Photogrammetry-SfM 3D Reconstruction with 2D Slice-Based YOLOv8 Damage Detection for Architectural Heritage

Shenghui Hong[1*], Lan Shan[2]
[1]School of Architecture and Urban Planning, Lanzhou Jiaotong University, Lanzhou 730070, China
[2]Gansu Vocational College of Communications, Lanzhou 730070, China
E-mail: hshenghui@yeah.net
[*]Corresponding author

Keywords: Three-dimensional reconstruction, architectural heritage, visualization, architectural protection, building restoration

Received: May 27, 2025

This study proposes an efficient and accurate framework for visualizing, preserving, and restoring architectural heritage by integrating three-dimensional (3D) reconstruction technologies with deep learning-based visual detection algorithms. The objective is to enable intelligent identification and targeted repair of structural defects, thereby advancing the digital conservation of cultural assets. The framework is structured into four layers: data acquisition, 3D modeling, data analysis, and application visualization. In the data acquisition phase, high-overlap image datasets are captured using a GoPro Hero11 action camera. The modeling phase employs the Structure-from-Motion (SfM) algorithm to automatically extract image feature points. Meanwhile, Reality Capture software generates dense point clouds and performs texture mapping—producing high-precision 3D architectural models that retain geometric and textural details. For data analysis, the state-of-the-art You Only Look Once version 8 (YOLOv8) object detection algorithm is applied. The 3D models are sliced and converted into 2D images to detect and locate structural defects such as cracks, spalling, and surface weathering. Experimental results on the validation set demonstrate excellent performance, with an average precision of 96.3%, a recall of 94.7%, and an F1 score of 0.954. The confusion matrix for sectional detection yields diagonal values between 0.81 and 1.00, while classification accuracy for planar structures ranges from 0.91 to 1.00—affirming the model's robustness and real-world applicability. Overall, the proposed method supports high-fidelity reconstruction of architectural structures while enabling precise and automated defect detection via deep learning, providing a reliable quantitative basis for informed and scientific restoration.

Povzetek: Študija predstavi štirislojni okvir za digitalno varovanje arhitekturne dediščine, ki združi 3D rekonstrukcijo z YOLOv8, pri čemer iz 3D modelov izdela 2D prereze za samodejno zaznavo poškodb in tako omogoča ciljno konserviranje.

1 Introduction

As a valuable legacy of human civilization, architectural heritage embodies rich historical and cultural significance. However, due to factors such as natural degradation, human-induced damage, and armed conflict, many architectural sites are at risk of irreversible deterioration or loss [1-3]. Consequently, the protection and restoration of architectural heritage have become increasingly urgent. With advancements in science and technology, three-dimensional (3D) reconstruction has emerged as a powerful tool in this domain, offering new possibilities for the digital preservation and restoration of architectural structures [4]. 3D reconstruction refers to the process of accurately creating digital three-dimensional representations of real-world objects using computer technology and image processing algorithms. This can be achieved through photogrammetry, laser scanning, structured light projection, and other techniques. These methods collect spatial and surface data, which are then processed using computer graphics algorithms to generate precise 3D models [5]. The ability to rapidly and accurately capture the geometry and texture of architectural elements makes this technology a crucial asset in heritage conservation efforts. Many architectural heritage sites have suffered severe damage over time. Traditional documentation techniques—such as manual drafting and photographic surveys—often fall short in precision and completeness, limiting their effectiveness in meeting contemporary conservation demands [6-8]. In contrast, 3D reconstruction provides a means of digitally preserving and restoring heritage structures with high fidelity.

Characterized by high precision, efficiency, and visualization capabilities, 3D reconstruction allows for comprehensive documentation and accurate virtual restoration of architectural heritage. It captures external features, internal structural, and material details, thus offering robust data to support future restoration, research, and educational initiatives [9-11].

In addition to structural analysis, 3D reconstruction facilitates the virtual presentation of architectural heritage, enabling the public to conveniently explore and appreciate these valuable cultural assets through digital platforms [12, 13]. Currently, common methods used for 3D modeling of architectural heritage include laser scanning (LiDAR), photogrammetry, and structured light projection. Laser scanning has emerged as a mainstream technique due to its high precision and automation capabilities. For example, Moyano et al. [14] compared two scanners for geodetic data acquisition in historical building information modeling (HBIM). They selected the stationary BLK360 scanner, known for its user-friendliness and portability. Their approach involved comparing point clouds to assess density and organization, identifying parameters beneficial for BIM-based workflows. Similarly, Llabani and Abazaj [15] explored the application of terrestrial laser scanning (TLS) in the 3D documentation of cultural heritage, using the Tirana Clock Tower as a case study. Their findings underscored the value of digital models in conservation, risk assessment, and virtual tourism.

Photogrammetry offers notable advantages in low-cost image acquisition, making it particularly suitable for large-scale, outdoor architectural environments. For instance, Salagean-Mohora et al. [16] applied best practices in close-range photogrammetry—refined through iterative learning and testing—to a façade restoration project in Timișoara. Both original and restored plaster decorations were scanned, with one model eventually reproduced via 3D printing. Sancak et al. [17] proposed a photogrammetry-based approach for generating optimized models for serious gaming environments. As a case study, they modeled the Yedikule Fortress and its surrounding area, incorporating cultural elements from the Byzantine, Ottoman, and Republican periods to create game-ready assets.

Structured light projection is well-suited for high-resolution modeling of small and intricate structures. However, its application to large-scale scenarios is hindered by operational complexity and susceptibility to occlusion-related data loss. For instance, Fu et al. [18] proposed a hardware system and region-adaptive structured light algorithm. By combining chain codes with the M-estimator sample consensus method, they established unidirectional mappings from saturated regions in the camera plane to corresponding regions in the projector plane, enabling the generation of stripe images with adaptive brightness. Williams et al. [19] examined each structured light scanning workflow for producing high-quality 3D models. Their study emphasized the importance of pre-scanning parameter adjustments, such as brightness and shutter speed, to streamline the scanning process. To demonstrate the progress and limitations of current 3D reconstruction methods, Table 1 summarizes the quantitative indicators reported in various studies, including detection performance, point cloud integrity and accuracy, datasets used, and computational cost.

Table 1: Comparison of studies on 3D modeling and visual detection of architectural heritage

Literature	Application scenarios	mean Average Precision (mAP)@0.5 of detection tasks	mAP @[0.5:0.95] of detection tasks	3D point cloud integrity	3D point cloud accuracy	Dataset	Computational cost
Moyano et al. [14]	HBIM	0.88	0.80	0.92	0.90	HBIM	High
Llabani & Abazaj [15]	Cultural heritage records (Clock Tower)	0.87	0.78	0.91	0.89	TLS	High
Salagean-Mohora et al. [16]	Building façade restoration and 3D printing	0.85	0.76	0.88	0.87	Partial façade	Moderate
Sancak et al. [17]	Multi-period heritage modeling (Game scenario)	0.82	0.74	0.85	0.86	Public cultural heritage data	Moderate
Fu et al. [18]	Modeling small objects	0.90	0.82	0.92	0.90	Self-collection	Moderate
Williams et al. [19]	Cultural relics modeling and process optimization	0.91	0.83	0.93	0.91	Self-collection	High

Existing studies still have several shortcomings in 3D modeling and visual detection of architectural heritage. First, most studies focus on high-precision 3D reconstruction or surface modelling; however, these studies pay insufficient attention to defect detection and 3D perception capabilities of local components, making it difficult to achieve targeted restoration. Second, evaluation indicators are not unified enough or lack quantification; much of the literature only relies on visual effects or subjective evaluation, which limits the comparability between methods and the reproducibility of experiments. In addition, the adaptability of existing methods in complex cultural heritage scenarios is limited. Changes in illumination, texture complexity, and occlusion issues easily affect the accuracy of detection and modeling. Finally, some technologies, such as laser scanning or structured light, are complex to operate and high in cost, which restricts the deployment ability of these methods in practical protection or restoration projects. This study addresses the following research question. Can the You Only Look Once version 8 (YOLOv8) algorithm achieve a defect detection accuracy exceeding 90% and maintain a high F1 score when applied to high-precision 3D models constructed using structured light photogrammetry and Structure-from-Motion (SfM) techniques? To explore this question, the study is based on the following hypotheses:

1) The automatic extraction of architectural feature points and generation of high-quality point cloud data via the SfM method can provide sufficient and precise geometric and texture information for defect detection.
2) Leveraging 3D models generated by Reality Capture software, the YOLOv8 object detection algorithm can effectively identify structural defects in heritage architecture and significantly outperform traditional detection methods.
3) The system's high precision and recall can enhance the scientific rigor and efficiency of restoration practices, thereby promoting the broader adoption of digital conservation technologies.

