The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIX-B4-2026
XXV ISPRS Congress 2026 "From Imagery to Understanding", Commission IV, 4–11 July 2026, Toronto, Canada

An Automated Recognition Framework for Surface Deterioration Features of Stone Sculptural Artifacts in the Yungang Grottoes Based on Deep Learning

Zhenyu Lin [1,2], Yuan Cheng [1,3*], Hanling Li [1], Jizhong Huang [1,3], Guanpeng Xia [1], Yue Zhang [1,3],Wanfu Wang [4], Bo Ning [5], Hongbin Yan [5]

[1] Institute for the Conservation of Cultural Heritage, School of Cultural Heritage and Information Management, Shanghai University, Shanghai, China, 200444 – shulzy@shu.edu.cn, chengyuan@shu.edu.cn, lihanling@shu.edu.cn, hjizhong@shu.edu.cn, xiaguanpeng@shu.edu.cn, 2019zhangy@shu.edu.cn
[2] School of Materials Science and Engineering, Shanghai University, Shanghai, China, 200444 – shulzy@shu.edu.cn
[3] Key Laboratory of Silicate Cultural Relics Conservation (Shanghai University), Ministry of Education – chengyuan@shu.edu.cn, hjizhong@shu.edu.cn, 2019zhangy@shu.edu.cn
[4] National Research Center for Conservation of Ancient Wall Paintings and Earthen Sites, Dunhuang Academy, Dunhuang, Gansu, China, 736200 – wwanfu@hotmail.com
[5] Yungang Research Institute, Datong, Shanxi, China, 037034 – ningbo007@163.com, 435135458@qq.com

*Corresponding Author: chengyuan@shu.edu.cn

Keywords: Stone Sculptural Artifacts, Yungang Grottoes, Deep Learning, Deterioration Features, Automated Recognition.

Abstract

Grotto temples are valuable World Cultural Heritage sites and provide important physical evidence for the study of ancient civilizations. However, prolonged exposure to natural and human factors has caused surface deterioration in these grottoes. Conventional deterioration identification methods have shown clear limitations in quantification, automation, and large-scale application. To address this issue, this study focused on the Yungang Grottoes, a World Cultural Heritage site. We collected and processed multisource historical and contemporary images and built a high-quality annotated dataset covering three typical deterioration features: peeling, crack, and human factor. We then developed an improved YOLO11 model and introduced a spatial attention mechanism to dynamically direct the model toward critical deteriorated regions, thereby improving detection accuracy. The experiments showed that the model achieved recognition confidence levels of 88.53%, 86.42%, and 84.78% for the three typical deterioration features, respectively. These results provide a new technical approach for automated deterioration identification of stone cultural relics in grotto temples.

1. Introduction

Grotto temples preserve valuable cultural remains such as murals, painted sculptures, stone statues, and sutra fragments. Long-term natural weathering and human activities have caused severe damage to many grotto relics, threatening both their structural stability and cultural value (Guo and Jiang, 2015). Therefore, investigating the surface deterioration of stone sculptures is important for understanding deterioration mechanisms and supporting conservation.

Due to the large-scale of grotto temples and the complexity of their environments, traditional survey methods are inadequate for large-scale, cross scale, and long-term monitoring (Chalikakis et al., 2011). Although 3D digital modeling and machine learning have advanced deterioration analysis, short-term assessments often fail to reflect actual deterioration patterns, and conventional machine learning usually relies on manually designed features (Fehér and Török, 2022; Zhang et al., 2024). Deep learning therefore provides a promising solution for automated deterioration identification from continuously accumulated multisource image data (Zhan et al., 2025).

Although deep learning has shown promise in cultural heritage conservation, its application to complex deterioration scenes such as the Yungang Grottoes remains challenging (Saravanan and Bhaskar, 2024). Existing studies mainly focus on standardized datasets or downstream restoration tasks, while early-stage deterioration identification under multiform deterioration, large-scale variation, blurred edges, and irregular contours remains insufficiently addressed (Karimi et al., 2025). In such cases, general recognition models often struggle to balance contextual perception and detail preservation, leading to missed fine cracks, inaccurate localization of large damaged areas, and limited multiscale boundary modeling (Fu et al., 2024; Mayya et al., 2025).

To address these challenges, this study proposes a deep learning framework for surface deterioration identification in Yungang stone carvings. A dedicated dataset was constructed from historical and contemporary multisource images, covering three typical deterioration categories: peeling, crack, and human factor. Considering the need for both object localization and fine-grained segmentation under irregular shapes and large-scale variation, YOLO11x-seg was adopted as the baseline for its balance of accuracy, efficiency, and instance-level segmentation capability. A spatial attention mechanism was further introduced to enhance attention to critical deterioration regions and improve recognition performance. Experiments on the test set verified the effectiveness of the proposed method.

