The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-M-9-2025
30th CIPA Symposium “Heritage Conservation from Bits:
From Digital Documentation to Data-driven Heritage Conservation”, 25–29 August 2025, Seoul, Republic of Korea

Dense 3D Reconstruction of Ancient Architectural Heritage by Integrating Cross-View Geometric Prior and Semantic Enhancement

Yongshuai Liu[1], Tao Shen[1], Liang Huo[1], Wenfei Shen[1]

[1] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China - 19932481104@163.com

Keywords: Ancient Building Protection, Crowd-sourced Images, Cross-view Geometric Prior, Semantic Enhancement

Abstract

Ancient buildings are in urgent need of protection due to multiple diseases such as structural crack expansion and painted fading. Multi-view 3D reconstruction of crowdsource images has been widely used due to its high flexibility and low cost. This study proposes a multimodal global optimization framework for the protection of ancient buildings to address the low accuracy of crowdsource image screening and the problems of paired view cumulative error and detail sparsity of the MASt3R method. First, the EfficientNet is improved based on the CBAM mechanism to improve its screening robustness in low-quality crowdsource images. Secondly, a cross-reference view block is designed to achieve cross-view geometric prior fusion through a multi-view attention mechanism. A semantic-guided matching enhancement strategy is further introduced to segment key areas of the building based on the SAM model to focus on detail reconstruction. Experiments show that this method improves the number of point clouds and surface density by 23.4% and 25.9% respectively, effectively solving the problems of geometric distortion and weak texture details in the reconstruction of typical architectural details such as brackets and plaques in traditional methods.

1. Introduction

As a material witness of historical civilization, the digital protection of ancient buildings has long faced technical adaptability challenges. Among traditional 3D reconstruction technologies, although laser scanning can achieve sub-millimeter accuracy, it is difficult to popularize due to equipment cost and operation complexity; aerial images are limited by airspace control and low-altitude resolution, making it difficult to capture detailed structures such as brackets and mortise and tenon joints; professional sequence image reconstruction methods are highly dependent on lighting conditions and shooting trajectories, and cannot meet the flexibility requirements of on-site inspections of heritage sites. In this context, crowd-source imaging technology has gradually become a research hotspot for the digitization of cultural heritage with its multi-source heterogeneous data acquisition capabilities and low-cost advantages, but the inherent quality heterogeneity and lack of perspective of unstructured data have led to significant accuracy bottlenecks in existing methods in the reconstruction of architectural details.

The current research mainly focuses on two dimensions: data acquisition optimization and reconstruction algorithm innovation. At the data level, the dynamic retrieval technology based on geo-fences achieves accurate recall of building main body images by fusing multi-source APIs (Cui,2019); the intelligent screening algorithm builds a quality assessment system for ancient building images by fusion modeling of deep features and classic descriptors, such as the improved YOLOv4(Bochkovskiy, A,2020) model that increases the recall rate of occluded samples to 81.2% through the DenseNet169 backbone network (Wang,2021). At the reconstruction algorithm level, in traditional explicit methods, the motion restoration structure (SfM) realizes disordered image reconstruction through feature matching and bundling adjustment, but is limited by repeated texture mismatching, and topological breaks often occur in high-frequency detail areas such as glazed tiles; although passive stereo vision methods (such as SGM-Nets) can handle radiation differences, they place strict requirements on image resolution and lighting consistency. The intervention of deep learning technology has opened up a new path for three-dimensional reconstruction. The PSGN network based on point cloud generation achieves single-view coarse-grained reconstruction through an encoder-decoder architecture, but there are blurred details on complex components such as brackets; voxel reconstruction methods (such as 3D-R2N2) use recurrent neural networks to process multi-view inputs, and their voxel representation method causes the memory overhead and computational complexity to grow cubically(Choy,C.B.,2016).Neural radiance field (NeRF) breaks through the geometric constraints of explicit reconstruction through implicit neural representation(Mildenhall,B.,2021). Mip-NeRF uses cone projection and integrated position encoding technology to improve the peak signal-to-noise ratio of repeated texture areas, but its dependence on dense view input and the sensor compatibility defects of hash coding make it difficult to adapt to the heterogeneous characteristics of Crowd-sourced images (Barron,J.T.,2021) .The recently proposed DUSt3R method abandons the traditional pose estimation process and directly constructs a dense three-dimensional correspondence field between image pairs(Wang, S.,2024). Its improved version MASt3R achieves sub-pixel matching through a local feature enhancement strategy(Duisterhof, B.,2024), but when extended to large-scale images, it exposes the problem of pairwise matching error accumulation, resulting in overall structural distortion.

In summary, existing methods face three contradictions in the reconstruction of ancient buildings: first, although explicit reconstruction methods (such as COLMAP) are physically interpretable, they are not adaptable enough to weak texture areas; second, although implicit neural expressions can depict complex surfaces, they are difficult to implement in engineering due to excessive computing resource requirements; third, although emerging unsupervised methods (such as MASt3R)

reduce the dependence on camera parameters, the paired matching paradigm has inherent defects in maintaining geometric consistency across perspectives. These contradictions are particularly prominent in ancient building scenes - the weathering cracks of wooden components require high-precision reconstruction, and painted patterns also require high color reproduction. Traditional methods are difficult to meet the needs of high-precision repair and monitoring. How to break through the dual bottlenecks of Crowd-sourced image data optimization and detail reconstruction accuracy has become a difficult problem that needs to be overcome in the field of digital heritage protection.

