and their harmonic mean - the comprehensive evaluation index $F1_{\text{score}}$. The computer system directly relies on the judgment count in the final cross modal mapping dictionary: let T_{pos} be true positives (i.e. correctly aligned spatial components and semantic entity pairs by the algorithm), F_{pos} be false positives (i.e. incorrectly bound entity pairs by the algorithm), and F_{neg} be false negatives (i.e. associated objects that already have corresponding relationships but were missed by the algorithm). The mathematical definition of a measurement system is as follows:

$$P_{\text{val}} = \frac{T_{\text{pos}}}{T_{\text{pos}} + F_{\text{pos}}} \tag{18}$$

$$R_{\text{val}} = \frac{T_{\text{pos}}}{T_{\text{pos}} + F_{\text{neg}}} \tag{19}$$

$$F1_{\text{score}} = 2 \times \frac{P_{\text{val}} \times R_{\text{val}}}{P_{\text{val}} + R_{\text{val}}} \tag{20}$$

At the same time, two sets of heterogeneous algorithms were established as comparison baselines, namely hard coded rule matching based on expert experience and traditional similarity matching based on local text or attribute editing distance. The core mechanism and verification framework of the above benchmarks are summarized in Table 2.

IV. RESULT ANALYSE

A. PERFORMANCE ANALYSIS OF DOMAIN KNOWLEDGE EXTRACTION MODEL

To verify the actual effectiveness of the joint deep learning model in extracting entity relationships from unstructured historical materials and archives, the evaluation process follows the precision, recall, and F1 value measurement

TABLE 1. Core hyperparameter configuration table

Module	Parameter	Symbol	Value
Joint Entity-Relation Extraction	Embedding Dimension	d_{embed}	768
Joint Entity-Relation Extraction	Loss Balance Coefficient	α	0.6
Point Cloud Segmentation	Search Radius	R_{search}	0.15 m
Point Cloud Segmentation	Sampling Points	N_{points}	32768
Cross-modal Alignment	Projection Dimension	d_{proj}	256
Cross-modal Alignment	Margin Parameter	m_{margin}	1.2
Cross-modal Alignment	Similarity Threshold	τ_{sim}	0.75
Optimization	Learning Rate	η	1.00E-04
Optimization	Batch Size	B	32

TABLE 2. Comparison of benchmark algorithms and evaluation frameworks

Method	Label	Core Mechanism	Evaluation Metrics
Rule-based Matching	Rule-based	Regular Expressions and Expert Rules	P_{val}, R_{val}, $F1_{\text{score}}$
Traditional Similarity	Traditional	Edit Distance and Term Frequency-Inverse Document Frequency	P_{val}, R_{val}, $F1_{\text{score}}$
Proposed Method	Ours	Cross-modal Contrastive Learning	P_{val}, R_{val}, $F1_{\text{score}}$

system established in Section 3.3. The manually annotated test set generated in the preprocessing stage was input into the model for forward inference, and the quantitative results showed significant advantages of this architecture in multi-dimensional domain knowledge extraction tasks. Thanks to the design of synchronously defining entity boundaries and classification relationships in the underlying shared feature space, the defect of cascading errors layer by layer in traditional pipeline architecture is effectively suppressed, and the representation ability of text semantic features is substantially optimized. In the fine-grained analysis of specific entity relationship categories, the model exhibits differentiated convergence characteristics in terms of extraction efficiency for different semantic dimensions. As shown in Figure 4, the extraction performance of cross relationship types exhibits high stability in all five core dimensions. Among them, the extraction effect of "spatial topology" relationships (such as excavation location and component attachment) is the most significant, with an precision rate of 93.8%, a recall rate of 94.6%, and an F1 value of 94.2%; The precision of the "temporal evolution" relationship (such as historical reconstruction and repair years) is 91.2%, the recall rate is 92.3%, and the F1 value is 91.8%.

Faced with the common problem of implicit references and long-range entity associations in local chronicles and repair logs, the model uses a joint loss function (see formula 5) for gradient backpropagation, forcing the hidden state matrix to enhance the capture of global context dependencies during the encoding stage. This punishment adjustment mechanism significantly exceeds the reliability threshold for extracting highly fragmented nested relationships such as "disease features" and "material composition", with F1 values of 89.5% and 88.7%, respectively. In addition, the performance of the "mining relationships" category is equally outstanding, with an precision rate of 92.5%, a recall rate of 93.7%, and an F1 score of 93.1%. The above results effectively solve the operational bottleneck of automated extraction of complex topological structures in rare historical corpora, laying a reliable semantic data foundation for subsequent cross modal alignment and knowledge reasoning applications.