2. Data

As shown in Figure 1, the Yungang Grottoes are located on the southern slope of Wuzhou Mountain, about 16 km southwest of Datong, Shanxi Province. Focusing on the deterioration of stone sculptures, this study compiled a multisource image dataset spanning more than a century. The historical dataset covers key visual records from the early 20th century to the present, including illustrated catalogs and archaeological survey reports, which were digitized using high-resolution scanning under controlled conditions to preserve image fidelity and fine surface details.

The second component consisted of contemporary multisource images. Using photographic equipment such as DSLR cameras, smartphones, and drones, we documented the surface deterioration of the Yungang Grottoes and acquired high-precision, high-resolution contemporary images from multiple sources. Together, the contemporary and historical images constituted a multisource image dataset for deterioration in the stone sculptures of the Yungang Grottoes.

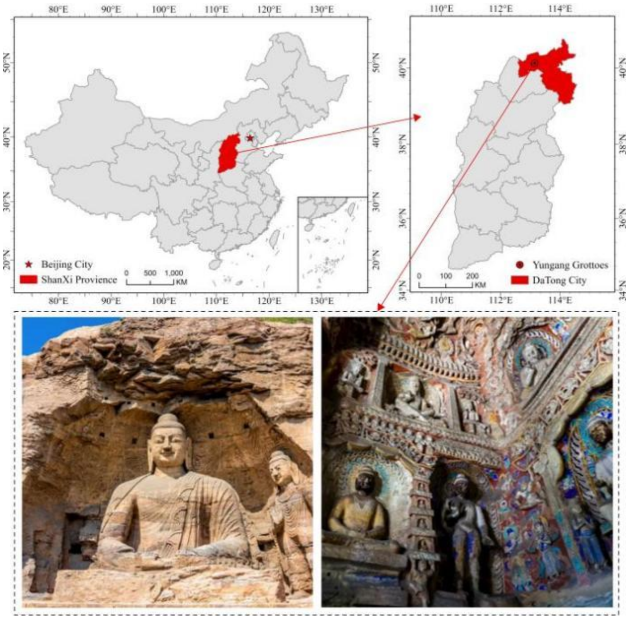

Figure 1. Geographic location and overview of the Yungang Grottoes.

3. Methodology

Surface deterioration of grotto carvings mainly includes peeling, crack, and human factor, as shown in Figure 2. To support accurate identification, this study analyzed the visual characteristics of these three categories and designed corresponding annotation strategies for pixel-level segmentation, thereby constructing a high-quality annotated dataset. To improve the efficiency of large-scale labeling, a semi-automatic annotation algorithm was developed to generate preliminary annotations that were refined with limited manual correction. In addition, target-centered cropping and multi-scale subimage extraction were applied to generate diverse training samples, enhancing the model's learning and generalization across different deterioration scales.

3.1 Image Preprocessing

Analysis of the multisource stone carving images showed that many samples were affected by Gaussian and salt-and-pepper noise, which reduced image quality and hindered deterioration analysis. Therefore, targeted denoising and enhancement methods were applied during preprocessing according to the noise characteristics of different images.

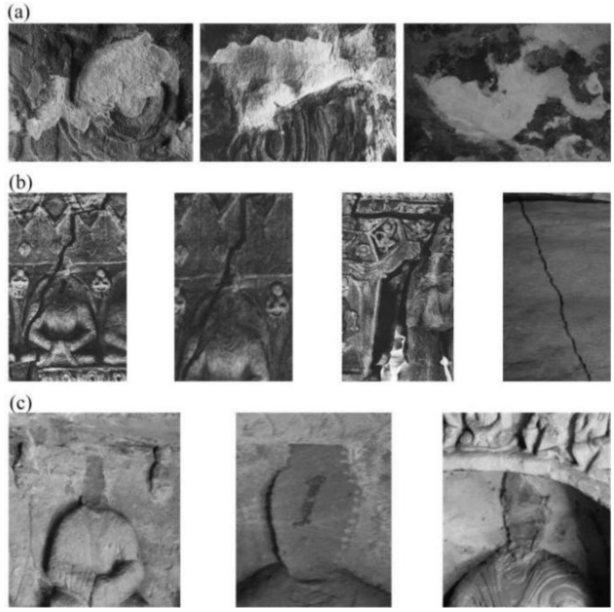

Figure 2. Three typical surface deterioration features: (a) peeling, (b) crack, and (c) human factor.