2. Data and Methods

2.1 Introduction to the research area and dataset

This study selected the Temple of Heaven in Beijing as the research object. This building complex is the largest and most complete imperial worship building complex in China during the Ming and Qing dynasties. Its unique circular plane layout, three-layer eaves and pointed roof structure, and complex wooden structure system pose challenges to the spatial analysis and shape restoration capabilities of the three-dimensional reconstruction algorithm. As a world cultural heritage, the digital protection of the Temple of Heaven has the dual significance of cultural inheritance and technological innovation. The natural aging of its wooden components, the changes in the microenvironment caused by tourist activities, and the maintenance of the integrity of the historical style all require high-precision three-dimensional models to support monitoring and restoration decisions. Based on text query, this study obtained a total of 5,000 images of the Temple of Heaven from Baidu Images and other open network platforms in batches to form the initial crowd-source image set. The optimal crowd-source image set after screening and processing by the method proposed in this paper contains 42 optimal crowd-source images, covering upward, horizontal and downward perspectives, and completing the presentation of the geometric shapes around the Hall of Prayer for Good Harvests, complex structural details, and diversity of tourist perspectives.

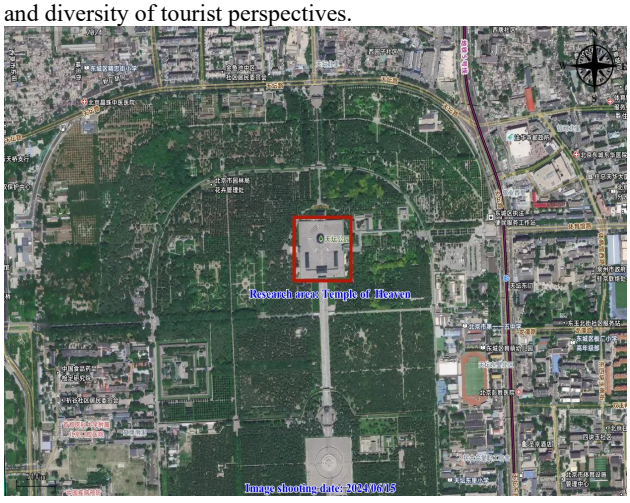

Figure 1. Study area (The Temple of Heaven is located in the Temple of Heaven Park in Dongcheng District, Beijing, China. Its architectural complex and landscape design have influenced the sacrificial group paradigm in East Asia. It carries the comprehensive expression of ancient Chinese philosophy, astronomy and ritual culture. The digital protection of the Temple of Heaven has the dual significance of cultural inheritance and technological innovation).

Figure 2. Schematic diagram of the main body and structural details of the Temple of Heaven.

2.2 Research methods

2.2.1 Crowd-source image screening method based on improved EfficientNet

In order to solve the problems of occlusion, blur and fine-grained feature recognition in the Crowd-sourced image screening task, this study proposes an EfficientNet model based on the improved CBAM (Convolutional Block Attention Module) attention mechanism. The SE module built into the traditional EfficientNet series model only adjusts the feature weights through channel attention, which is difficult to deal with the complex spatial occlusion problem in Crowd-sourced images. The CBAM module can dynamically enhance the feature response of key areas by combining channel and spatial attention mechanisms. This paper proposes to embed CBAM into the EfficientNet backbone network, and replace the original SE (Squeeze-and-Excitation) module in EfficientNet with the CBAM dual attention module to improve the model's ability to focus on the features of key areas of ancient buildings and the model's screening robustness in low-quality Crowd-sourced images.

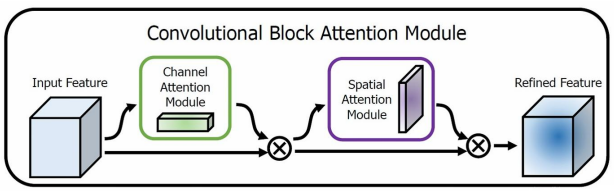

Figure 3. Overview of CBAM,The module has two sequential sub-modules: channel and spatial attention mechanism modules,,through CBAM to adaptively refine the intermediate feature maps at each,convolutional block of the deep network.

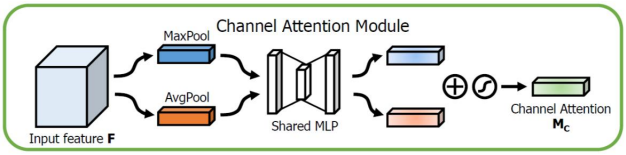

Figure 4. Channel attention module: the channel dimension remains unchanged and the spatial dimension is compressed. This module focuses on the meaningful information in the input image.

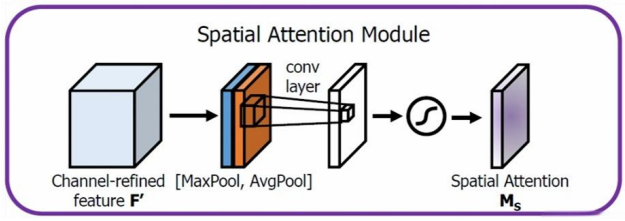

Figure 5. Spatial attention module: the spatial dimension remains unchanged and the channel dimension is compressed. This module focuses on the location information of the target.