B. VERIFICATION OF SEMANTIC SEGMENTATION ACCURACY FOR 3D COMPONENTS

For the accuracy evaluation of parsing digital twin models into independent component sets, this paper adopts Intersection over Union (IoU) as the core quantitative indicator. The original 3D point cloud in the validation set can be input into a multi-layer perceptron network, and the generated semantic cluster set can be spatially overlapped with the real annotated entity for calculation. The mathematical definition of intersection to union ratio is:

$$\text{IoU} = \frac{A_{\text{pred}} \cap A_{\text{gt}}}{A_{\text{pred}} \cup A_{\text{gt}}} \tag{21}$$

In the formula, A_{pred} represents the predicted segmentation area, and A_{gt} represents the actual annotated area. The test data shows that within the set component category space, the global mean intersection over union (mIoU) exhibits significant convergence characteristics as the number of sampling points increases. To analyze the segmentation robustness under different numbers of sampling points, the experiment conducted control variable testing using the farthest point sampling algorithm with five gradient point cloud input scales. As shown in Figure 5, the operation trajectory reveals a nonlinear mapping relationship between sampling density and segmentation accuracy. When the number of sampling points is 2048, the loss of geometric details leads to chaos in the output of the network classification probability matrix, and the overall mIoU is only 64.1%, of which the arch of wooden architecture component is most significantly affected due to its complex shape, with IoU

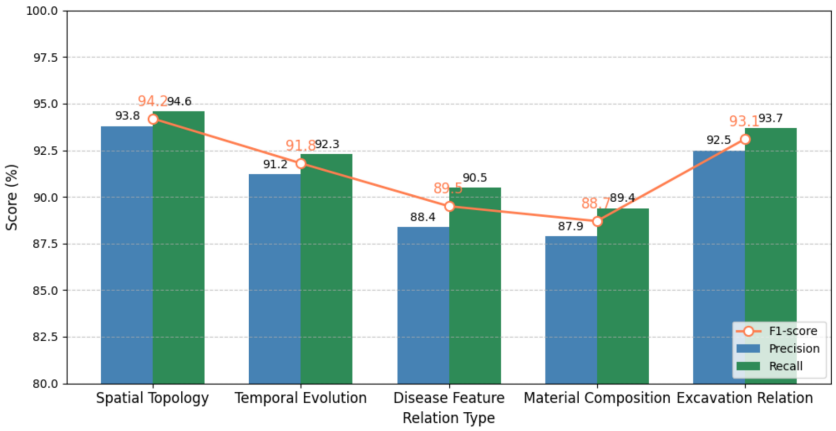

FIGURE 4. Extraction performance across relationship types (precision, recall, F1)

as low as 60.3%. As the sampling scale increases to 4096 points, the information entropy of the feature aggregation matrix begins to saturate, and the overall mIoU jumps to 78.5%. Due to the stable macroscopic topology structure of the column base component, the IoU first breaks through 84.6%.

When the number of sampling points reaches 8192, the three performance curves show obvious inflection point characteristics, with the overall mIoU reaching 86.2%, the arch of wooden architecture component IoU rising to 81.5%, and the column base component IoU breaking through 90.1%. The threshold has been confirmed as the convergence critical point, and increasing the sampling density significantly slows down the accuracy gain. At the scale of 16384 points, the overall mIoU is 88.9%, the arch of wooden architecture component IoU is 84.7%, and the column base component IoU is 92.4%. When the number of sampling points reached 32768, the performance curve entered a stable period, with the overall mIoU stable at 89.7%, the arch of wooden architecture component IoU maintained at 86.5%, and the column base component IoU peak reached 93.2%.

The above convergence characteristics indicate that the 8192 point sampling scale can achieve the optimal balance between geometric fidelity and computational complexity. For macroscopic topological structures such as column foundations and platform foundations, the peak IoU of boundary segmentation reaches 93.2%; For the complex arch of wooden architecture and mortise tenon node, depending on the fine constraint of searching the ball radius in the local feature extraction link, it effectively prevents the edge point cloud feature from being excessively smooth, and finally keeps its intra class IoU at a high level of 86.5%. This high fidelity segmentation result under large-scale sampling conditions minimizes the coordinate drift error of the bounding box vertex set, demonstrating that this deconstruction mechanism can provide highly stable and reliable physical anchor points for subsequent spatial semantic alignment algorithms.