3.1.1 Histogram Equalization: After denoising, image enhancement was performed using histogram equalization, logarithmic transformation, gamma correction, and dark channel dehazing according to the characteristics of different images. Histogram equalization was used to improve grayscale distribution through gray-level mapping (Jha et al., 2023).

$$S_k = T(r_k) = \sum_{j=0}^{k} p_r(r_j) = \sum_{j=0}^{k} \frac{n_j}{n} \tag{1}$$

where → r_k = gray level k of the input image
$p_r(r_j)$ = probability of gray level r_j
n_j = number of pixels with gray level r_j
n = total number of pixels in the image
S_k = gray level after transformation
$T(r_k)$ = transformation function

As shown in Figure 3, histogram equalization expanded the grayscale distribution and enhanced overall image contrast, making carving edges more prominent and facilitating subsequent deterioration analysis.

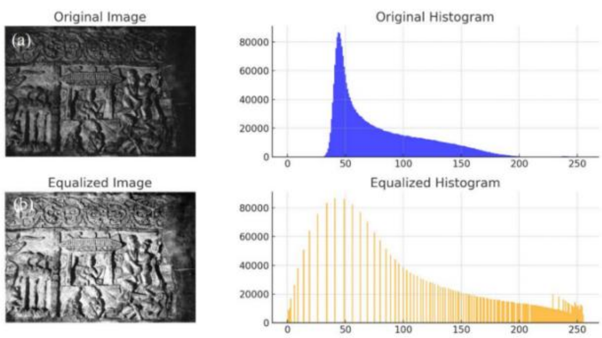

Figure 3. Results of histogram equalization for the carving image: (a) Original image, (b) Equalized image.

https://doi.org/10.5194/isprs-archives-XLIX-B4-2026-703-2026

3.1.2 Logarithmic Transformation: Logarithmic transformation was used to enhance image contrast by expanding low-gray regions and highlighting dark details while compressing high-gray regions (Jourlin and Pinoli, 2001).

$$s = c\log(1 + r) \tag{2}$$

where → r = original pixel value
s = transformed pixel value
c = constant

3.1.3 Gamma Transformation: Gamma transformation was applied as a nonlinear gray-level correction method to enhance image contrast and adjust details in low- or high-gray regions (Rahman et al., 2016). As shown in Figure 4, different γ values control whether low- or high-gray regions are emphasized.

$$g(x, y) = c\, f(x, y)^{\gamma} \tag{3}$$

where → $f(x, y)$ = pixel value at image location (x, y)
$g(x, y)$ = transformed pixel value
c, γ = constant

3.1.4 Dark Channel Dehazing: Dark channel dehazing was applied to reduce haze effects and improve image clarity based on the dark channel prior (Sabir et al., 2020).

$$J^{dark}(x) = \min_{y \in \Omega(x)} \left(\min_{c \in \{r,g,b\}} J^{c}(y) \right) \tag{4}$$

where → $J^{c}(y)$ = pixel value in channel c
$c \in \{r, g, b\}$ = color channels
$\Omega(x)$ = window centered at x

Based on the dark channel prior, haze effects were removed and image details were recovered through transmission estimation. As shown in Figures 5(a) and 5(b), dark channel dehazing effectively improved the clarity of historical carving images affected by haze, light scattering, or exposure problems.

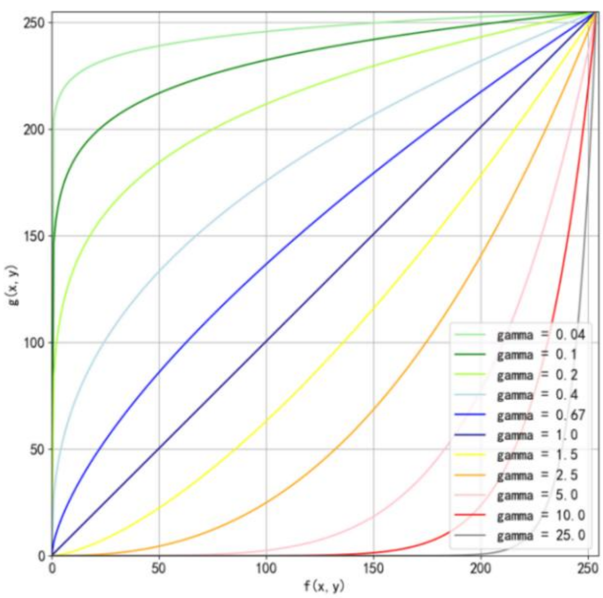

Figure 4. Gamma transformation curves for different values of γ.

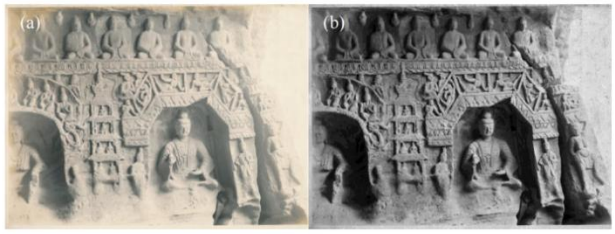

Figure 5. Comparison of the carving image before and after dark channel dehazing: (a) Original image, (b) Dehazed image.