This study firstly used the text query method and the Python Scrapy framework to crawl 5000 images containing the keyword "Temple of Heaven" from Baidu Images, Flickr and other open network platforms (resolution $\geqslant$ 1920 × 1080). Secondly, the images outside the Beijing area were filtered out by geo-tag filtering, and the data with latitude and longitude within the range of (116.413° E, 39.884° N) ± 0.02° were retained to ensure spatial correlation. Then, the quality of the above-mentioned retained data was initially screened, and the image blur index (Laplacian variance < 100) and illumination uniformity index (standard deviation after histogram equalization > 50) were calculated using OpenCV, and 2406 low-quality images were eliminated. Finally, the main position of the Hall of Prayer for Good Harvests in the image was identified based on the YOLOv9 target detection model, and the shooting angle was calculated by the perspective projection matrix. The upward angle was used to capture the details of the glazed tile brackets; the horizontal angle showed the proportion of the building facade; the downward angle showed the roof caisson structure, and the Crowd-sourced images that met the reconstruction perspective requirements were retained.

In this study, CBAM-EfficientNet is used to screen target images after removing low-quality images. The specific implementation process and network structure of this method are shown in Figure 6.

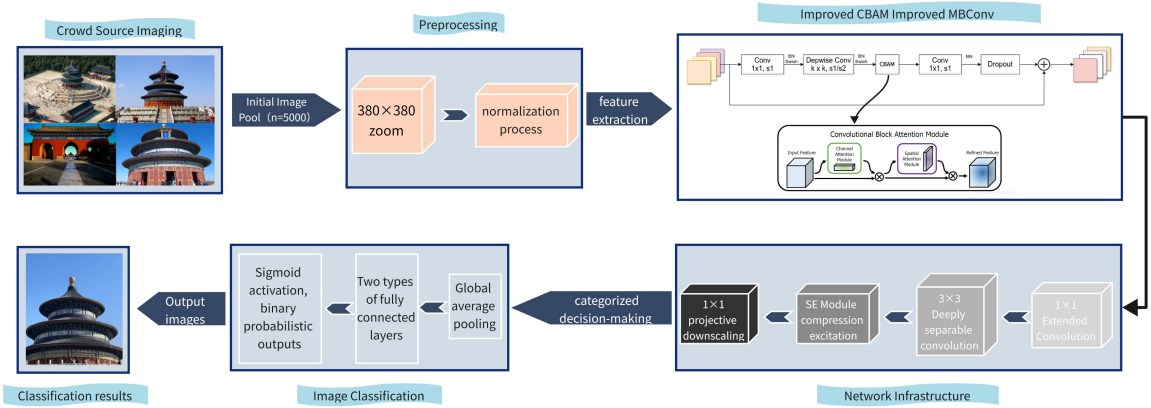

Figure 6. EfficientNet ancient building crowd-sourced images screening architecture based on CBAM dual attention mechanism.

Specifically, first, in the channel attention generation stage, global average pooling and global maximum pooling are performed on the input feature map $F \in R^{H\times W\times C}$ to generate channel descriptors and $D_{max} \in R^{1\times 1\times C}$. The channel attention weight $\mathrm{M_c(F)}$ is generated by a two-layer MLP with shared weights (the hidden layer dimension is $C/16$), and the formula is:

$$M_c(F) = \sigma\left(MLP\left(D_{avg}\right) + MLP(D_{max})\right) \quad (1)$$

Among them, σ is the Sigmoid activation function, and the MLP structure is implemented with reference to the CBAM standard.

Secondly, in the spatial attention enhancement stage, the channel-weighted feature map $F^{'} = M_c(F) \odot F$ is average pooled and max pooled along the channel axis to generate two-channel feature maps $F_{avg} \in R^{H\times W\times 1}$ and $F_{max} \in R^{H\times W\times 1}$. After splicing, a 7 × 7 convolution is performed to generate the spatial attention weight $M_s\left(F^{'}\right)$. This design draws on the classic structure of the CBAM spatial attention module.

$$M_s\left(F^{'}\right) = \sigma\left(Conv_{7\times 7}\left(\left[F_{avg}; F_{max}\right]\right)\right) \quad (2)$$

Then, in the feature fusion stage, the channels are multiplied element-wise with the spatial attention weights to output the final enhanced features:

$$F^{''} = M_s\left(F^{'}\right) \odot F^{'} \quad (3)$$

Finally, in the residual connection optimization stage, the original depthwise separable convolution and dilation-contraction structure of MBConv is retained, and the CBAM module is inserted after the deep convolution layer and before the residual connection to ensure that the attention mechanism acts on the high-order semantic information after feature extraction.

This study uses EfficientNet-B0 as the baseline model and optimizes the feature extraction capability by embedding the CBAM dual attention mechanism. The experimental data uses the Temple of Heaven Crowd-sourced images dataset (5000 images, including 30% occluded samples), and the training set and test set are divided into 8:2. In the data preprocessing stage, the geometric distortion of the image is first corrected by the affine transformation of OpenCV to eliminate the perspective deformation caused by the difference in shooting angles; secondly, CLAHE (Contrast-Limited Adaptive Histogram Equalization) is used to normalize the image illumination to alleviate the interference of low-light or over-exposed areas on model training; finally, random horizontal flipping, ± 15° rotation and Gaussian noise injection (σ=0.05) are used to enhance data diversity.