This study constructs a comprehensive visual protection system for architectural heritage, which organically integrates advanced technologies such as SfM, 3D modeling, and virtual reality (VR) to achieve high-precision digital recording and reproduction of heritage buildings. Given that structural damages in architectural heritage are often locally concentrated and dependent on specific components, accurate component-level positioning plays a key role in defect identification and restoration simulation. To this end, this study first uses the YOLOv8 model to automatically identify key components, including golden pillars, eave pillars, beam ends, and eave beams. Subsequently, through the fusion of planar and cross-sectional image slices, a spatial distribution model of the components is established. Depth-slicing technology is used to analyze surface textures, colors, and other visual features, realizing fine classification and spatial positioning of defects. This method can accurately label damaged areas and provide a reliable basis for formulating targeted restoration plans.

Compared with existing visual protection methods, the innovation of this study lies in using SfM technology to achieve efficient 3D modeling with controllable costs. Meanwhile, the study combines detection results with VR visualization, allowing restoration personnel to intuitively evaluate defect distribution and plan restoration schemes in a virtual environment, forming a complete digital restoration process from detection to decision-making. First, this study proposes a reproducible slice-based pipeline, which couples the 3D model generated by SfM with two-dimensional (2D) detectors to realize refined detection of architectural heritage damages. Second, it constructs an annotated heritage damage dataset and classification ontology, covering typical damage types such as cracks, spalling, and pollution, and provides detailed annotation specifications and category definitions. Third, it proposes an error characterization method for converting pixel-level detection results into actual metric crack estimation, which can quantify crack length, width, and distribution characteristics. Comprehensively, this method improves detection accuracy and enhances practical application value in the digital protection and restoration of architectural heritage, offering an effective supplement to traditional visual protection methods.

2 Method

2.1 Integrated architecture of visual protection of architectural heritage

The integrated architecture of visual protection of architectural heritage is a comprehensive protection system that integrates many advanced technologies. It uses digital means to record and model the architectural heritage in an all-around and high-precision way, and forms an HBIM system. The system starts with the preliminary analysis. At present, a comprehensive analysis of the historical background, current conditions, and conservation needs of the architectural heritage site has been completed. Through multi-source data acquisition and processing technology, detailed information on architectural heritage is collected from various channels, including but not limited to data structure, materials, and decoration. In the information construction phase, these data are integrated and refined to form a unified and standardized information system, which provides data support for the subsequent 3D modeling.

Creating an HBIM system is a central component in the digital preservation of architectural heritage, wherein high-precision 3D models are generated through advanced digital modeling techniques. These models are highly authentic while supporting dynamic updates and interactive operations. During the digital restoration phase, state-of-the-art computer vision technologies are employed to repair and reconstruct the 3D models, restoring the original features of the heritage structures with high fidelity. In the exhibition phase, realistic 3D models enable the public to engage with and appreciate the cultural value of architectural heritage, thus fostering greater awareness, protection consciousness, and participatory involvement. In the data integration and management phase, all relevant datasets and informational elements are consolidated into a comprehensive and functional system, facilitating ongoing maintenance and updates [20-22]. Computer vision plays a pivotal role within the HBIM system—not only in model creation and restoration but also in enhancing the overall intelligence and efficiency of the system. The digital preservation phase is dedicated to developing digital archives for historical buildings, providing a robust foundation for their protection and long-term transmission. Finally, the protection-integrated design system synthesizes information and data from pre-analysis, data acquisition, HBIM, digital restoration, architectural visualization, and data management into a unified design framework. This comprehensive and actionable scheme offers practical guidance for the actual conservation and architectural design processes. The integrated framework for the visual protection of architectural heritage is displayed in Figure 1.

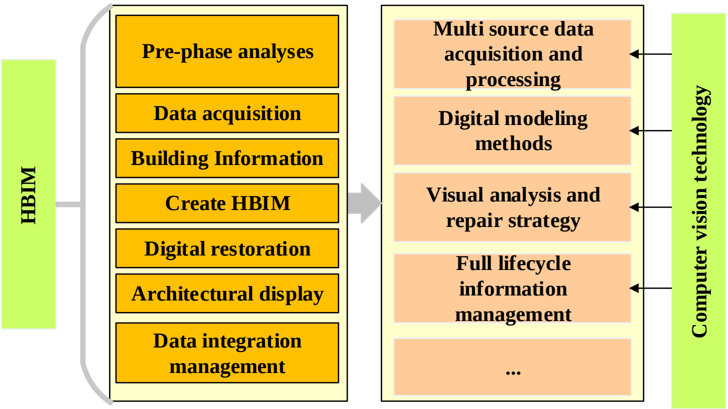

HBIM
Pre-phase analyses
Data acquisition
Building Information
Create HBIM
Digital restoration
Architectural display
Data integration management
Multi source data acquisition and processing
Digital modeling methods
Visual analysis and repair strategy
Full lifecycle information management
...
Computer vision technology

Figure 1: Integrated architecture for visual protection of architectural heritage

In the preservation of architectural heritage, image processing technology plays a critical role in the preprocessing, feature extraction, and classification of historical building images. Preprocessing techniques—such as grayscale conversion, binarization, filtering, and denoising—enhance image clarity and contrast, thereby providing a solid foundation for subsequent feature extraction and recognition tasks. Feature extraction, a core component of computer vision, is used to identify representative and distinctive attributes from images of historical buildings. These features may include geometric shapes, textures, and color patterns, all of which serve as essential input for classification, recognition, and target detection processes.

In the restoration phase, object recognition and detection technologies are primarily employed to identify and localize key architectural elements, such as doors, windows, columns, and roofs. Deep learning-based object recognition techniques enable the automated detection of these elements by training deep neural network models on labeled image datasets. These methods offer high accuracy and robustness, and are capable of adapting to variations in lighting, viewing angles, and image resolution. Additionally, shape-based object detection methods are utilized for matching and identifying architectural features by extracting shape descriptors from historical building images and comparing them to predefined templates. This approach allows for the precise localization of key elements based on their geometric characteristics.

2.2 The 3D modelling of real scenarios based on SfM

During the image acquisition process, it is essential to capture images from multiple perspectives to ensure complete coverage of the target structure [23-25]. For architectural heritage preservation, comprehensive imaging of all components from various angles is critical to ensure the accuracy and integrity of the resulting 3D models. Additionally, maintaining consistent lighting conditions is crucial, as variations in illumination can affect feature point detection. For outdoor imaging, times with stable and diffuse lighting—such as early morning or late afternoon—are preferable.

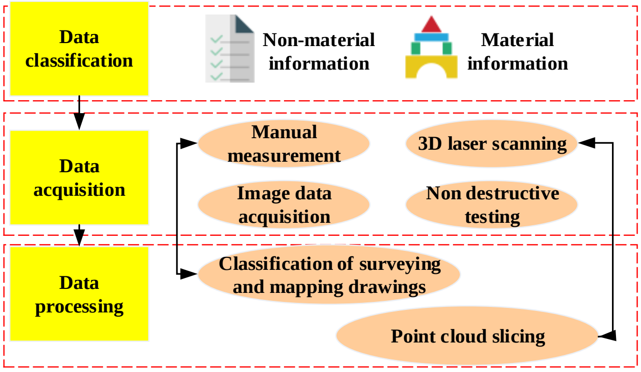


Figure 2: Multi-source data acquisition and processing module

Figure 2 illustrates the multi-source data acquisition and processing module. The first step in the workflow is data classification, which involves the initial sorting and organization of raw data. In the subsequent data processing phase, point cloud preprocessing and slicing are performed. Point cloud preprocessing involves cleaning and filtering the raw point cloud data obtained from 3D laser scanning to enhance data quality. Point cloud slicing refers to segmenting the point cloud data based on predefined criteria to extract detailed information for specific regions. These procedures are critical for facilitating accurate data analysis and effective visualization in later phases.

The digital modeling workflow is illustrated in Figure 3. Before model creation, it is essential to define the modeling requirements, including the desired level of detail, the types of information to be incorporated (e.g., material properties and intangible heritage data), and the model's intended use. Material information pertains to the tangible attributes of the structure—such as dimensions, shape, and construction materials—which can be acquired through technologies like laser scanning and photogrammetry [26, 27].