C. PERFORMANCE EVALUATION OF CROSS MODAL SPACE SEMANTIC ALIGNMENT ALGORITHM

To verify the actual effectiveness of the cross modal space semantic alignment algorithm in crossing the "semantic gap", this paper imports the 3D component vectors extracted through feature extraction and the knowledge graph entity node vectors into the alignment network in batches for matching testing. The evaluation system adopts the precision, recall, and F1 value indicators defined in Section 3.3 to quantify the reliability of the mapping between physical coordinates and historical text nodes. The test data feedback shows that the algorithm achieved a precision of 95.6% in entity component mapping tasks, a recall rate of 94.2%, and a comprehensive F1 score of 94.9%. The achievement of this high-performance indicator is mainly attributed to the synergistic effect of nonlinear projection constraints and contrastive learning mechanisms in the continuous latent space. By introducing strict margin parameters in the contrastive loss function, the network forcibly pushes away negative sample pairs that are morphologically similar but semantically contradictory in the shared high-dimensional space, while greatly compressing the Euclidean distance between positive sample pairs, and resolving the spatial discreteness caused by modal heterogeneity at the level of the underlying feature manifold.

In the horizontal comparison with two sets of heterogeneous benchmark algorithms, the superiority of the dynamic alignment mechanism has been further confirmed. As shown in Figure 6, the three methods exhibit significant gradient differences in various indicators. The hard coded rule matching method based on expert experience is limited by manually preset regular expressions and cannot generalize the flexible changes in component aliases in historical archives, resulting in serious missing matching phenomena. Its precision is 80.2%, recall rate is only 62.7%, and F1 value is as low as 70.4%. Although the traditional similarity matching method based on local text or attribute editing distance has improved its precision to 86.5%, it is prone to binding erroneous entities with similar name editing distances due to the lack of three-dimensional spatial

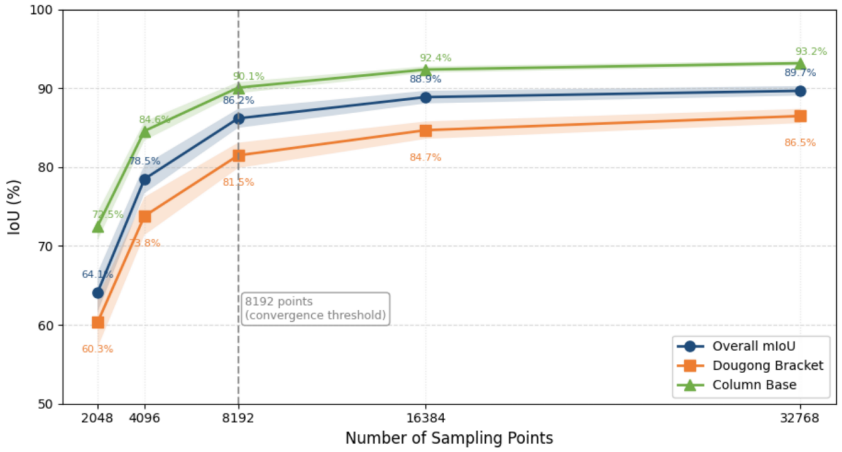

FIGURE 5. Convergence curves of Intersection over Union (IoU) for point cloud segmentation at different sampling scales

topology perception ability, resulting in a large number of false positive mismatches, with a recall rate of 79.3% and an F1 value of 82.7%.

In contrast, the cross modal depth model proposed in this paper utilizes a dynamic cosine boundary threshold determination mechanism to directly calculate the dense vector similarity that integrates spatial geometry and topological relationships within the shared latent space. The formal definition of matching decision logic is:

$$\text{Match} = \begin{cases} 1, & \text{if } S_{\cos}(v_{3D}, v_{KG}) \geq \tau, \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

In the formula, $S_{\cos}$ represents the cosine similarity function, v_{3D} and v_{KG} represent the three-dimensional component vector and knowledge graph entity vector, respectively, and τ is the preset similarity threshold. This mechanism completely avoids the semantic bias caused by surface string alignment, enabling the proposed method to significantly surpass the benchmark model in three core evaluation indicators: precision (95.6%), recall (94.2%), and F1 score (94.9%). The above results effectively solve the problem of the inability to directly address and map physical coordinates to historical texts, providing a highly reliable cross modal indexing foundation for subsequent intelligent management and dissemination applications.