The specific configuration of the experimental environment is shown in Table 1.The optimizer selects AdamW (initial

learning rate 3×10^{-4}, weight decay 1×10^{-4}), and the batch size is set to 32. The loss function uses Focal Loss (γ=2, σ=0.25) to alleviate the category imbalance problem caused by occluded samples, and cooperates with cosine annealing learning rate scheduling and early stopping strategy (patience value=15) to prevent overfitting. To verify the effectiveness of this research method, three groups of comparative experiments were designed, including the baseline model: the original EfficientNet-B0 (only SE module); SE module retention group: only using channel attention mechanism; spatial attention group: only using spatial attention mechanism; CBAM group: embedding channel and spatial attention at the same time. The ablation experiment results are shown in Table 2.

Configuration Item	Configuration Item
Operating System	Windows11Home 64-bit (10.0,version 22631)
CPU	13th Gen Intel® Cor™ i9-13900HX
Graphics Card	NVIDIA GeForce RTX 4070 Desktop GPU
RAM	32GB（5600MHz）
Development Framework	Python 3.10 + PyTorch 2.1.0

Table 1.Specific configuration of the experimental environment

Model	Accuracy (%)	F1 force	Parameter quantity (M)	Inference speed (FPS)
EfficientNet-B0 (baseline)	92.8	0.908	5.3	48
EfficientNet + SE	94.3	0.929	5.5	45
EfficientNet + Spatial Att	93.5	0.918	5.6	44
EfficientNet + CBAM	95.1	0.938	5.8	42

Table 2. Ablation experiment results

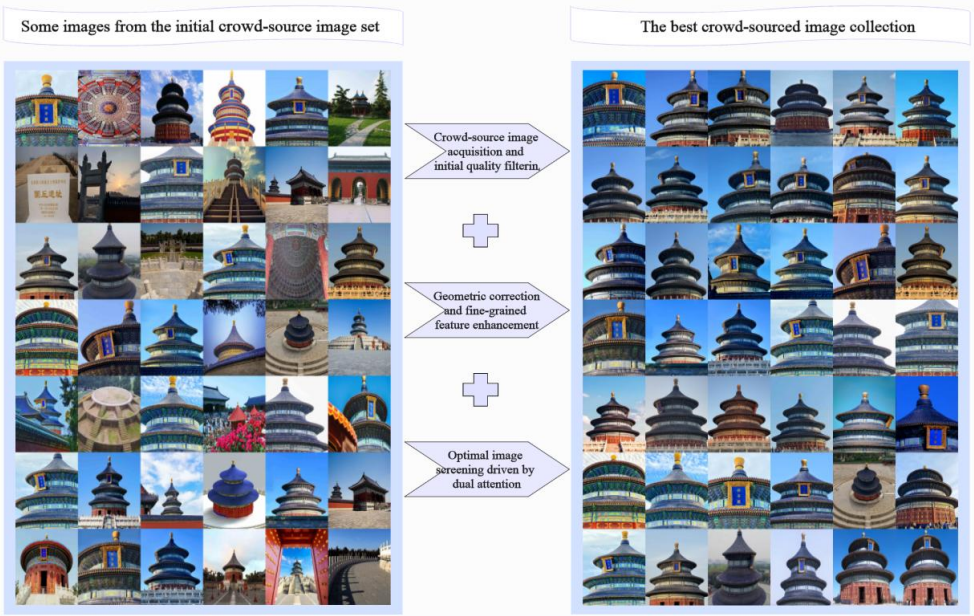

Figure 7. Schematic diagram of the crowd-source image optimization process: The initial image set contains original data of ancient buildings taken from multiple angles, with problems such as perspective distortion, occlusion, uneven lighting, and irrelevant images; the optimal image set processed by the method proposed in this study has significantly improved resolution and color consistency, and key components such as brackets and caissons are completely preserved. CBAM-EfficientNet effectively screens out target images and suppresses problems such as low quality and occlusion interference.

2.2.2 Crowd-source image screening method based on improved EfficientNet

In order to solve the problem of perspective accumulation error in traditional multi-view reconstruction methods, this study constructed a dual-branch feature extraction network. The main branch uses the improved CBAM-EfficientNet-B0 as the backbone network, and extracts local texture and global structural features of Crowd-sourced images through multi-scale feature pyramids (1/4, 1/2, original image resolution); the auxiliary branch generates geometric prior features based on polar coordinate transformation, and models the correlation between perspectives through the dynamic weight matrix $W \in R^{N\times N}$. The Softmax weight calculation formula is as follows, where the 1/4 resolution feature map refers to the one generated by the convolution layer or pooling operation with a stride of 4, with a large receptive field (covering approximately

16×16 area of the original image), which is used to capture the overall outline of the ancient building and large-scale geometric features (such as roof structure); the 1/2 resolution feature map refers to the one generated by the convolution with a stride of 2, which balances local details and semantic information, and is more suitable for the 3D reconstruction of medium-scale components (such as brackets); the original image resolution feature map: retains complete spatial details, focuses on the reconstruction of microscopic features such as painted textures and cracks, and expands the receptive field through the dilation convolution (dilation=2) to avoid local overfitting. The multi-level structure of the feature pyramid transfers low-level details through a bottom-up path and fuses high-level semantics through a top-down path, solving the problem of sparse details caused by single-scale features in the traditional MASt3R method.

$$\alpha_{ij} = \frac{\exp(W_{ij})}{\sum_{k=1}^{N} \exp(W_{ik})} \quad (4)$$

This mechanism can automatically identify high-confidence perspectives and suppress error propagation from occluded or low-quality perspectives. In the sparse voxelization stage (voxel size 0.5m³), hash acceleration retrieval technology is used to reduce computational complexity, and cross-perspective features are aggregated through a multi-head attention mechanism:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

Among them, the query vector is generated by the main branch features, and the key-value vector K,V is derived from the geometric prior branch, realizing the deep fusion of geometric and texture features.