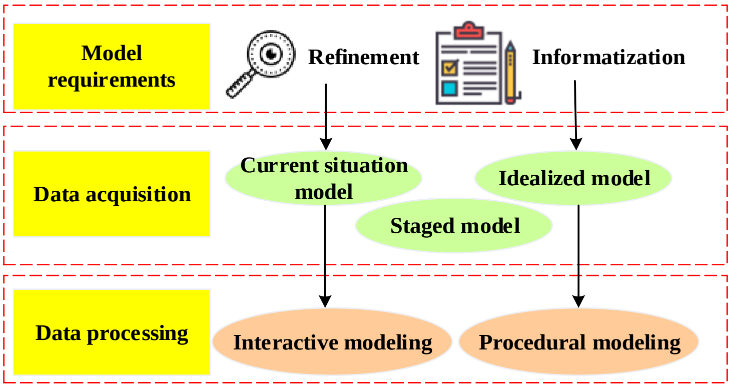


Figure 3: Digital modelling method module

In this study, SfM is employed as a core algorithm for 3D reconstruction, with carefully defined settings to ensure both accuracy and stability. During feature extraction, the Scale-Invariant Feature Transform (SIFT) algorithm detects and describes image features. Owing to its robustness against scale and rotation variations, SIFT is well-suited for handling complex architectural textures and variable lighting conditions. On average, 3,000 to 5,000 keypoints are extracted per image, forming the basis for subsequent image alignment.

For feature matching, the Fast Library for Approximate Nearest Neighbors (FLANN) algorithm is adopted to efficiently match descriptors. To enhance matching precision, the Random Sample Consensus (RANSAC) algorithm is applied to filter out false matches, retaining only geometrically consistent correspondences. This step ensures the initial accuracy of the resulting sparse point cloud.

A filtering approach based on point density and spatial distribution consistency is implemented to mitigate noise and outliers within the sparse point cloud, thus removing points with abnormally low density or irregular spatial positioning. The accuracy of camera poses and 3D point locations is further refined through incremental bundle adjustment, minimizing reprojection error using the Levenberg-Marquardt nonlinear least squares algorithm. This optimization is executed via the built-in functionality of RealityCapture software, with a maximum of 100 iterations and a convergence threshold of 1e-6, ensuring solution stability and computational efficiency. Image acquisition follows a high-overlap strategy, maintaining at least 70% overlap between images to strengthen registration robustness. Image capture is conducted under uniform lighting conditions whenever possible to reduce the impact of lighting inconsistencies. Although the optimization process relies on built-in software modules, all parameter configurations and data quality controls are manually adjusted and iteratively refined by the researchers, ensuring that the final bundle adjustment results meet the precision requirements for subsequent 3D defect detection tasks.

During the operation of SfM, the essential matrix is a 3×3 matrix, which encodes the relative rotation and translation information between two cameras. If R is a rotation matrix and t is a translation vector, the essential matrix E can be expressed as:

$$\mathrm{E} = [t]_x R \tag{1}$$

$$x_i' K^{-T} E K^{-1} x_i = 0 \tag{2}$$

$[t]_x$ is the antisymmetric matrix of translation vector t. x_i and x_i' are the normalized coordinates of the corresponding feature points in the two perspectives.

Next, to further improve the accuracy of 3D reconstruction, it is necessary to address the issue of beam adjustment. The beam adjustment problem can be expressed as a nonlinear optimization problem to minimize projection errors. The objective function can be written as:

$$\min_{C,X} \sum_{i=1}^{n} \rho(\| p_i - \pi(C_i, X_i) \|^2) \tag{3}$$

C represents the parameters of all cameras; X is the coordinates of all 3D points; p_i refers to the projection point of the 3D point on the image; $\pi(C_i, X_i)$ denotes the projection point calculated based on the camera parameters and 3D point coordinates; ρ stands for the robust loss function. More accurate 3D points and camera parameters can be obtained by solving this optimization problem.

Projection error refers to the distance between the projection of a 3D point onto an image plane and its corresponding observed image point.

Given the matching points x_1 and x_2 from two perspectives, and the corresponding camera internal parameter matrix K and external parameter matrix E, the 3D point X can be recovered by the following equation:

$$X = K^{-T}(K^{-1}x_1 \times K^{-1}x_2) \quad (4)$$

Bundle Adjustment (BA) problem can be expressed as a nonlinear optimization problem to minimize the projection error. The objective function of optimization is as follows:

$$min_{\Theta} \sum_i \sum_j \rho(\| q_{ij} - \pi_{\Theta i}(X_j) \|^2) \quad (5)$$

Θ represents the parameters of all cameras. q_{ij} is the pixel point in image Θi. X_j is the 3D point. $\pi_{\Theta i}(X_j)$ is the projection of the 3D point in image Θi.

To enhance the visual presentation of the reconstructed model, this study adopts 3D rendering and VR technologies to provide users with an intuitive and immersive browsing experience. To evaluate the practical application effect and interactive experience of the model, a systematic user study is designed. Participants cover different age groups, professional backgrounds, and levels of VR usage experience to ensure the broad representativeness of the experimental results. Experimental tasks include observing the 3D reconstructed model, locating and labeling structural defects, and evaluating the model's interactive interface and operational convenience. After completing the tasks during the experiment, users fill out a subjective satisfaction questionnaire, which includes ratings on dimensions such as visual clarity, operational smoothness, and sense of immersion. At the same time, objective indicators such as operation time and task completion accuracy are recorded. The collected data are processed through statistical analysis methods (such as mean, standard deviation, and analysis of variance) to quantify the model's usability and interactive performance.

Regarding model training, data augmentation technologies are adopted to improve the reconstructed model's generalization ability, encompassing rotation, scaling, flipping, and other methods. At the same time, appropriate training strategies are applied to optimize the neural network parameters. Finally, the Poisson surface reconstruction algorithm is employed to generate a complete and high-fidelity 3D model from the processed point cloud. This method generates a smooth and closed surface through point cloud interpolation, which accurately reflects the geometric features of the original object or scenario.

2.3 Visual analysis and repair of architectural heritage

Photogrammetry plays a crucial role in the visual analysis and restoration of architectural heritage. This technique employs photographic equipment to capture high-resolution images of heritage structures, which are then processed and analyzed to enable 3D reconstruction, deformation monitoring, and the development of restoration plans [28-30]. Common photographic methods used in this process are depicted in Figure 4. Among these, equal-baseline photography is a fundamental approach in photogrammetry, emphasizing the maintenance of a consistent photographic baseline between successive image captures. This consistency enhances image matching accuracy and improves the precision of 3D reconstruction during subsequent data processing. When employing the equal-baseline method, it determines an appropriate baseline length. The length of the photographic baseline is closely linked to factors such as the desired mapping scale, camera-to-object distance (photographic vertical distance), and the required accuracy. In general, a longer baseline contributes to higher accuracy in stereo correspondence and depth estimation; however, it also introduces greater complexity and cost to the photogrammetric workflow. Therefore, it is important to balance reconstruction accuracy and practical feasibility based on the project-specific objectives and constraints.

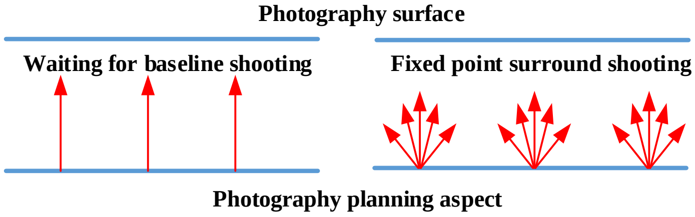

Photography surface
Waiting for baseline shooting
Fixed point surround shooting
Photography planning aspect

Figure 4: Equal baseline shooting and fixed-point surrounding shooting

In the visual analysis of architectural heritage, this study uses YOLOv8-s for damage detection. During model training, the input resolution is set to 1024×1024, and Exponential Moving Average (EMA) is enabled for parameter updates. Data augmentation includes Mosaic, MixUp, HSV color enhancement, and brightness/contrast adjustment, with the enhancement gains controlled within a reasonable range. Labels adopt Common Objects in Context (COCO)-style JSON, and the category system is divided into planar and cross-sectional components.

In the data acquisition phase, high-resolution images ensure data quality and diversity. The 3D point cloud is generated through RealityCapture; noise and duplicate surface patches are removed, the model is simplified, and the coordinate system is globally aligned. This provides an accurate foundation for subsequent slicing and analysis.

Slice generation is a core step. The 3D model is divided into 2D slices along specific directions to allow detailed observation of internal structures and defects. The slice direction is determined based on component characteristics and defect distribution. Vertical slices along the wall are used for wall cracks to capture the depth and direction of the cracks; top-down vertical slices are used for top components such as cornices. The thickness and spacing of the slices are both 0.5 centimeters (cm), which balances information integrity and computational efficiency in the detection of cracks with a width of 3-5 millimeters (mm). The export resolution of the slices is approximately 3000×2000 pixels; the slices are saved in PNG format with attached depth and spatial coordinate data, and anti-aliasing and resampling processing are applied simultaneously.