D. SYSTEM MULTIMODAL RETRIEVAL PERFORMANCE AND RENDERING EFFICIENCY

To verify the usability and lightweight features of the system after integrating a large knowledge graph, the testing environment injected continuous burst traffic simulating Poisson distribution to perform concurrent stress testing. To alleviate the high computational overhead caused by massive multi-hop inference of nodes, the underlying retrieval engine called the index free adjacency mechanism of the native graph database, and combined it with a memory level non relational cache pool to perform hot data preloading on high-frequency triplets. The core performance evaluation focuses on two key indicators: the average response time T_{resp} for cross modal semantic retrieval and the real-time rendering frame rate R_{fps} for 3D scenes. Their mathematical definitions are as follows:

$$T_{\text{resp}} = \frac{1}{N} \sum_{i=1}^{N} (t_{\text{end}}^{i} - t_{\text{start}}^{i}) \tag{23}$$

$$R_{\text{fps}} = \frac{1}{\Delta t} \sum_{j=1}^{M} F_j \tag{24}$$

In the formula, N represents the total number of requests, t_{start}^{i} and t_{end}^{i} represent the start and end timestamps of the i-th request, M represents the total number of frames rendered within the statistical time window Δt, and F_j is the rendering flag of the j th frame.

During the testing period when the concurrent request volume increased from 100 to 1000 in a stepwise manner, the system memory stack did not overflow or crash due to increased data complexity. As shown in Figure 7, the response time and rendering frame rate exhibit predictable convergence characteristics as the concurrent load changes. In the low load range (100-300 QPS), the time consumption of node addressing and dictionary lookup operations in cross modal semantic retrieval is effectively compressed, and the average response time remains stable in the range of 85ms to 115ms. The 3D scene rendering frame rate remains at a high level of 68 FPS to 72 FPS.

As the concurrent load gradually climbs to the mid to high range (400-700 QPS, Questions Per Second), the response time shows a linear growth trend, with the median gradually increasing from 125ms to 145ms, and the frame rate showing slight fluctuations, with the median remaining in the range of 63 FPS to 68 FPS. When the concurrent request volume reaches the maximum load (800-1000 QPS), the system performance enters a steady saturation state, and the global average response time converges to 142ms, successfully dropping to the preset baseline of 150ms. The output speed of the rendering of multi-terminal twin scenes

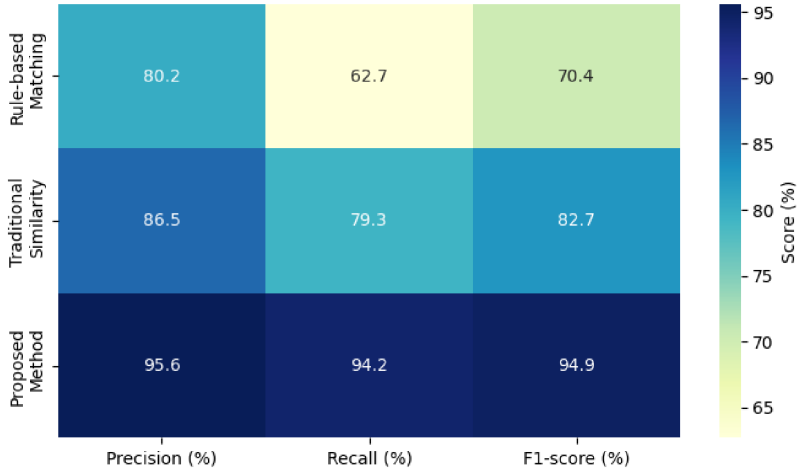

FIGURE 6. Comparison heatmap of multiple matching algorithms in terms of precision, recall, and F1 score

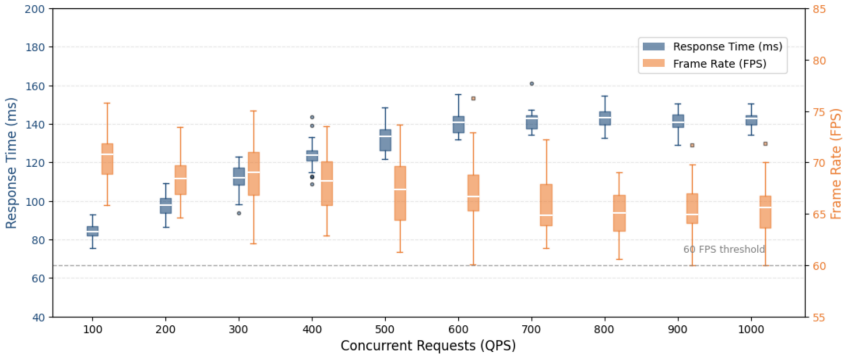

FIGURE 7. Box plot distribution of average response time and real-time rendering frame rate under concurrent requests

has slightly decreased, but it has consistently exceeded the smoothness threshold of 60 FPS, with the lowest median frame rate remaining around 65 FPS.