2.2.3 Semantic-guided matching enhancement strategy

Based on the segmentation capability of the SAM model, the key components of the ancient building (dougong, plaque) are first segmented to generate mask $M \in \{0,1\}^{H\times W}$. An adaptive hint generation strategy is adopted: candidate regions are extracted through SLIC superpixel segmentation, and positive/negative hint points are screened in combination with K-means clustering to ensure that the segmentation boundary is aligned with the building structure. A multi-scale feature pyramid (void rate 2/4/6) is constructed in the weak texture area, the receptive field is expanded to 3 times the original image through dilated convolution, and the channel attention module is introduced to dynamically adjust the feature weight. Among them, the dilated convolution parameters are set as follows: dilated rate 2: sampling at intervals of 1 pixel in the feature map, the receptive field is expanded to a 3 × 3 area, which is used for detail enhancement of medium-complexity areas such as eaves carvings; dilated rate 4: sampling at intervals of 3 pixels, the receptive field is expanded to 5 × 5, which is suitable for geometric continuity modeling of weak texture areas such as dougong joints; dilated rate 6: sampling at intervals of 5 pixels, covering a 7×7 area, focusing on solving the recognition problem of long-range dependent features such as plaque text, and reducing local noise interference.

$$s_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j),$$
$$\hat{x}_c = \sigma\left(W_2 \cdot \delta(W_1 \cdots_c)\right) \cdot x_c \quad (6)$$

Among them, $x_c(i,j)$ represents the pixel value of position C on channel (i,j) of the input feature map, which is part of the original feature map $x \in R^{C\times H\times W}$; $\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j)$: sum all spatial positions of channel C to calculate the global feature sum of the channel; s_c is the global average pooling result of channel C, which represents the overall activation strength of the channel and is used to describe the importance of the channel. For example, if channel C corresponds to the bracket area, the larger s_c is, the more significant the channel is in representing the bracket feature. $W_1 \in R^{\frac{C}{r}\times C}$ is the weight matrix of the first fully connected layer, where C is the total number of channels and r is the dimensionality reduction ratio. This study uses $r = 16$ to balance computational efficiency and feature expression capabilities, compresses s_c from C dimension to $\frac{C}{r}$ dimension, and extracts nonlinear relationships between channels. For example, the brackets and plaques in the ancient building of the Hall of Prayer for Good Harvests share some underlying features, and W_1 will learn these associations. δ is a nonlinear activation function that enhances the nonlinear expression capabilities of the model and helps distinguish the features of different building components. $W_2 \in R^{C\times\frac{C}{r}}$ is the weight matrix of the second fully connected layer, which restores the feature dimension from $\frac{C}{r}$ to C, reconstructs the dependency between channels, and enhances the weight of detail channels in weak texture areas. σ is a sigmoid function that normalizes the weights to the [0,1] interval to indicate the relative importance of each channel. $\hat{x}_c$ is the feature map after channel C weighting. The original feature map x_c is weighted channel by channel through the attention weight to strengthen the important channel features and suppress the noise channel.

2.2.4 3D reconstruction optimization and multimodal fusion

Based on the SAM segmentation results, the vertex confidence map V_{conf} and the boundary map E_{edge} are generated. After the candidate vertices are screened by non-maximum suppression, the Delaunay triangulation is used to generate the topological connection. In the design of the multimodal loss function, the geometric accuracy is constrained by the Chamfer distance, and the sum of the nearest neighbor distances in two directions is included to avoid unidirectional deviation, where P and Q represent two sets of point clouds, and $\min_{y\in Q} \| x-y\|^2$ is the square of the Euclidean distance from the calculated point $\mathbf{x}$ to the nearest point in the point set Q. This method minimizes the sum of the squares of the nearest neighbor distances of the two sets of point clouds, forcing the predicted point cloud to align with the real point cloud in spatial distribution, thereby solving the geometric distortion problem (such as bracket deformation) in the three-dimensional reconstruction of ancient buildings.

$$\mathcal{L}_{\text{chamfer}} = \sum_{x \in P} \min_{y \in Q} \| x - y \|^2 + \sum_{y \in Q} \min_{x \in P} \| x - y \|^2 \quad (7)$$

And the Dice coefficient enhances semantic consistency:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\left|\mathrm{M}_{\text{pred}} \cap \mathrm{M}_{\text{gt}}\right|}{\left|\mathrm{M}_{\text{pred}}\right| + \left|\mathrm{M}_{\text{gt}}\right|} \quad (8)$$