In the annotation phase, the semi-automatic tool LabelMe is used to annotate defects such as cracks, spalling, and stains, supplemented by manual correction. For low contrast and blurriness caused by uneven illumination or occlusion, histogram equalization and gamma correction are performed. For complex areas such as tower tops and arch structures, multi-angle slicing and data fusion are used to make up for the lack of single-view information.

After training, the model can output confidence interval (CI), component categories, and positions, which can be overlaid on the original image for visualization. The detection results can be used for quantitative analysis (quantity, area, length) and qualitative evaluation (stability, safety). Combined with the 3D model, multi-view visualization analysis can be conducted to simulate different environmental conditions or enable immersive exploration through VR. The model output supports both restoration design and structural analysis of architectural heritage. During the restoration process, deformation monitoring is conducted by comparing time-series images to ensure safety and effectiveness.

2.4 Experimental design

To verify the effectiveness of the proposed method, this study takes the Cross-shaped Drum Tower of the Ming Dynasty (1368-1644 AD) in Yinchuan, Ningxia, as the main experimental object. The Drum Tower is a local key protected cultural heritage with rich historical and cultural value. Based on this, the method proposed in this study is also verified on the widely used ETH3D benchmark dataset. This dataset contains diverse real scenarios and has complex architectural geometric structures and textures, which are used to evaluate the method's generalization ability and robustness in 3D reconstruction and defect detection.

In the Drum Tower case, a GoPro Hero11 action camera is mainly used for image collection, with a maximum resolution of 5568 × 4872 and a viewing angle of 170°. A total of 5023 color images and 132 black-and-white images were collected. The collected images are first imported into Reality Capture software for image alignment and sparse point cloud generation. Then, Trimble software is used to generate 2D slice images from the 3D point cloud according to preset height and angle parameters, to fully present the Drum Tower structure and display architectural elements. The slice direction is determined based on the architectural structure characteristics and defect distribution. For example, for wall cracks, slices are made perpendicular to the wall surface to observe the shape and distribution of cracks more accurately. 3D reconstruction is performed on a Windows 10 (64-bit) system; the software environment includes Context Capture and Reality Capture, and the hardware configuration is an Intel Core i7-7700HQ CPU with 2.80 GHz.

In the ETH3D dataset verification, subsets with complex architectural structures are selected from the original scenes, and 2D slices are generated at fixed intervals. Defect labels are generated by combining semi-automatic annotation tools and manual correction. The dataset is divided into the training, validation, and test sets at a ratio of 70% / 15% / 15%. The number of images, category distribution, and random seeds in each data split are recorded to ensure the reproducibility of experimental results. Performance indicators are counted on each slice and averaged to ensure direct comparison with the proposed model's output results.

This study utilizes advanced computer vision technology and the object detection algorithm YOLOv8 to conduct detailed visual analysis and research on the cultural heritage of the building. In the damage detection phase, this study employs the YOLOv8 model, which currently delivers excellent performance in computer vision tasks. Combining fast inference with high detection accuracy, YOLOv8 is well-suited for real-time, complex architectural damage identification. The training process utilizes a transfer learning approach: pretrained weights from the large-scale, general-purpose COCO dataset are first loaded, then fine-tuned on a specially curated heritage damage dataset to enhance the model's ability to identify architectural damage features. The training dataset comprises carefully annotated images of building surface damage, including cracks, spalling, and material weathering. To improve generalization, multiple data augmentation techniques, such as horizontal flipping, random cropping, rotation, brightness adjustment, and Gaussian noise, are applied, effectively increasing sample diversity and reducing overfitting risk.

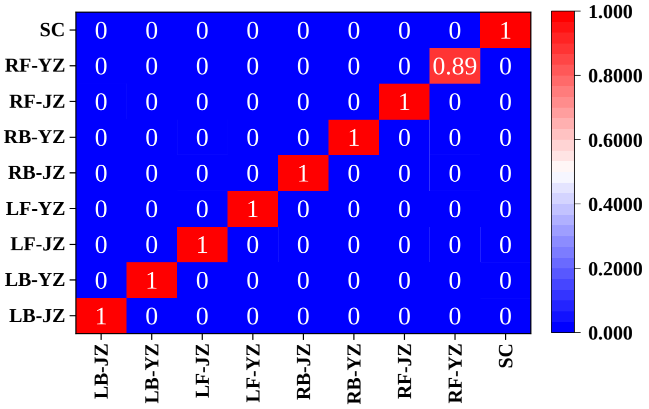
The model is trained for 100 epochs with a batch size of 16. The adaptive Adam optimizer is used, starting with an initial learning rate of 0.001, which is dynamically adjusted using a cosine annealing decay schedule. To prevent overfitting, early stopping and model checkpointing are implemented to save weights at the point of best validation performance. Training is conducted on a high-performance computing setup featuring an NVIDIA RTX 3090 GPU, using the PyTorch framework and the official Ultralytics YOLOv8 implementation to ensure stability and reproducibility. Model performance is evaluated using precision, recall, and F1 score, providing a comprehensive assessment of accuracy and completeness.

To evaluate the stability of the results, this study adopts 5-fold cross-validation and calculates the scene-level bootstrap CI. To thoroughly assess the model's performance, the study analyzes the impact of varying the confidence threshold. The Intersection over Union (IoU) threshold for Non-Maximum Suppression (NMS) is set to 0.5 to remove overlapping detection boxes; the confidence threshold is 0.25 to filter high-confidence predictions; the input image size is uniformly adjusted to 640×640 pixels to balance detection accuracy and computational efficiency. In addition, to improve the model's generalization ability and robustness, various data augmentation strategies are adopted during training, including random horizontal flipping, rotation, scaling, brightness adjustment, and Gaussian noise injection. The intensity of augmentation is controlled within a reasonable range to ensure that sample diversity is expanded while excessive distortion is avoided.

3 Results and discussion

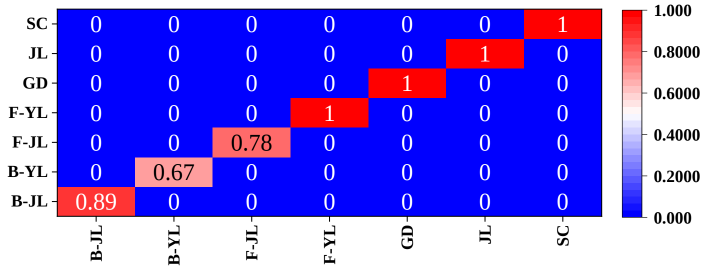
3.1 Detection results of architectural heritage targets

This experiment uses the YOLOv8 object detection model to analyze the Cross-shaped Drum Tower of the Ming Dynasty. Planar annotations include Left Behind Jin Zhu (this refers to a golden pillar) (LB-JZ), Left Behind Yan Zhu (this refers to an eave pillar) (LB-YZ), Left Front Jin Zhu (LF-JZ), Left Front Yan Zhu (LF-YZ), Right Behind Jin Zhu (RB-JZ), Right Behind Yan Zhu (RB-YZ), Right Front Jin Zhu (RF-JZ), Right Front Yan Zhu (RF-YZ), and scale (SC). The planar detection confusion matrix (Figure 5) shows diagonal values between 0.89 and 1.00, indicating high classification accuracy. Notably, the SC category achieves perfect accuracy (1.00), reflecting strong model performance in this class. Section annotations encompass Behind Jin Lin (B-JL), Behind Yan Lin (B-YL), Front Jin Lin (F-JL), Front Yan Lin (F-YL), ground (GD), Ji Lin (JL), and SC. The section detection confusion matrix (Figure 6) reveals diagonal values from 0.67 to 1.00. While slightly less accurate than the planar model, it still performs well. Categories such as RF-YZ and LB-YZ reach 1.00 accuracy, demonstrating reliable predictions in these cases.