3.2 Semi-Automatic Annotation Algorithm for Stone Carving Deterioration

Since no public dataset for stone carving deterioration is available, large-scale annotation was required. However, manual annotation is time-consuming and inefficient, and annotation quality directly affects model training. Therefore, a semi-automatic annotation algorithm was developed to improve labeling efficiency while ensuring annotation accuracy and consistency.

In the deterioration region extraction module, Canny edge detection, gray-level co-occurrence matrix texture analysis, and gray-threshold segmentation were combined for preliminary identification of deterioration regions. Morphological operations with a 7×7 rectangular structuring element were then applied to refine the segmentation results by filling small holes, smoothing contours, and removing noise. The extracted regions were converted into polygonal contours and saved as JSON annotations with three labels: peeling, crack, and human factor. Final annotations were obtained after manual review and refinement.

To verify annotation quality, an annotation inspection module was developed to visualize polygonal boundaries of deterioration regions in each image. Different categories were distinguished by color and optional text labels, including peeling, crack, and human factor. As shown in Figure 6, the annotation polygons accurately covered the target regions, including areas with complex boundaries, providing reliable data support for subsequent model training.

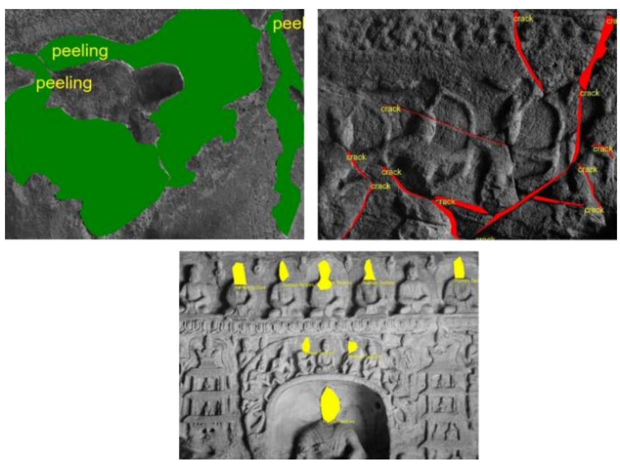

Figure 6. Visualization of annotations for three types of deterioration regions.

3.3 Automated Recognition Model for Deterioration Features

3.3.1 Automated Recognition of Deterioration Features Based on a Deep Learning Model: To identify peeling, crack, and human factor in Yungang stone carvings, this study adopted YOLO11x-seg as the baseline for efficient feature extraction and segmentation. However, its performance is still limited in complex backgrounds and irregular deterioration patterns. Therefore, CBAM (Convolutional Block Attention Module) was introduced to enhance feature representation and improve recognition performance (Woo et al., 2018). The optimized network structure is shown in Figure 7.

3.3.2 Spatial Perception Mechanism for Deterioration Recognition: As shown in Figure 8, CBAM models feature saliency through cascaded channel and spatial attention. To enhance the extraction of multiscale deterioration features, attention mechanisms were introduced at multiple network scales. Since all images were converted into single-channel grayscale images during preprocessing, only the spatial attention module was retained in the final architecture.

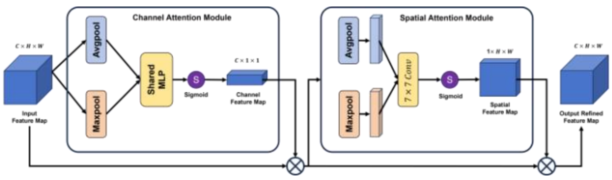

Figure 8. Convolutional Block Attention Module.

The spatial attention mechanism was introduced to guide the model toward key regions associated with peeling, crack, and human factor. By combining average and max pooling across channels, the module captured spatial saliency cues of different deterioration patterns, such as large continuous peeling regions, fine linear cracks, and human factor damage concentrated in upper image areas. A subsequent 7×7 convolution generated spatial weights to emphasize important locations and improve spatial feature representation, as shown in Figure 9.

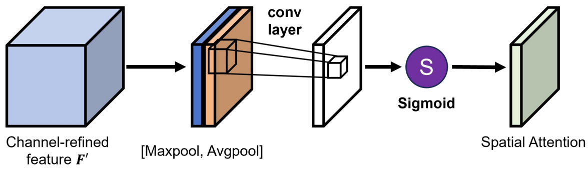

Figure 9. Spatial attention module.