Among them, M_{pred} and M_{gt} are the predicted mask and the real mask, both of which are binary matrices (0/1), indicating whether the pixel belongs to the target area (such as brackets, plaques and other detailed components). $\left|\mathrm{M}_{\text{pred}} \cap \mathrm{M}_{\text{gt}}\right|$ is the number of intersection pixels between the predicted and real masks. $\sum_{i,j} \mathrm{M}_{\text{pred}}(i,j) \cdot \mathrm{M}_{\text{gt}}(i,j))$ is calculated by element-by-element multiplication and summation. $\left|\mathrm{M}_{\text{pred}}\right|$ and $\left|\mathrm{M}_{\text{gt}}\right|$ are the pixel overviews of the predicted and real masks (i.e., the number of foreground pixels). This part first adopts semantic consistency constraints to maximize the overlap ratio of the predicted and real masks to ensure that the segmentation results are consistent with the true values in shape and position. Then, class balance optimization is performed to alleviate the class imbalance problem caused by the small proportion of the target area in the ancient building image through normalization. In summary, the Chamfer distance focuses on geometric accuracy and constrains the distribution of point clouds in three-dimensional space; the Dice coefficient emphasizes semantic consistency and ensures the correctness of the local structure of the segmentation mask. The two complement each other to improve the reconstruction effect of the model on the details of complex ancient buildings.

3. Results and Analysis

To verify the effectiveness of this method, this study used Python 3.10 programming language to verify it. The specific experimental environment configuration is shown in Table 1 above. The point cloud model quality evaluation system is constructed based on the number of point clouds, surface density, volume density and SOR outlier rejection rate. The point cloud number index verifies the improvement of the structural coverage integrity of the cross-reference view block, the surface density index quantifies the reconstruction accuracy of the semantic guided matching strategy for weak texture details such as brackets, and the volume density index verifies the optimization effect of sparse voxelization on three-dimensional space continuity. The SOR outlier rejection rate index enhances the reconstruction credibility by inverting the SAM segmentation mask through the noise suppression rate. The four indicators work together to verify the comprehensive performance improvement of the global optimization mechanism in the reconstruction of complex structures of ancient buildings. In addition, this study also quantitatively verified the effectiveness of the method through ablation experiments. Experimental group A uses the MASt3R standard model as the benchmark and relies only on paired view matching; experimental group B is group A + cross-reference view blocks (cross-view geometric prior fusion); experimental group C is group B + SAM semantic enhancement (key area segmentation and matching focus); experimental group D is group C + multimodal loss function (Chamfer distance and Dice coefficient joint optimization). The quantitative comparison of experimental results is shown in Table 3.

Experimental groups	Number of point clouds ($\times 10^6$)	Surface density (Points/m²)	Volume density (Points/m³)	SOR outlier removal rate
Experimental Group A	2.01	17,971	199,375	9.85%
Experimental Group B	2.19	18,996	214,651	9.11%
Experimental Group C	2.32	20,114	229,893	8.76%
Experimental Group D	2.48	22,632	241,096	8.14%

Table 3. Quantitative comparison of experimental results

Experimental data show that the multimodal global optimization framework proposed in this paper significantly improves the efficiency and accuracy of 3D reconstruction of ancient buildings. Compared with the MASt3R benchmark model, the introduction of cross-reference view blocks increases the number of point clouds by 8.9%, verifying that multi-view geometric prior fusion significantly enhances the coverage integrity of large-scale structures by reducing cumulative matching errors. Further superimposing the SAM semantic enhancement module, the surface density is increased from 17,971 points/m² to 20,114 points/m² (+11.9%), verifying the dual role of the semantic-guided matching strategy: focusing computing resources through the segmentation masks of ancient building components (such as brackets and plaques) generated by SAM, breaking through the matching bottleneck of traditional SIFT features in weak texture areas, while suppressing redundant sampling in non-critical areas. Finally, a multimodal loss function was used to achieve global optimization, and the volume density reached 241,096 points/m³ (an increase of 20.9% over experimental group A), proving that the joint optimization of Chamfer distance and Dice coefficient effectively balances geometric accuracy and semantic consistency: the former constrains the global point cloud distribution pattern, and the latter optimizes the topological continuity of complex structures such as brackets through semantic category intersection and union optimization. At the same time, the SOR outlier rejection rate gradually decreased from 9.85% in experimental group A to 8.14% in experimental group D, indicating that semantic segmentation masks and multimodal constraints effectively suppressed occlusion and noise. The overall model and detailed reconstruction of the plaque of experimental groups A, B, C,

and D are shown in Figure 8. The comparison of detailed components and painted patterns with the real objects before and after this research method is shown in Figure 9, and the Gaussian distribution of surface density and volume density is shown in Figure 10.