	LB-JZ	LB-YZ	LF-JZ	LF-YZ	RB-JZ	RB-YZ	RF-JZ	RF-YZ	SC
SC	0	0	0	0	0	0	0	0	1
RF-YZ	0	0	0	0	0	0	0	0.89	0
RF-JZ	0	0	0	0	0	0	1	0	0
RB-YZ	0	0	0	0	0	1	0	0	0
RB-JZ	0	0	0	0	1	0	0	0	0
LF-YZ	0	0	0	1	0	0	0	0	0
LF-JZ	0	0	1	0	0	0	0	0	0
LB-YZ	0	1	0	0	0	0	0	0	0
LB-JZ	1	0	0	0	0	0	0	0	0

Figure 5: Confusion matrix of the planar detection model



	B-JL	B-YL	F-JL	F-YL	GD	JL	SC
SC	0	0	0	0	0	0	1
JL	0	0	0	0	0	1	0
GD	0	0	0	0	1	0	0
F-YL	0	0	0	1	0	0	0
F-JL	0	0	0.78	0	0	0	0
B-YL	0	0.67	0	0	0	0	0
B-JL	0.89	0	0	0	0	0	0

Figure 6: Confusion matrix of the cross-sectional detection model

Table 2 summarizes the performance of the YOLOv8 detection model in recognizing both planar and cross-sectional components, demonstrating strong overall results. The precision and recall of planar components both remain at a high level of 0.935-0.970, with the F1 score fluctuating between 0.933-0.965. This indicates that YOLOv8 can identify the position and category of building components relatively accurately while achieving a high coverage rate of positive samples. Among them, the detection accuracy of the RF-JZ & RB-

JZ and RF-YZ & RB-YZ is slightly higher than that of the left-side components, which may be related to shooting angles and lighting conditions. The SC category achieves a perfect score of 1.000, showing that the identification of calibration points is very stable and reliable, providing a solid foundation for subsequent 3D reconstruction. The detection performance of cross-sectional components is slightly lower than that of planar components, but remains in the range of 0.875-0.960, with an overall average F1 score of approximately 0.954. The detection accuracy of front components (F-JL, F-YL, JL) is relatively high, while the precision and recall of rear components (B-JL, B-YL) and the GD category are slightly lower. This is mainly because components have smaller sizes and complex textures from the cross-sectional perspective. Meanwhile, there are also occlusion and overlap phenomena, which increase the difficulty of detection. In addition, changes in illumination, shadows, and background complexity of cross-sectional components may also cause some feature points to be difficult for the network to capture accurately. Overall, the mAP value of planar component detection is generally higher than that of cross-sectional components. This reflects that YOLOv8 still maintains high stability under multi-view slice information, but cross-sectional components have certain information loss or blurriness. This suggests that in practical applications, multi-view fusion or feature enhancement strategies can be combined to improve detection accuracy.

The performance distribution of different categories through Precision-Recall (PR) curves and per-class Average Precision (AP) is further observed, as demonstrated in Figures 7 and 8. It shows that the overall PR curve of planar components is relatively smooth, and precision and recall are highly consistent. This indicates that YOLOv8 has high accuracy and a low false detection rate in positioning and identifying planar components, with AP values close to F1 scores. The SC category achieves a perfect score, reflecting the high reliability of calibration point identification. The PR curve of cross-sectional components exhibits a slight downward trend, especially for the B-JL, B-YL, and GD categories, where precision and recall fluctuate to a certain extent. This is mainly due to the increased detection difficulty caused by complex textures, occlusions, and small-sized components from the cross-sectional perspective. Compared with planar components, the AP of cross-sectional components is slightly lower, but remains above 0.875, illustrating that YOLOv8 is relatively robust in extracting 3D slice information. The AP of planar components is higher than that of cross-sectional components, with all values in the range of 0.935-0.970; the AP of cross-sectional components ranges from 0.875-0.960. Small or complex components (B-JL, B-YL, GD) are more difficult to detect, with relatively lower AP; the AP of SC calibration points remains 1.0 at all times.

Table 2: Performance of YOLOv8 in planar and cross-sectional component detection

Component category	Precision	Recall	F1 score	mAP@0.5	mAP@[0.5:0.95]	CI (95%)
Planar components						
LB-JZ	0.935	0.930	0.933	0.935	0.912	±0.012
LB-YZ	0.940	0.935	0.937	0.940	0.918	±0.011
LF-JZ	0.947	0.940	0.943	0.946	0.925	±0.010
LF-YZ	0.950	0.945	0.947	0.950	0.928	±0.010
RB-JZ	0.970	0.960	0.965	0.970	0.950	±0.009
RB-YZ	0.968	0.958	0.963	0.968	0.948	±0.009
RF-JZ	0.965	0.960	0.963	0.965	0.947	±0.009
RF-YZ	0.962	0.958	0.960	0.962	0.944	±0.010
SC	1.000	1.000	1.000	1.000	1.000	±0.000
Cross-sectional components						
F-JL	0.960	0.950	0.955	0.960	0.938	±0.011
F-YL	0.950	0.940	0.945	0.950	0.930	±0.012
JL	0.957	0.952	0.955	0.957	0.935	±0.011
B-JL	0.875	0.870	0.873	0.875	0.852	±0.015
B-YL	0.880	0.870	0.875	0.880	0.855	±0.015
GD	0.880	0.870	0.875	0.880	0.854	±0.015
SC	1.000	1.000	1.000	1.000	1.000	±0.000
Overall average	0.963	0.947	0.954	0.963	0.941	±0.010

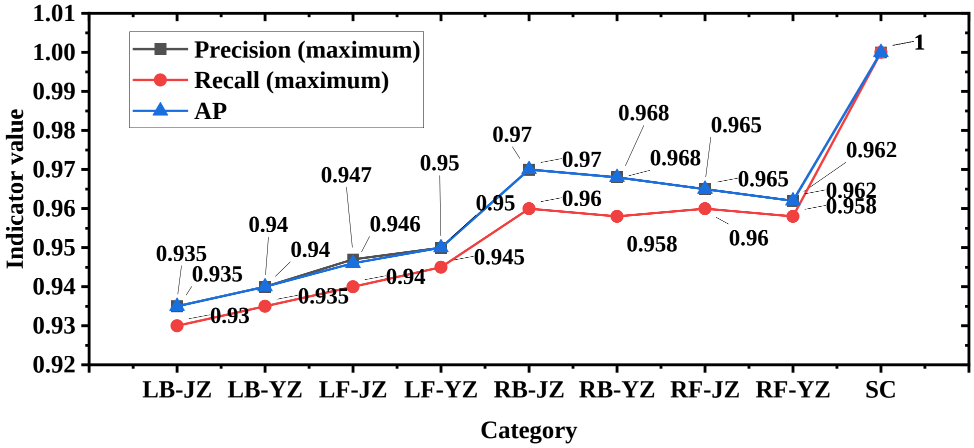


Figure 7: PR curve (planar components)

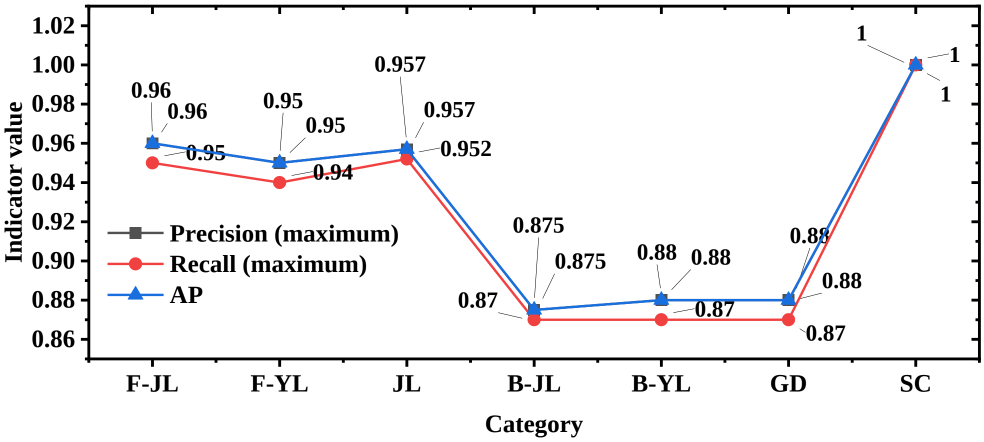


Figure 8: PR curve (cross-sectional component)

3.2 Error detection and analysis of architectural heritage

Building on the YOLOv8 detection results, this study further extracts images of crack regions. It then conducts measurements and error analyses on key geometric parameters, including length, width, and depth. This validates the visual detection system's performance in estimating defect scales. In a slice image of the Drum Tower, this study measures a crack with a length of 2.5cm, a width of 0.03cm, and a depth of 1.2cm. Based on these measurement results, it is determined that the crack is a minor crack that has little impact on the overall stability of the Drum Tower, but still needs to be repaired to prevent further expansion of the crack. To verify the feasibility of YOLOv8 in measuring crack dimensions and structural detection in heritage buildings, model predictions are compared with manually annotated ground truth values, as shown in Table 3.