3.4 Model Evaluation

The proposed model was evaluated on three deterioration categories, namely peeling, crack, and human factor, in terms of both detection and segmentation performance. Bounding boxes and segmentation masks were used to assess target localization and shape delineation, while Precision, Recall, mAP50, mAP50-95, and F1-score were adopted as quantitative metrics.

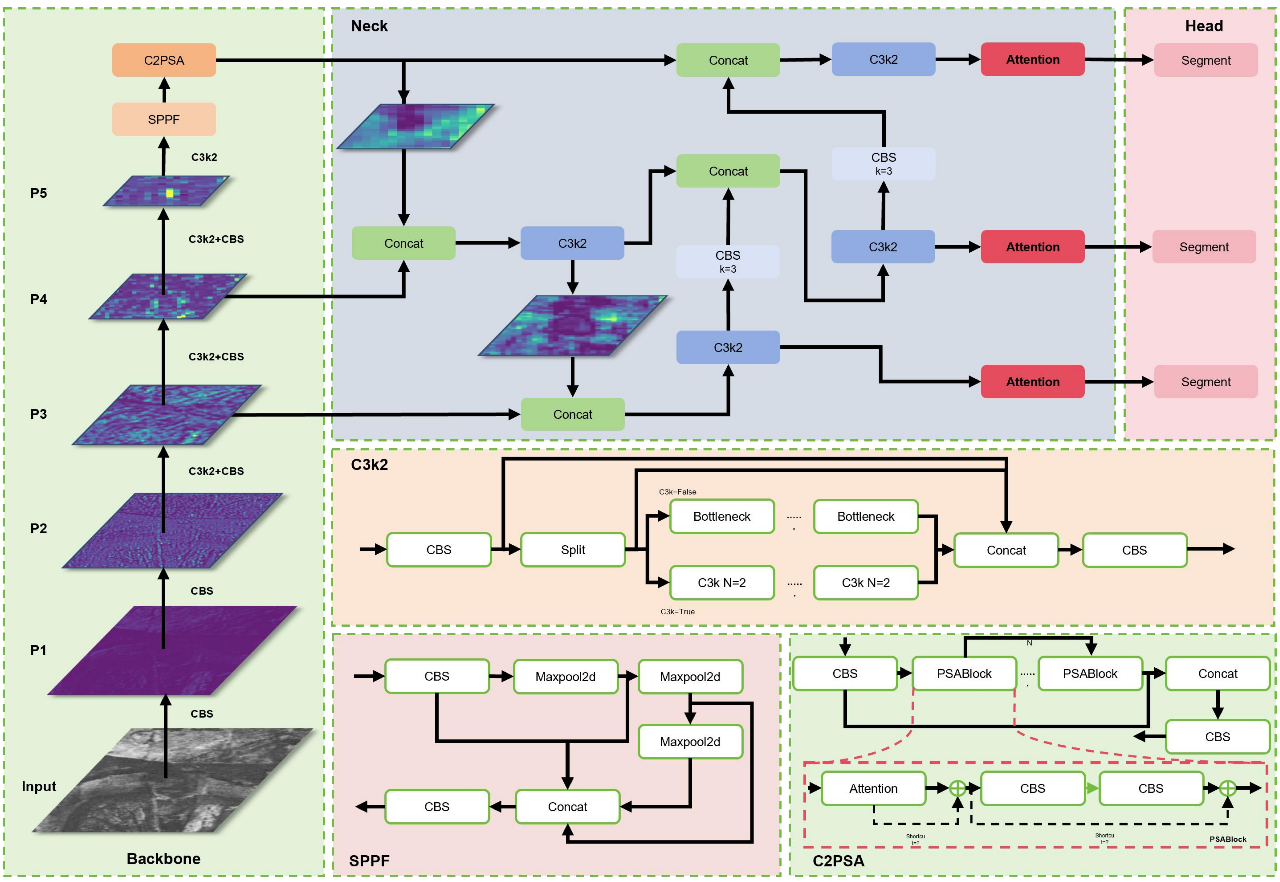

Figure 7. Architecture of the automated deterioration feature recognition model.

https://doi.org/10.5194/isprs-archives-XLIX-B4-2026-703-2026 | © Author(s) 2026. CC BY 4.0 License.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$mAP_{50} = \frac{1}{N}\sum_{i=1}^{N} AP_{50,i} \tag{7}$$

$$mAP_{50-95} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{j=1}^{K} AP_{i,j} \tag{8}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

4. Experiments and Results

4.1 Experimental Setting

Model training was performed using PyTorch with Python 3.9 on an NVIDIA GeForce RTX4090D GPU and an AMD Ryzen 9950 X16-Core Processor. The main training settings were 1000 epochs, a batch size of 16, an input size of 640×640, and an initial learning rate of 0.0015 with cosine annealing. These settings were selected to balance convergence stability, model performance, and computational cost.