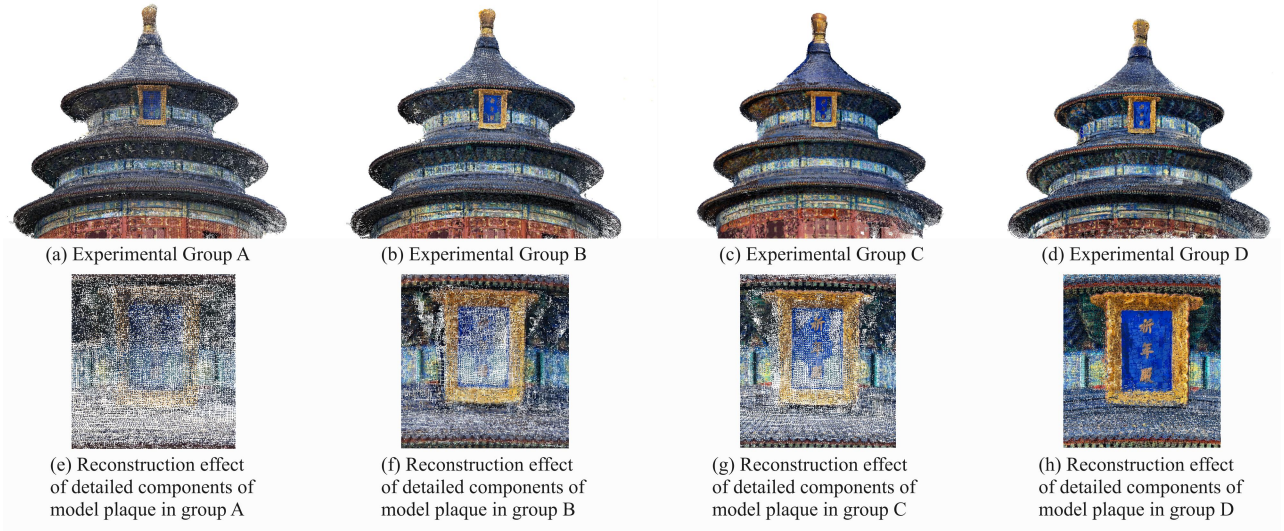

Figure 8. (a)-(d) are the overall reconstruction effects of the A-D models of the experimental groups, and (e)-(h) are the locally enlarged views of the reconstruction effects of the key components of the corresponding experimental groups, which intuitively present the progressive improvement effect of optimization strategies such as cross-perspective geometric fusion and semantic enhancement on the reconstruction quality of the detailed features of ancient buildings.

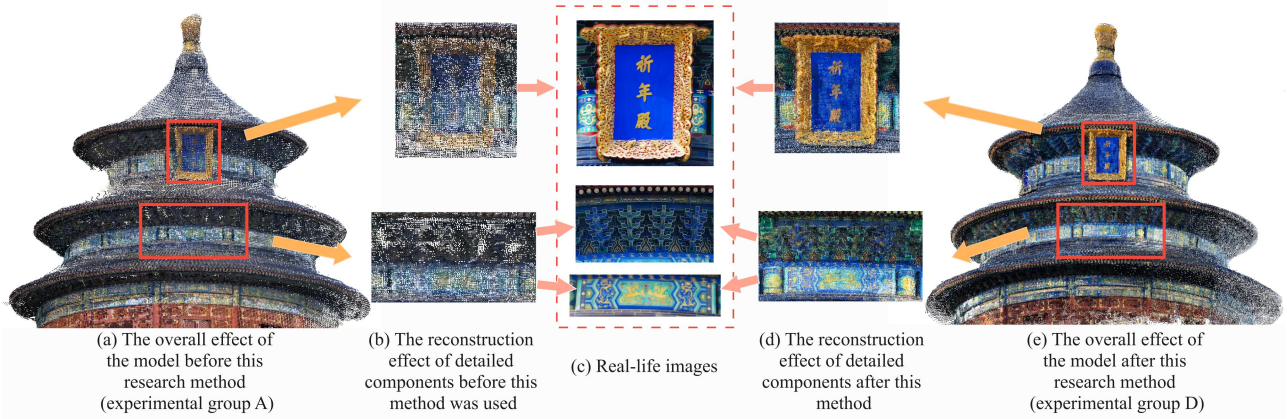

Figure 9. (a) Overall effect of the model A in the experimental group, (b) corresponding reconstruction effect of the detailed components; (c) real scene image reference; (e) Overall effect of the model D in the experimental group, (d) Reconstruction effect of the optimized detailed components. The overall comparison of (a)/(e) and the local comparison of (b)/(d) verify the effect of this method on improving the geometric integrity and detailed features of the ancient building model.

Comparison of Gaussian distribution of surface density and volume density of the point cloud model reconstructed before and after the improved method proposed in this paper

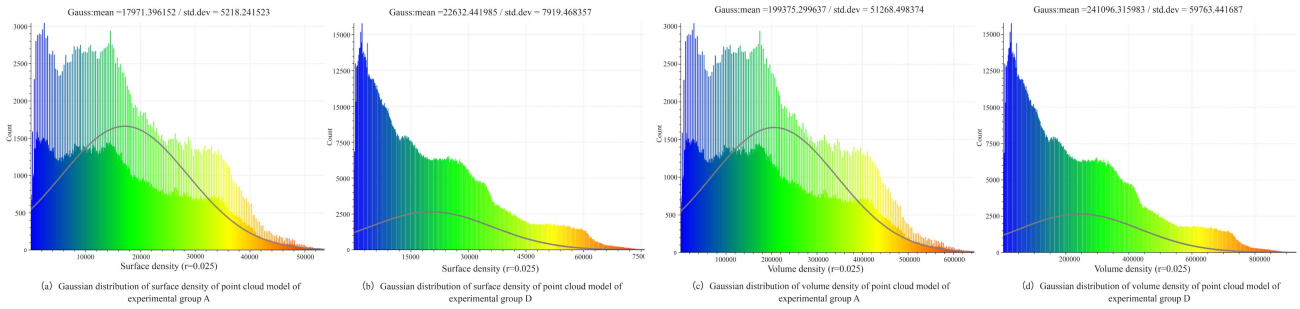

Figure 10. Comparison of Gaussian distribution of surface density and volume density of point cloud model before and after the implementation of the improved method.