In the verification of rendering efficiency in the twin visualization dimension, in order to eliminate the terminal computing bottleneck caused by high-density point clouds and complex grid analysis, the front-end graphics pipeline has fully taken over and enabled dynamic hierarchical detail scheduling mechanism and visual cone hardware removal algorithm. This combination strategy calculates the spatial vector distance between the virtual camera viewpoint and the 3D component in real-time, dynamically assigns video memory weights, and actively unloads and removes polygon patches outside the field of view and occluded areas. The rendering optimization mechanism can be formally expressed as:

$$R_{\text{opt}} = R_{\text{raw}} \cdot I(v_{\text{cam}} \cdot n_{\text{face}} > \theta_{\text{cull}}) \tag{25}$$

In the formula, R_{opt} represents the optimized rendering load, R_{raw} is the original rendering load, v_{cam} is the camera line of sight vector, n_{face} is the polygon normal vector, θ_{cull} is the rejection threshold, and $I(\cdot)$ is the indicator function.

The rendering flow control logs from multiple terminal platforms confirm that even in extreme scenarios where thousands of concurrent retrieval tasks and complex spatial roaming interactions overlap, the system still exhibits excellent anti-crash robustness and availability boundaries. The dispersion degree of the response time box plot slightly increases with the increase of load and tends to stabilize, while the proportion of outliers is controlled, indicating that the system still maintains stable service quality under high concurrency conditions. The above results fully verify that the DT-KG deep fusion architecture can effectively support the engineering requirements of large-scale concurrent user access while ensuring real-time interactive experience.

E. VERIFICATION OF APPLICATION EFFICIENCY FOR SMART MANAGEMENT AND COMMUNICATION IN DUAL SCENARIOS

To verify the effectiveness of this system in breaking the disconnect between expert management and mass communication applications in the actual business chain, the underlying DT-KG fusion architecture is synchronously mounted on the professional operation and maintenance management flow and the public display terminal. On the intelligent protection management end, the system relies on continuous physical monitoring data from the Internet of Things and combines it with the knowledge graph ontology to perform dynamic graph inference. When the stress or environmental

temperature and humidity readings of a specific three-dimensional component deviate from the safety threshold, an alarm signal is generated through a preestablished cross modal mapping dictionary to instantly retrieve and highlight the associated semantic nodes

in the graph, automatically extracting the previous repair records, material composition, and aging evolution graph of the physical component. This mechanism utilizes data-driven automated correlation analysis to replace the traditional manual retrieval of dispersed archives, effectively supporting real-time disease warning and precise investigation and intervention.

On the public cultural dissemination side, the underlying logic of immersive interaction is directly taken over by the spatial semantic bidirectional mapping engine. When users trigger interactive operations on a specific twin grid in a 3D visualization viewport, the front-end collects spatial coordinates and converts them into feature vectors. Through a latent space alignment model, the corresponding historical allusions, morphological evolution, and related semantic clusters of cultural relics are recalled in real time. Then, dynamic viewpoint guidance and 3D information flow overlay technology are used for scene reconstruction. This mechanism of directly addressing high-dimensional semantic knowledge from spatial entities completely reshapes the linear display path of traditional static text and monolithic models, and constructs a network information feedback loop that can dynamically derive based on user perspectives.

To rigorously quantify the comprehensive application effectiveness of the above operating mechanism, a single digital twin platform and an independent graph retrieval system are introduced as comparison baselines in the evaluation process, and a multidimensional radar graph evaluation matrix is constructed. The assessment criteria of this matrix are strictly divided into five core dimensions: disease traceability efficiency, public interaction duration, knowledge acquisition richness, warning accuracy, and narrative coherence. Let the evaluation vector space be $V_{\text{eval}} = [E_{\text{trace}}, T_{\text{inter}}, R_{\text{know}}, A_{\text{alert}}, C_{\text{narr}}]$, where each component corresponds to the normalized scores of the five indicators mentioned above.

As shown in Figure 8, the integrated system achieves a warning accuracy of 96%, significantly higher than the 90% of a single digital twin system and the 68% of a single knowledge graph system; The efficiency of disease traceability has been recorded at 95%, far exceeding the 70% of a single knowledge graph system. In the dimension of mass communication, the public interaction duration and narrative coherence of the integrated system reach 90% and 91% respectively, effectively filling the gap of a single digital twin system in such semantic intensive indicators (65% and 55% respectively). In terms of knowledge acquisition richness, the integrated system scored 92%, overcoming the deficiency of information scarcity in a single digital twin system (only 60%). The above results confirm the core value of cross modal fusion mechanism in completely resolving "data silos" with clear engineering quantification results, indicating that the system not only approaches the theoretical judgment limit in terms of professional warning accuracy and disease traceability efficiency, but also demonstrates intergenerational advantages in terms of interaction duration and narrative coherence in the public dimension.