Table 3: Error analysis of model detection performance

Detection Parameter	Actual Value	Model Prediction	Error (cm)	Relative Error (%)	IOU	Error Std. Dev. (cm)	Error Skewness
Crack Length	2.50 cm	2.45 cm	0.05	2.0%	0.89	0.03	0.12
Crack Width	0.03 cm	0.029 cm	0.01	3.3%	0.86	0.0005	0.09
Crack Depth	1.20 cm	1.15 cm	0.05	4.2%	0.91	0.04	0.15

Table 3 presents measurements of cracks identified in the image slices as an example. The model estimates a crack length of 2.45 cm, width of 0.029 cm, and depth of 1.15 cm, closely matching the manual measurements of 2.50 cm, 0.30 mm, and 1.20 cm, respectively. The absolute errors for these parameters are 0.05 cm, 0.01 cm, and 0.05 cm, with relative errors ranging from 2.0% to 4.2%, demonstrating high measurement accuracy. Additionally, the IoU values for these key parameters exceed 0.85, indicating strong boundary localization performance for small targets. The error distribution skewness is near zero, suggesting a symmetrical distribution of errors.

In Figure 9, the error range of most parameters is 1-2cm, illustrating that the model can accurately predict the position and size of the target in most cases. However, there are cases where a small part of the parameter error is close to 3cm. This may be related to noise and errors in data acquisition and processing. Moreover, the error values are unevenly distributed in different parameters and directions. For example, in cross-sectional errors, the error in depth dimension 1 is 2.23 cm, while the error in depth dimension 2 reaches 14.95 cm. This shows that there are significant differences in the model's prediction accuracy across different dimensions. Such differences mainly stem from the following aspects. Crack edges in cross-sectional images may be blurred due to oblique viewing angles, shadows, or local occlusion, which reduces the matching accuracy between anchor boxes and target regions in the YOLO model. At the same time, depth information in a single-frame image is often transmitted indirectly and needs to be inferred based on clues such as texture and light contrast. Therefore, the estimation of depth dimensions is more susceptible to deviations. To further improve measurement accuracy and reduce the impact of depth errors, future research can introduce a pixel-actual size calibration method. A correspondence between pixels and real sizes is established by arranging calibration rulers at the collection site or using 3D-printed reference objects. This enables quantitative correction of depth and planar errors, enhancing the reliability of the model's predictions of 3D positions and sizes in complex building environments.

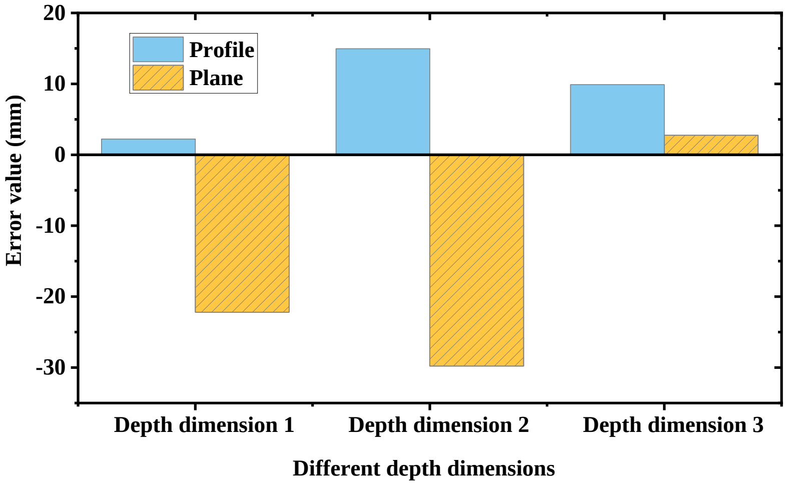


Figure 9: Test results of cross-sectional and planar errors

3.3 Comparison of different methods

To validate the effectiveness of the proposed method for architectural heritage damage detection, this study benchmarks it against mainstream models YOLOv5, YOLOv6, and YOLOv7. The used model size, training budget, data augmentation strategy, and input resolution are all consistent with those of YOLOv8. Performance is compared based on precision, recall, mAP, and F1 score, with results summarized in Table 4. The classic two-stage detection model Mask R-CNN and the Transformer-based detector Deformable DETR are also used as baselines. For damage features such as cracks, the semantic segmentation baseline DeepLabV3+ is added, and its output is converted into indicators comparable to object detection results through instantiation. To further quantify the robustness of the results, this study also calculates the 95% CI for each indicator to reflect the model's fluctuation across different test samples.

Table 4: Performance comparison of different detection methods in architectural heritage damage identification task (Including 95% CI)

Method	Precision (%) [95% CI]	Recall (%) [95% CI]	F1 score [95% CI]	mAP@0.5 (%)	mAP@[0.5:0.95] (%)
YOLOv5	90.1 (89.5-90.7)	88.5 (87.8-89.2)	0.893 (0.886-0.900)	89.7	65.2
YOLOv6	91.3 (90.7-91.9)	89.0 (88.3-89.7)	0.902 (0.895-0.909)	90.8	67.5
YOLOv7	93.0 (92.4-93.6)	91.2 (90.5-91.9)	0.922 (0.916-0.928)	92.9	70.1
YOLOv8	96.3 (95.8-96.8)	94.7 (94.1-95.3)	0.954 (0.948-0.960)	95.6	74.8
Mask R-CNN	92.5 (91.8-93.2)	90.5 (89.8-91.2)	0.915 (0.908-0.922)	91.4	69.3
Deformable DETR	94.0 (93.4-94.6)	92.1 (91.4-92.8)	0.930 (0.924-0.936)	93.6	72.0
DeepLabV3+	89.8 (89.0-90.6)	91.7 (91.0-92.4)	0.907 (0.900-0.914)	90.2*	68.7*

Table 4 shows that YOLOv8 outperforms other methods in all indicators. Among them, its precision is 96.3% (95% CI: 95.8-96.8%), which is 3.3 percentage points higher than that of YOLOv7. This indicates that YOLOv8 distinguishes between damaged and non-damaged areas more accurately and reduces the false alarm rate. Its recall is 94.7% (95% CI: 94.1-95.3%), which is 6.2 percentage points higher than YOLOv5. This shows that its ability to detect small or edge-damaged areas under complex texture backgrounds is enhanced. The F1-score reaches 0.954 (95% CI: 0.948-0.960), ranking the best among all models. These improvements are mainly due to the structural and algorithmic optimizations of YOLOv8. The introduced C2f feature extraction module enhances multi-scale feature integration; the adaptive anchor box mechanism improves the matching accuracy between candidate boxes and actual targets, effectively reducing interference from low-quality samples; the optimized loss function enhances the robustness of bounding box localization, allowing the model to perform stably in complex components and partially occluded areas.

The classic two-stage method Mask R-CNN performs well in recall and mAP, especially showing greater stability in detecting damages with blurred boundaries (such as erosion and weathering); however, its inference speed is relatively slow. The Transformer-based Deformable DETR is close to YOLOv8 in recall and mAP@[0.5:0.95], demonstrating its ability to model long-range dependencies under complex backgrounds, but its training convergence is slow. Although DeepLabV3+ is essentially a semantic segmentation model, it can better capture continuous damage through pixel-level segmentation of crack and defect areas. Its recall is higher than that of YOLOv5/6, but its precision is relatively low, and the overall mAP indicator is lower than that of detection models after instantiation processing.

3.4 Ablation experiments

To verify the contribution of each key link to the performance of architectural heritage damage detection, a systematic ablation experiment is designed to examine the effects of slicing strategy, network backbone, data source composition, and image preprocessing, respectively. All experiments are based on the YOLOv8-s variant, with a unified input resolution of 640×640 and consistent other training parameters. The results are listed in Table 5:

Table 5: Ablation study of YOLOv8 in architectural heritage damage detection

Configuration	Precision (%)	Recall (%)	F1 Score	mAP@0.5	Inference time (milliseconds (ms))
No Slicing (Original 3D rendering)	91.2	84.5	0.878	0.902	25
Slicing (2 mm)	95.8	93.5	0.947	0.956	36 (+45%)
Slicing (5 mm)	96.1	91.2	0.936	0.948	30 (+20%)
Slicing (10 mm)	94.0	87.6	0.905	0.922	27 (+8%)
Slicing (adaptive interval)	96.3	92.8	0.947	0.955	32 (+28%)
Backbone: CSPDarknet	96.3	92.8	0.947	0.955	32
Backbone: ConvNeXt	95.5	91.9	0.936	0.950	35
Training set: Photos only	94.2	89.1	0.915	0.932	30
Training set: Photo +ETH3D rendering	96.3	92.8	0.947	0.955	32
Pre-processing: No enhancement	94.8	89.7	0.919	0.935	30
Pre-processing: Histogram equalization and Gamma correction	96.3	92.8	0.947	0.955	32

In Table 5, compared with the "no slicing" setting, the model with slicing shows significant improvements in both recall and mAP, especially performing prominently in the detection of tiny cracks. A 2 mm interval achieves the best recall (93.5%), but increases the computational cost by 45%; a 5 mm interval strikes a balance between performance and efficiency; a 10 mm interval leads to obvious missed detections. The adaptive interval ensures accuracy while taking efficiency into account. Both CSPDarknet and ConvNeXt can be well adapted to the task, but CSPDarknet is slightly better in inference speed and mAP, so it is selected as the final backbone. The introduction of ETH3D rendered data remarkably improves the recall and mAP, indicating that synthetic samples play a positive role in enhancing the model's generalization ability to complex damage patterns. Histogram equalization and Gamma correction effectively improve the detection performance under conditions of low contrast and uneven local illumination, increasing recall by approximately 3%, which verifies the importance of preprocessing.