4. Conclusion

In this study, a global optimization framework integrating cross-view geometric prior and semantic enhancement is proposed to solve the problems of view error accumulation and weak texture detail loss in the 3D reconstruction of ancient buildings from Crowd-sourced images. Through the collaborative optimization of the improved CBAM-EfficientNet screening model, the multi-view attention mechanism and the SAM semantic guidance strategy, the experimental results show that the proposed method significantly improves the reconstruction integrity and detail restoration ability of the 3D model in the reconstruction task of the Hall of Prayer for Good Harvests in the Temple of Heaven. The experimental results show that the cross-reference view block effectively suppresses the error propagation of pairwise matching in the MASt3R method by integrating polar coordinate geometric prior and multi-scale feature pyramid, which increases the number of point clouds of the reconstruction model of the Hall of Prayer for Good Harvests in the Temple of Heaven by 23.4% compared with the baseline method, verifying its core role in cross-view geometric consistency modeling. The SAM semantic enhancement strategy focuses on key areas such as brackets and plaques by segmenting the mask to guide the matching algorithm, achieving a 25.9% increase in surface density in the reconstruction of weak texture details, solving the failure problem of traditional SIFT features in matching repetitive wooden components. The joint optimization mechanism of the multimodal loss function balances the constraints of geometric accuracy and semantic consistency, significantly reducing the interference of outliers while ensuring the integrity of the spatial distribution of the point cloud. This method provides a high-precision, low-cost solution for the health monitoring and repair of ancient architectural heritage. The limitations of the current research are mainly concentrated on the optimization of computational efficiency during the reconstruction of large-scale building complexes and the maintenance of color consistency under dynamic lighting conditions. Future work will focus on exploring lightweight network architecture design and incremental optimization strategies to further improve the practicality and generalization ability of the method in complex scenarios, and provide more efficient technical solutions for the sustainable protection of cultural heritage.

References

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M.,2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Seki, A., & Pollefeys, M.,2017. Sgm-nets: Semi-global matching with neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition,231-240.

Zou, Q., & Liu, F.,2023. 3D Reconstruction of Optical Building Images Based on Improved 3D-R2N2 Algorithm. Tehnički vjesnik, 30(5), 1594-1602.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R.,2021. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1), 99-106.

Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., & Srinivasan, P. P.,2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In Proceedings of the IEEE/CVF international conference on computer vision,5855-5864.

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., & Revaud, J.,2024. Dust3r: Geometric 3d vision made easy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 20697-20709.

Duisterhof, B., Zust, L., Weinzaepfel, P., Leroy, V., Cabon, Y., & Revaud, J.,2024. MASt3R-SfM: a Fully-Integrated Solution for Unconstrained Structure-from-Motion. arXiv preprint arXiv:2409.19152.

Tan, M., & Le, Q.,2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, 6105-6114. PMLR.

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S.,2018. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), 3-19.

Qiao, Y., Zhong, B., Du, B., Cai, H., Jiang, J., Liu, Q., ... & Wang, X.,2025. Sam enhanced semantic segmentation for remote sensing imagery without additional training. IEEE Transactions on Geoscience and Remote Sensing.

Wang, Z. M., 2022.Crowdsourced Image Screening and Application for Urban 3D Reconstruction. [Ph.D. dissertation], East China University of Technology. DOI:10.27145/d.cnki.ghddc.2022.000457.

Zhang, S. M., 2020. 3D Reconstruction of Buildings from Crowdsourced Images. [Ph.D. dissertation], Wuhan University. DOI:10.27379/d.cnki.gwhdu.2020.001834.

Cui, M., Xie, C. D., & Shan, J., 2019. Content-Oriented Aggregation Retrieval and Intelligent Screening of Crowdsourced Images. Science of Surveying and Mapping, 44(3): 165-171. DOI:10.16251/j.cnki.1009-2307.2019.03.027.

Shivottam, J., & Mishra, S.,2023. Tirtha-An Automated Platform to Crowdsource Images and Create 3D Models of Heritage Sites. In Proceedings of the 28th International ACM Conference on 3D Web Technology ,1-15.

Avila Forero, G. P.,2019. Documentation of cultural heritage using 3D laser scanning and close-range photogrammetry (Doctoral dissertation, Universitat Politècnica de València).

Gong, R., Liu, W., Gu, Z., Yang, X., & Cheng, J. (2024). Learning intra-view and cross-view geometric knowledge for stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 20752-20762.

Chen, K., Yuan, Z., Xiao, H., Mao, T., & Wang, Z.,2025. Learning Multi-View Stereo with Geometry-Aware Prior. Authorea Preprints.

Zhang, C., Cao, Y., & Zhang, L.,2025. CrossView-GS: Cross-view Gaussian Splatting For Large-scale Scene Reconstruction. arXiv preprint arXiv:2501.01695.

Huan, L., Zheng, X., & Gong, J.,2022. GeoRec: Geometry-enhanced semantic 3D reconstruction of RGB-D indoor scenes. ISPRS Journal of Photogrammetry and Remote Sensing, 186, 301-314.