V. CONCLUSION

To bridge the semantic gap between the geometric physical data of industrial machinery and the historical records of manufacturing techniques, this paper proposes a DT-KG deep fusion model based on cross-modal ontology mapping. This framework integrates the high-fidelity fiber morphology and mechanical kinematics of industrial heritage into a comprehensive digital system, effectively reconciling the disconnection between smart protection management and public cultural dissemination. In terms of specific implementation steps, the system first overcomes the low resource text limitation by using a joint information extraction model with parameter sharing, and automatically extracts entity relationships from unstructured archives to construct a domain knowledge graph; At the same time, using multilayer perception and local feature perception mechanisms, high fidelity 3D point clouds are accurately parsed into independent semantic component sets with spatial bounding boxes. Subsequently, this article designed a core cross modal space semantic dynamic alignment algorithm, relying on the nonlinear projection and contrastive loss mechanism in the continuous latent space to forcibly optimize the distribution structure of heterogeneous vectors. Finally, the cosine similarity threshold was used to achieve precise binding and bidirectional data-driven between three-dimensional physical coordinates and historical knowledge nodes. The quantitative experiment and actual deployment results confirm the progressiveness and robustness of the deep fusion mechanism. In the core cross modal mapping task, the spatial semantic alignment algorithm effectively resolves the spatial discreteness caused by modal heterogeneity, and the mapping precision of entity component reaches 95.6%. In addition, under the concurrent stress test of intelligent management and dissemination dual scenarios, the average response time of cross modal semantic retrieval of the system was successfully compressed to within 150ms, and the real-time rendering efficiency of multi-terminal twin scenes remained stable at over 60 FPS. In summary, the system constructed in this article has completely broken the data silos at the bottom of cultural heritage. It not only supports real-time structural health warnings based on material degradation and fiber state transitions on the expert side, but also reshapes the immersive semantic interaction path driven by three-dimensional viewpoints on the public side, establishing a new engineering paradigm for the intelligent protection and deep-level digital inheritance of multimodal cultural heritage.

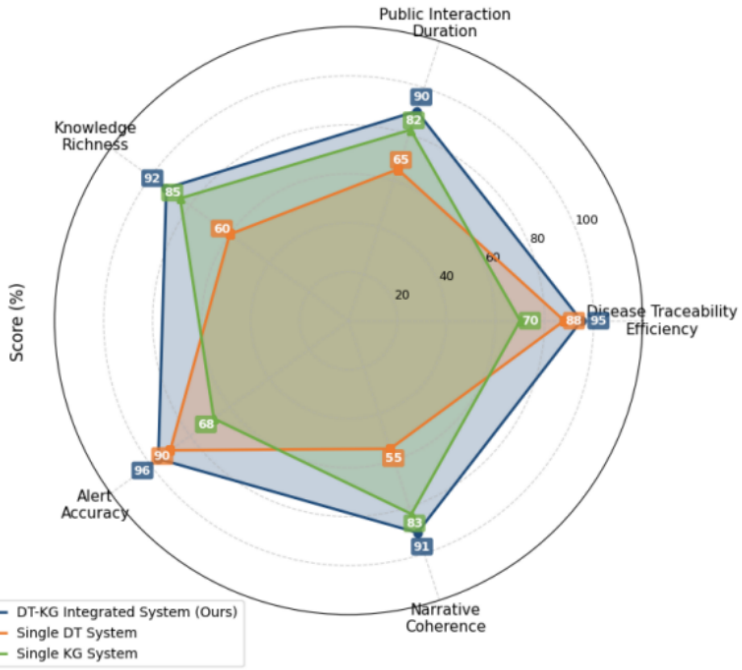

FIGURE 8. Radar diagram of comprehensive application effectiveness of DT-KG fusion system and single system

AUTHOR CONTRIBUTIONS

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

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] A. Felicetti, A. Himmiche, and M. Somenzi, "Knowledge Graphs and Artificial Intelligence for the Implementation of Cognitive Heritage Digital Twins," Applied Sciences, vol. 15, no. 18, pp. 1-34, 2025, doi: 10.3390/app151810061.

[2] Y. Huang Y, S. Yu S, J. Chu J, et al., "Using knowledge graphs and deep learning algorithms to enhance digital cultural heritage management," Heritage Science, vol. 11, no. 1, pp. 1-26, 2023, doi: 10.1186/s40494-023-01042-y.