4.2 Ablation Study

To evaluate the effect of the spatial attention mechanism, an ablation study was conducted on four models: YOLO11x-seg, CBAM-YOLO11x, YOLOv8x-seg, and CBAM-YOLOv8x. YOLOv8x-seg was included as a comparison model because of its architectural similarity to YOLO11x-seg, which helped reduce structural bias and clarify the contribution of spatial attention.

As shown in Table 1, the spatial attention mechanism consistently improved detection performance across all three deterioration categories, and CBAM-YOLO11x achieved the best overall results. Compared with the baseline, CBAM-YOLO11x improved Precision, Recall, mAP50, mAP50-95, and F1-score for peeling by 3.39%, 1.80%, 1.16%, 1.77%, and 4.15%, respectively; for crack by 3.15%, 4.12%, 2.77%, 5.13%, and 4.27%; and for human factor by 3.32%, 1.50%, 0.92%, 1.38%, and 2.58%. These results indicate that spatial attention helped the model focus on deterioration-related target regions, thereby improving overall recognition accuracy.

4.3 Testing Experiment

The optimized CBAM-YOLO11x-seg model was further evaluated on the independent test set to assess its practical recognition performance under complex deterioration scenarios. The model outputs included class predictions, bounding box localization results, and confidence scores, which together reflected its ability to identify different deterioration types and locate their corresponding regions. In particular, the confidence scores were used to indicate the reliability of each predicted target, providing an intuitive measure of the model ’ s recognition certainty for peeling, crack, and human factor deterioration.

As shown in Table 2, the confidence distribution demonstrates strong overall recognition performance across the three deterioration categories. Among the 1,387 valid detected targets, 47.15% achieved confidence scores higher than 90%, indicating that nearly half of the recognition results were produced with very high certainty. In addition, 46.58% of the detections fell within the confidence range of 60% – 90%, suggesting that most remaining targets were still identified with relatively reliable confidence levels. Only 6.27% of the detections had confidence scores below 60%, which accounts for a small proportion of the total results. These findings indicate that the optimized CBAM-YOLO11x-seg model maintained high predictive reliability and stable recognition performance on the test set.

Model	Type	Precision (%)	Recall (%)	mAP50 (%)	mAP 50-95 (%)	F1-Score (%)
YOLO11x-seg	peeling	93.16	82.79	90.07	70.27	87.64
YOLO11x-seg	crack	92.44	81.59	89.32	66.49	87.69
YOLO11x-seg	human factor	92.06	84.62	89.89	70.25	88.22
CBAM-YOLO11x	peeling	96.55	84.59	91.23	72.04	91.79
CBAM-YOLO11x	crack	95.59	85.71	92.09	71.62	91.96
CBAM-YOLO11x	human factor	95.38	86.12	90.81	71.63	90.80
YOLOv8x-seg	peeling	87.14	77.72	82.33	64.93	82.16
YOLOv8x-seg	crack	89.01	79.81	85.75	66.69	86.53
YOLOv8x-seg	human factor	86.50	81.60	85.03	60.21	82.51
CBAM-YOLOv8x	peeling	91.11	80.24	88.69	68.93	85.36
CBAM-YOLOv8x	crack	92.81	83.07	89.77	68.46	88.11
CBAM-YOLOv8x	human factor	93.52	82.48	87.69	65.21	87.65

Table 1. Performance Comparison of Different Models for the Recognition of Three Deterioration Features.

Type	Total Targets	≥90% Count (Ratio)	60%-90% Count (Ratio)	<60% Count (Ratio)	Average Confidence (%)
Peeling	282	216 (76.39%)	51 (18.06%)	15 (5.55%)	88.53%
Crack	847	403 (47.45%)	393 (46.43%)	51 (6.12%)	86.42%
Human Factor	258	35 (13.57%)	202 (78.29%)	21 (8.14%)	84.78%
Total	1387	654 (47.15%)	646 (46.58%)	87 (6.27%)	86.09%

Table 2. Statistics of Confidence Levels for Various Deterioration Types in the Test Set.

As shown in Figure 10, the model accurately detected and segmented the large peeling regions in Figure. 10(a) and Figure. 10(b), indicating good localization and segmentation capability for prominent peeling damage. However, in Figure. 10(c), some small peeling areas were missed. This limitation may be related to the insufficient number of training samples containing subtle peeling features, which weakened the model’s feature extraction stability for small regions and led to missed detections.