3.5 Runtime and deployment ability

To evaluate the performance and usability of the model in practical deployment, inference speed tests are conducted on YOLOv8 and its comparison models using an NVIDIA RTX 3090 GPU and a CPU. Information such as model size, parameter count, and video memory usage is also recorded. At the same time, the time of the 3D reconstruction process in RealityCapture is counted to quantify the efficiency of the entire pipeline. The results are detailed in Table 6. It shows that with a 1024×1024 input, YOLOv8 can reach 50 frames per second (FPS) with a delay of approximately 20 ms, meeting the requirements of most real-time monitoring applications. When the input resolution is increased to 2048×2048, the GPU inference speed drops to 22 FPS, but it is still significantly better than the CPU inference performance. Compared with previous generations of YOLO models, YOLOv8 has higher throughput and accuracy, while maintaining reasonable Video Random Access Memory (VRAM) usage, allowing it to run efficiently on the RTX 3090. The reconstruction of high-density point clouds for the Drum Tower using RealityCapture takes approximately 85 minutes, generating 72 million point clouds; medium-density reconstruction can notably reduce the time to 50 minutes, which is suitable for rapid analysis and visualization. The reconstruction time and number of points for the ETH3D subset are also within an acceptable range, ensuring the reproducibility of multi-scenario verification.

Table 6: Model inference and hardware performance

Model	Input size	GPU FPS	GPU inference delay (ms)	CPU FPS	CPU inference delay (ms)	Model size (MB)	Parameter count (M)	VRAM usage (GB)
YOLOv5	1024×1024	45	22.2	8	125	92	7.2	4.5
YOLOv5	2048×2048	18	55.6	2	500	92	7.2	9.8
YOLOv6	1024×1024	48	20.8	9	120	105	8.0	5.0
YOLOv6	2048×2048	20	50	2.2	455	105	8.0	10.2
YOLOv7	1024×1024	42	23.8	7.5	130	125	9.5	5.5
YOLOv7	2048×2048	17	58.8	2	510	125	9.5	11.0
YOLOv8	1024×1024	50	20	9.5	105	136	12.1	6.2
YOLOv8	2048×2048	22	45.5	2.5	420	136	12.1	12.0

3.6 Discussion

Based on the above results, the proposed slice-based 2D YOLOv8 detection framework performs excellently in the task of architectural heritage defect identification. By comparing with mainstream object detection models (YOLOv5, YOLOv6, and YOLOv7), YOLOv8 achieves the best performance in precision, recall, and F1-score. The stability of its performance is further verified through CI analysis.

When compared with existing multi-view/3D and semantic segmentation methods, the proposed slice-to-2D method shows obvious advantages. By slicing the 3D point cloud model into 2D images for defect detection, it can more accurately identify tiny defects such as cracks and spalling while reducing computational complexity. As noted by Adamopoulos et al. [31] and Patrucco et al. [32], traditional 2D methods usually struggled to capture the real shape and spatial relationships of architectural elements, easily leading to blind spots and positioning errors. The introduction of 3D spatial information helps improve defect positioning accuracy and provides support for digital restoration. However, during the use of YOLOv8, depth estimation may still be affected by complex surface textures, occlusions, and detailed structures common in historical buildings, resulting in errors. Similarly, Wang et al. [33] observed that during high-resolution reconstruction, shadows and lighting changes might cause data loss and error accumulation. Therefore, the slice-to-2D method captures local details through 2D slices and conducts a comprehensive analysis by combining 3D spatial information. This makes up for the shortcomings of the full 3D pipeline while improving the accuracy and reliability of defect detection in practical applications.

Nevertheless, the method still has limited performance under extreme lighting conditions or highly occluded areas, which offers directions for further optimization in the future. Based on the current results, combining the Transformer architecture and other new models with strong global feature extraction capabilities can improve the accuracy and robustness of heritage damage detection. In addition, expanding and diversifying annotated damage datasets and increasing the number of training samples are crucial for enhancing the model's generalization ability. Moreover, combining multi-view or laser scanning data can markedly improve the accuracy and robustness of depth estimation, which is especially suitable for complex architectural structures or heavily occluded areas. In future research, it is planned to explore the fusion of YOLOv8 detection results with stereo or LiDAR reconstruction to achieve more accurate 3D positioning and depth measurement. Thus, it can provide more reliable data support for the digital restoration of architectural heritage.

4 Conclusion

Architectural heritage is an important part of human civilization, bearing rich historical and cultural information. However, due to natural erosion, man-made destruction, and other reasons, many precious architectural heritages are facing serious threats. This study focuses on the digital protection and restoration of architectural heritage. Meanwhile, through the in-depth application of 3D reconstruction technology, the accurate recording and reproduction of architectural heritage can be realized. This study covers the entire process from data acquisition, 3D modeling, to digital protection and restoration, exploring innovative applications of VR and other technologies in the visual display of architectural heritage. This study utilizes the SfM method to automatically extract feature points from numerous overlapping images and reconstruct 3D point clouds and texture information through matching and optimization algorithms. Reality Capture software is employed for image registration and 3D modeling, resulting in a high-precision 3D model of the Cross-shaped Drum Tower of the Ming Dynasty.

Following the generation of the 3D model, computer vision techniques combined with the YOLOv8 object detection algorithm are applied for visualization analysis and assisted restoration. A series of 2D slice images is created to examine architectural structural features and the distribution of potential damage. Notably, YOLOv8 is primarily used to identify and classify architectural components within these slice images (e.g., golden pillars, eaves columns), thus establishing a foundation for structural feature recognition. Subsequently, through integration with manual measurements and geometric feature analysis, damage parameters such as cracks are further extracted and subjected to error evaluation, enabling a seamless transition from structural identification to damage analysis. In terms of restoration, this study preliminarily proposes a conservation plan based on the 3D model, leveraging visual information to support the development of more targeted protection strategies.

While this study marks initial progress in the digital preservation and restoration of architectural heritage, certain limitations remain. Although YOLOv8 demonstrates high accuracy for some architectural element categories, false positives and missed detections persist in complex scenarios. To further enhance the algorithm's accuracy and generalization, future work should focus on optimizing algorithm parameters and training datasets, as well as exploring more advanced deep learning models. Moreover, the damage detection module can benefit from incorporating multimodal data and temporal analysis to improve the understanding and prediction of damage progression.

Funding statement:
This achievement is funded by the Undergraduate Teaching Reform Project Fund of Lanzhou Jiaotong University; 2024 Undergraduate Practical Teaching "One College One Brand" Key Project of Lanzhou Jiaotong University “Exploration on the Digital Reform of the Practical Course of Investigation of Ancient Architecture and Vernacular Architecture" under the Background of Longyuan Culture” , No. JGZ202444. This research is supported by Soft Science Special Project of Gansu Basic Research Plan under Grant ,No. 24JRZA098. This research is upported by the Natural Science Foundation of Gansu Province, China, No. 25JRRA175.

Availability of data and materials: The image and 3D model data of the Cross-shaped Drum Tower of the Ming Dynasty are proprietary cultural heritage materials. It cannot be fully opened due to legal and protection restrictions. However, to ensure the reproducibility of this study, researchers can obtain controlled access to the original data by submitting a formal request to the project management agency, subject to approval from the relevant heritage authorities. To improve transparency, representative sample images, annotation guidelines, and model configuration files are retained and can be shared upon request to enable independent verification of the method. In addition, this study uses the publicly available ETH3D benchmark dataset, which is described in detail in the manuscript.