[3] T. Fan, H. Wang, and T. Hodel, "CICHMKG: a large-scale and comprehensive Chinese intangible cultural heritage multimodal knowledge graph," Heritage Science, vol. 11, no. 1, pp. 1-18, 2023, doi: 10.1186/s40494-023-00927-2.

[4] Z. Li and H. Zhang, "The research progress on cultural heritage and digital humanities data, and its key technological pathways," Geographical Research Bulletin, vol. 4, no. 1, pp. 758-761, 2025.

[5] S. Battina, "Navigating geometric complexity in digital heritage: a review of AI-based semantic segmentation," Journal of Information Technology in Construction (ITcon), vol. 30, no. 70, pp. 1707-1727, 2025, doi: 10.36680/j.itcon.2025.070.

[6] M. Angheluta L, I. Popovici A, and C. Ratoiu L, "A Web-based platform for 3d visualization of multimodal imaging data in cultural heritage asset documentation," Heritage, vol. 6, no. 12, pp. 7381-7399, 2023, doi: 10.3390/heritage6120387.

[7] B. Ranjgar, A. Sadeghi-Niaraki, M. Shakeri, et al., "Cultural heritage information retrieval: past, present, and future trends," IEEE Access, vol. 12, no. 1, pp. 42992-43026, 2024, doi: 10.1109/ACCESS.2024.3374769.

[8] Z. Rao and G. Wang, "Smart Data-Enabled Conservation and Knowledge Generation for Architectural Heritage System," Buildings, vol. 15, no. 12, pp. 2122-2140, 2025, doi: 10.3390/buildings15122122.

[9] S. Yang, M. Hou, and S. Li, "Three-dimensional point cloud semantic segmentation for cultural heritage: a comprehensive review," Remote Sensing, vol. 15, no. 3, pp. 548-573, 2023, doi: 10.3390/rs15030548.

[10] I. Gagliardi and T. Artese M, "Exploring and visualizing multilingual cultural heritage data using multi-layer semantic graphs and transformers," Electronics, vol. 13, no. 18, pp. 3741-3768, 2024, doi: 10.3390/electronics13183741.

[11] S. Ferro, R. Giovanelli, M. Leeson, et al., "A novel NLP-driven approach for enriching artefact descriptions, provenance, and entities in cultural heritage," Neural Computing and Applications, vol. 37, no. 25, pp. 21275-21296, 2025, doi: 10.1007/ s00521-025-11449-2.

[12] E. Ntafotis, E. Zidianakis, N. Partarakis, et al., "Extraction of event-related information from text for the representation of cultural heritage," Heritage, vol. 5, no. 4, pp. 3374-3396, 2022, doi: 10.3390/heritage5040173.

[13] J. Li and A. Bikakis, "Towards a semantics-based recommendation system for cultural heritage collections," Applied Sciences, vol. 13, no. 15, pp. 8907-8931, 2023, doi: 10.3390/app13158907.

[14] A. Carriero V, A. Gangemi, L. Mancinelli M, et al., "Pattern-based design applied to cultural heritage knowledge graphs: ArCo: The knowledge graph of Italian Cultural Heritage," Semantic Web, vol. 12, no. 2, pp. 313-357, 2021, doi: 10.3233/ SW-200422.

[15] C. Li, S. Zhao, Y. Xuan, et al., "Metadata description framework for Chinese paper-cutting as traditional art in the context of intangible cultural heritage inheritance," Journal of Documentation, vol. 81, no. 3, pp. 657-683, 2025, doi: 10.1108/ JD-09-2024-0223.