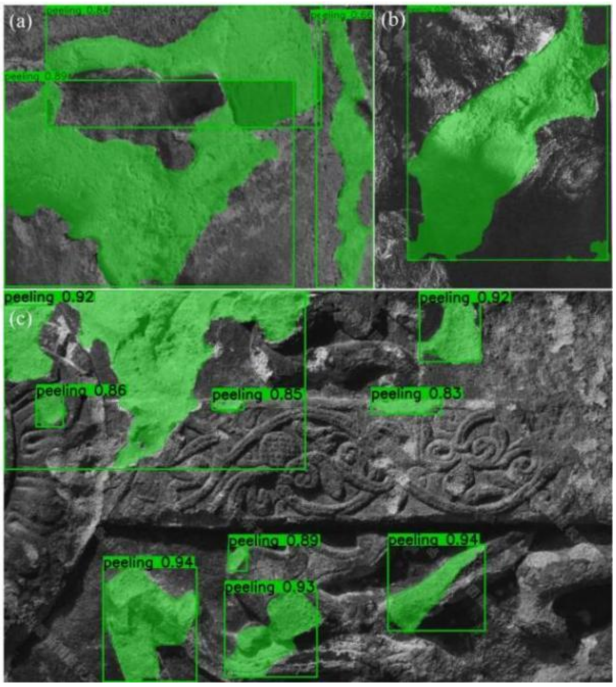

Figure 10. Specific Deterioration Feature Recognition Results for the Peeling Part of the Test Set: (a) The first large area peeling, (b) The second large-area peeling, (c) Small-area peeling.

As shown in Figure 11, most microcracks in Figure. 11(a) and (b) were correctly detected with high confidence, although a few fine cracks were still missed. Some black shadow boundaries on the carved surface were visually similar to cracks, but the model did not misclassify them, indicating good discriminative ability. In Figure. 11(b), well-repaired grouted cracks were not identified as active cracks, while the fracture-type cracks in Figure. 11(c) were accurately detected with an average confidence above 92%, demonstrating robust crack recognition across different scales.

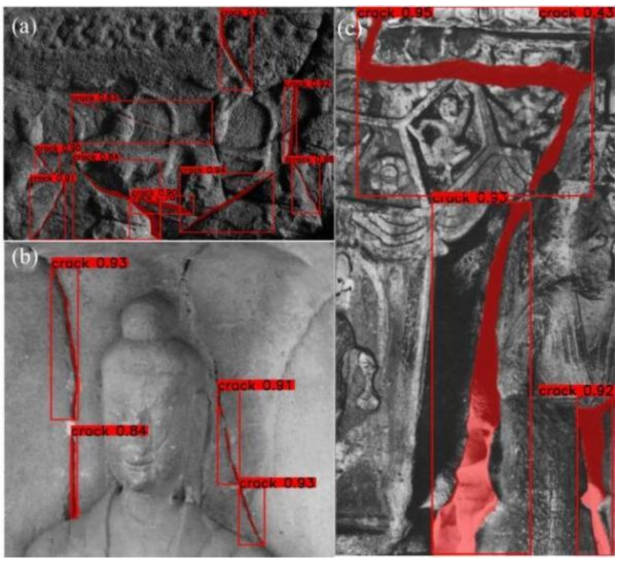

Figure 11. Specific Deterioration Feature Recognition Results for the Crack Part of the Test Set: (a) The first small-area crack, (b) The second small-area crack, (c) Large-area crack.

As shown in Figure 12, the detected human factor regions were mainly concentrated in clearly damaged areas, such as statue heads, and the bounding boxes were accurately localized. Although most detections fell within the 0.6-0.9 confidence range rather than above 0.9, the model still stably identified the main damaged regions, indicating good recognition and localization performance for non-natural deterioration features.

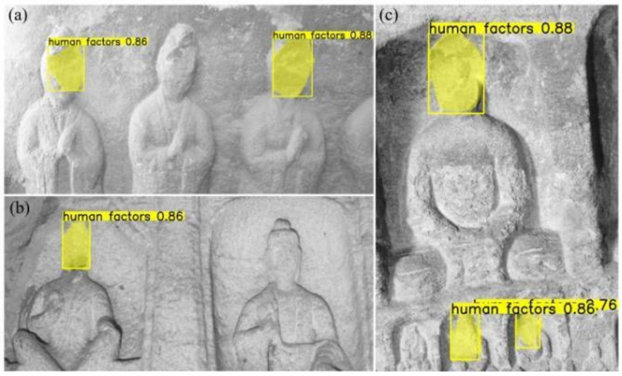

Figure 12. Specific Deterioration Feature Recognition Results for the human factor Part of the Test Set: (a) The first human factor area, (b) The second human factor area, (c) The third human factor area.