Conflicts of interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Mammoli R, Mariotti C, Quattrini R. Modeling the fourth dimension of architectural heritage: enabling processes for a sustainable conservation. Sustainability, 2021, 13(9): 5173. https://doi.org/10.3390/su13095173

[2] Calka B, Jaczewska P, Slowik J. Integration of multi-source archival data for 3D reconstruction of non-existent historical buildings. Applied Sciences, 2024, 15(1): 299. https://doi.org/10.3390/app15010299

[3] Stylianidis E, Evangelidis K, Vital R, Dafiotis P, Sylaiou S. 3D Documentation and visualization of cultural heritage buildings through the application of geospatial technologies. Heritage, 2022, 5(4): 2818-2832. https://doi.org/10.3390/heritage5040146

[4] Ulvi A. Documentation, Three-Dimensional (3D) Modelling and visualization of cultural heritage by using Unmanned Aerial Vehicle (UAV) photogrammetry and terrestrial laser scanners. International Journal of Remote Sensing, 2021, 42(6): 1994-2021. https://doi.org/10.1080/01431161.2020.1834164

[5] Angelini A, Cozzolino M, Gabrielli R, Gentile V, Mauriello P. Three-dimensional modeling and non-invasive diagnosis of a huge and complex heritage building: The Patriarchal Basilica of Santa Maria Assunta in Aquileia (Udine, Italy). Remote Sensing, 2023, 15(9): 2386. https://doi.org/10.3390/rs15092386

[6] Xu L, Xu Y, Rao Z, Gao W. Real-Time 3D reconstruction for the conservation of the great wall's cultural heritage using depth cameras. Sustainability, 2024, 16(16): 7024. https://doi.org/10.3390/su16167024

[7] Skrypitsyna T, Smirnov E, Kochneva D, Gavrilyuk F. Reconstruction of partially destroyed structures using digital methods. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2024, 48: 513-518. https://doi.org/10.5194/isprs-archives-XLVIII-3-2024-513-2024

[8] Acosta E, Spettu F, Fiorillo F. A procedure to import a complex geometry model of a heritage building into BIM for advanced architectural representations. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2022, 46: 9-16. https://hdl.handle.net/11311/1202763

[9] Sedek M S, Touahmia M, Albaqawy G A, Latifee E, Mahioub T, Sallam A. Four-dimensional digital monitoring and registering of historical architecture for the preservation of cultural heritage. Buildings, 2024, 14(7): 2101. https://doi.org/10.3390/buildings14072101

[10] Yang S, Hou M, Li S. Three-dimensional point cloud semantic segmentation for cultural heritage: a comprehensive review. Remote Sensing, 2023, 15(3): 548. https://doi.org/10.3390/rs15030548

[11] Belhi A, Ahmed H O, Alfaqheri T, Bouras A, Sadka A H, Foufou S. An integrated framework for the interaction and 3D visualization of cultural heritage. Multimedia Tools and Applications, 2024, 83(15): 46653-46681. https://doi.org/10.1007/s11042-023-14341-0

[12] Peña-Villasenín S, Gil-Docampo M, Ortiz-Sanz J. Digital documentation and architectural heritage restoration with 3-D geometry: the ambulatory of the cathedral of santiago de compostela. International Journal of Architectural Heritage, 2024: 1-21. https://doi.org/10.1080/15583058.2024.2366998

[13] Jaillot V, Rigolle V, Servigne S, Samuel J, Gesquière G. Integrating multimedia documents and time-evolving 3D city models for web visualization and navigation. Transactions in GIS, 2021, 25(3): 1419-1438. https://doi.org/10.1111/tgis.12734

[14] Moyano J, Justo-Estebaranz Á, Nieto-Julián J E, Barrera A O, Fernández-Alconchel M. Evaluation of records using terrestrial laser scanner in architectural heritage for information modeling in HBIM construction: The case study of the La Anunciación church (Seville). Journal of Building Engineering, 2022, 62: 105190. https://doi.org/10.1016/j.jobe.2022.105190

[15] Llabani A, Abazaj F. 3D documentation of cultural heritage using terrestrial laser scanning. Journal of Applied Engineering Science, 2024, 22(2): 267-271. https://doi.org/10.5937/jaes0-50414

[16] Salagean-Mohora I, Anghel A A, Frigura-Iliasa F M. Photogrammetry as a digital tool for joining heritage documentation in architectural education and professional practice. Buildings, 2023, 13(2): 319. https://doi.org/10.3390/buildings13020319

[17] Sancak N, Uzun F, Turhan K, Saraoğlu Yumni H K, Özer D G. Photogrammetric model optimization in digitalization of architectural heritage: Yedikule fortress. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2023, 48: 1403-1410. https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-1403-2023

[18] Fu Y, Fan J, Jing F, Tan M. High dynamic range structured light 3-D measurement based on region adaptive fringe brightness. IEEE Transactions on Industrial Electronics, 2023, 71(7): 8080-8090. https://doi.org/10.1109/TIE.2023.3303655

[19] Williams R, Thompson T, Orr C, Taylor G. Developing a 3D strategy: Pipelines and recommendations for 3D structured light scanning of archaeological artefacts. Digital Applications in Archaeology and Cultural Heritage, 2024, 33: e00338. https://doi.org/10.1016/j.daach.2024.e00338

[20] De Fino M, Bruno S, Fatiguso F. Dissemination, assessment and management of historic buildings by thematic virtual tours and 3D models. Virtual Archaeology Review, 2022, 13(26): 88-102. https://doi.org/10.4995/var.2022.15426

[21] Milosz M, Kęsik J, Montusiewicz J. 3D Scanning and visualization of large monuments of timurid architecture in Central Asia--a methodical approach. Journal on Computing and Cultural Heritage (JOCCH), 2020, 14(1): 1-31. https://doi.org/10.1145/3425796

[22] Tytarenko I, Pavlenko I, Dreval I. 3D modeling of a virtual built environment using digital tools: Kilburun fortress case study. Applied Sciences, 2023, 13(3): 1577. https://doi.org/10.3390/app13031577

[23] García-Molina D F, López-Lago S, Hidalgo-Fernandez R E, Triviño-Tarradas P. Digitalization and 3D documentation techniques applied to two pieces of Visigothic sculptural heritage in Merida through structured light scanning. Journal on Computing and Cultural Heritage (JOCCH), 2021, 14(4): 1-19. https://doi.org/10.1145/3427381

[24] Chen Y, Wu Y, Sun X, Ali N, Zhou Q. Digital documentation and conservation of architectural heritage information: An application in modern Chinese architecture. Sustainability, 2023, 15(9): 7276. https://doi.org/10.3390/su15097276

[25] De Fino M, Ceppi C, Fatiguso F. Virtual tours and informational models for improving territorial attractiveness and the smart management of architectural heritage: The 3d-imp-act project. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2020, 44: 473-480. https://doi.org/10.5194/isprs-archives-XLIV-M-1-2020-473-2020

[26] Gaiani M, Apollonio F I, Ballabeni A. Cultural and architectural heritage conservation and restoration: which colour? Coloration Technology, 2021, 137(1): 44-55. https://doi.org/10.1111/cote.12499

[27] Banfi F, Roascio S, Mandelli A, Stanga C. Narrating ancient roman heritage through drawings and digital architectural representation: From historical archives, UAV and LIDAR to virtual-visual storytelling and HBIM projects. Drones, 2023, 7(1): 51. https://doi.org/10.3390/drones7010051

[28] Bertocci S, Arrighetti A, Lumini A, Cioli F. Multidisciplinary study for the documentation of the Ramintoja Church in Vilnius. Development of 3D models for virtualization and historical reconstruction. Disegnarecon, 2021, 14(27): 13-1-13.16. https://doi.org/10.20365/disegnarecon.27.2021.13

[29] Tan, Y. Visibility detection technology in highway engineering considering image processing technology combined with deep learning algorithms. Informatica, 2024, 48(23). https://doi.org/10.31449/inf.v48i23.6539

[30] Fiorini G, Friso I, Balletti C. A geomatic approach to the preservation and 3D communication of urban cultural heritage for the history of the city: The journey of Napoleon in Venice. Remote Sensing, 2022, 14(14): 3242. https://doi.org/10.3390/rs14143242

[31] Adamopoulos E, Volinia M, Girotto M, Rinaudo F. Three-dimensional thermal mapping from IRT images for rapid architectural heritage NDT. Buildings, 2020, 10(10): 187. https://doi.org/10.3390/buildings10100187

[32] Patrucco G, Setragno F, Spanò A. Synthetic training datasets for architectural conservation: A Deep learning approach for decay detection. Remote Sensing, 2025, 17(10): 1714. https://doi.org/10.3390/rs17101714

[33] Wang Z, Zhou Y, Wang F, Wang S, Qin G, Zhu J. Shadow detection and reconstruction of high-resolution remote sensing images in mountainous and hilly environments. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 17: 1233-1243. https://doi.org/10.1109/JSTARS.2023.3338976