[16] J. Siqueira and L. Martins D, "Workflow models for aggregating cultural heritage data on the web: A systematic literature review," Journal of the Association for Information Science and Technology, vol. 73, no. 2, pp. 204-224, 2022, doi: 10.1002/ asi.24498.
[17] L. Qin, S. Chen, J. Huang, et al., "Statistical system of cultural heritage tourism information based on image feature extraction technology," Mathematical Problems in Engineering, vol. 2022, no. 1, pp. 1-12, 2022, doi: 10.1155/2022/5250853.
[18] T. Li, "A digital cultural heritage tourism information statistics system based on image feature extraction technology," Computer-Aided Design and Applications, vol. 21, no. S16, pp. 1-18, 2024, doi: 10.14733/cadaps.2024.S16.1-18.
[19] Y. Zhang, Y. Wang, Z. Guo, et al., "A Knowledge Graph-Guided and Multimodal Data Fusion-Driven Rapid Modeling Method for Digital Twin Scenes: A Case Study of Bridge Tower Construction," ISPRS International Journal of Geo-Information, vol. 15, no. 1, pp. 27-47, 2026, doi: 10.3390/ijgi15010027.
[20] K. Mirzaei, M. Arashpour, E. Asadi, et al., "Automatic generation of structural geometric digital twins from point clouds," Scientific reports, vol. 12, no. 1, pp. 22321-22336, 2022, doi: 10.1038/s41598-022-26307-7.
[21] X. Pan, Q. Lin, S. Ye, et al., "Deep learning based approaches from semantic point clouds to semantic BIM models for heritage digital twin," Heritage Science, vol. 12, no. 1, pp. 1-17, 2024, doi: 10.1186/s40494-024-01179-4.
[22] A. Basu, S. Paul, S. Ghosh, et al., "Digital restoration of cultural heritage with data-driven computing: A survey," IEEE access, vol. 11, no. 1, pp. 53939-53977, 2023, doi: 10.1109/ACCESS.2023.3280639.
[23] J. Yin, "Application of intelligent image recognition and digital media art in the inheritance of black pottery intangible cultural heritage," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 6, pp. 1-15, 2024, doi: 10.1145/3597430.
[24] Q. Cheng, Y. Zhou, P. Fu, et al., "A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, no. 1, pp. 4284-4297, 2021, doi: 10.1109/JSTARS.2021.3070872.
[25] M. Hou, Z. Zhang, C. Liu, et al., "Semantic alignment network for multi-modal emotion recognition," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 9, pp. 5318-5329, 2023, doi: 10.1109/TCSVT.2023.3247822.
[26] M. Pal, M. Kumar, R. Peri, et al., "Meta-learning with latent space clustering in generative adversarial network for speaker diarization," IEEE/ACM transactions on audio, speech, and language processing, vol. 29, no. 1, pp. 1204-1219, 2021, doi: 10.1109/TASLP.2021.3061885.
[27] M. Moradi, P. Moradi, A. Faroughi, et al., "Two-level attention mechanism with contrastive learning for heterogeneous graph representation learning," Expert Systems with Applications, vol. 273, no. 1, pp. 1-15, 2025, doi: 10.1016/ j.eswa.2025.126751.
[28] D.-L. Li, S.-K. Lee, and Y.-T. Liu, "Printed document layout analysis and optical character recognition system based on deep learning," Scientific Reports, vol. 15, no. 1, pp. 1-15, 2025, doi: 10.1038/s41598-025-07439-y.
[29] A. Fateh, M. Fateh, and V. Abolghasemi, "Enhancing optical character recognition: Efficient techniques for document layout analysis and text line detection," Engineering Reports, vol. 6, no. 9, Art. no. e12832-e12858, 2024, doi: 10.1002/eng2.12832.
[30] L. Gu, H. Zhang, and X. Wu, "Surveying and mapping of large-scale 3D digital topographic map based on oblique photography technology," Journal of Radiation Research and Applied Sciences, vol. 17, no. 1, pp. 1-10, 2024, doi: 10.1016/ j.jrras.2023.100772.
[31] T. Zhou, L. Lv, J. Liu, et al., "Application of UAV oblique photography in real scene 3d modeling," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 43, no. 1, pp. 413-418, 2021, doi: 10.5194/isprs-archives-XLIII-B2-2021-413-2021.
[32] H. Ikeuchi and H. Saito, "Shape estimation using location-unknown distance sensors: Iterative-closest-point-based approach," IEEE Sensors Journal, vol. 22, no. 11, pp. 10740-10750, 2022, doi: 10.1109/JSEN.2022.3168002.
[33] W. Yuan, C. Liu, T. Wang, et al., "Multi-source 3D point clouds fusion for potential rock mass hazard evaluation in highsteep rock slopes," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 48, no. 1, pp. 1663-1668, 2025, doi: 10.5194/isprs-archives-XLVIII-G-2025-1663-2025.
[34] Q. Shen, Y. Liu, R. Chen, et al., "The atmospheric vertical detection of large area regions based on interference signal denoising of weighted adaptive Kalman filter," Sensors, vol. 22, no. 22, pp. 8724-8738, 2022, doi: 10.3390/s22228724.
[35] Y. Su, J. Han, C. Ma, et al., "Adaptive Kalman filter based on online ARW estimation for compensating low-frequency error of MHD ARS," IEEE Transactions on Instrumentation and Measurement, vol. 73, no. 1, pp. 1-10, 2024, doi: 10.1109/ TIM.2024.3375962.