5. Conclusions

This study developed an automated framework for recognizing surface deterioration features in Yungang stone carvings by integrating a multisource image database, a semi-automatic annotation method, and a spatial-attention-enhanced YOLO11x-seg model. A total of 1,586 historical and modern images were collected, and a high-quality annotated dataset containing 1,609 images was constructed for three deterioration categories: peeling, crack, and human factor. Experimental results showed that the proposed model achieved average Precision, Recall, mAP50, mAP50-95, and F1-score values of 95.84%, 85.47%, 91.38%, 71.76%, and 91.52%, respectively, while ablation experiments confirmed that spatial attention consistently improved performance. Nevertheless, the model still missed some very small deterioration features. Future work will focus on improving small-target recognition by further integrating traditional image processing with deep learning.

Acknowledgement

This research was supported by Scientific and Technological Research Project of Cultural Relics of State Administration of Cultural Heritage (No. 2023ZCK014) and Project of Key Laboratory of Silicate Cultural Relics Conservation (Shanghai University), Ministry of Education (SCRC2024ZZ02ZD, SCRC2025ZZ04YQ). The authors thank Shanghai Technical Service Center of Science and Engineering Computing at Shanghai University for computational resources.

References

Chalikakis, K., Plagnes, V., Guerin, R., Valois, R., Bosch, F. P., 2011. Contribution of geophysical methods to karst-system exploration: an overview. *Hydrogeology Journal*, 19(6), 1169–1180. https://doi.org/10.1007/s10040-011-0746-x.

Fehér, K., Török, Á., 2022. Detecting short-term weathering of stone monuments by 3D laser scanning: lithology, wall orientation, material loss. *Journal of Cultural Heritage*, 58, 245–255. https://doi.org/10.1016/j.culher.2022.10.012.

Fu, X., Angkawisittpan, N., 2024. Detecting surface defects of heritage buildings based on deep learning. *Journal of Intelligent Systems*, 33(1), 20230048. https://doi.org/10.1515/jisys-2023-0048.

Guo, F., Jiang, G., 2015. Investigation into rock moisture and salinity regimes: implications of sandstone weathering in Yungang Grottoes, China. *Carbonates and Evaporites*, 30(1), 1–11. https://doi.org/10.1007/s13146-014-0191-8.

Jha, K., Sakhare, A., Chavhan, N., & Lokulwar, P. P. 2023. A review on image enhancement techniques using histogram equalization. *AIDE-2023 and PCES-2023*, *497*, 497. https://doi.org/10.1088/1742-6596/1019/1/012026.

Jourlin, M., Pinoli, J. C., 2001. Logarithmic image processing: the mathematical and physical framework for the representation and processing of transmitted images. *Advances in Imaging and Electron Physics*, 115, 129–196. https://doi.org/10.1016/S1076-5670(01)80095-1.

Karimi, N., Mishra, M., Lourenço, P. B., 2025. Automated surface crack detection in historical constructions with various materials using deep learning-based YOLO network. *International Journal of Architectural Heritage*, 19(5), 581–597. https://doi.org/10.1080/15583058.2024.2376177.

Mayya, A. M., Alkayem, N. F., 2025. Triple-stage crack detection in stone masonry using YOLO-ensemble, MobileNetV2U-net, and spectral clustering. *Automation in Construction*, 172, 106045. https://doi.org/10.1016/j.autcon.2025.106045.

Rahman, S., Rahman, M. M., Abdullah-Al-Wadud, M., Al-Quaderi, G. D., Shoyaib, M., 2016. An adaptive gamma correction for image enhancement. *EURASIP Journal on Image and Video Processing*, 2016(1), 35. https://doi.org/10.1186/s13640-016-0138-1.

Sabir, A., Khurshid, K., Salman, A., 2020. Segmentation-based image defogging using modified dark channel prior. *EURASIP Journal on Image and Video Processing*, 2020(1), 6. https://doi.org/10.1186/s13640-020-0493-9.

Saravanan, T. J., Bhaskar, B., 2024. Automated evaluation of degradation in stone heritage structures utilizing deep vision in synthetic and real-time environments. *Journal of Building Engineering*, 98, 111117. https://doi.org/10.1016/j.jobe.2024.111117.

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. https://doi.org/10.48550/arXiv.1807.06521.

Zhan, J., Meng, Y., Zhang, L., Li, K., Yan, F., 2025. Research on computer vision in intelligent damage monitoring of heritage conservation: the case of Yungang Cave Paintings. *npj Heritage Science*, 13(1), 50. https://doi.org/10.1038/s40494-025-01567-4.

Zhang, R., Huang, J., Cheng, Y., Zhang, Y., 2024. Machine learning based prediction for bulk porosity and static elastic modulus of Yungang Grottoes sandstone. *Heritage Science*, 12(1), 359. https://doi.org/10.1186/s40494-024-01459-z.