

Article

Object Detection of Northern Chinese Rock Art Images Using YOLOv8 with Omni-Dimensional Dynamic Convolution

Lizhong Guo ¹ and Fei Ju ^{2,*} ¹ School of Arts, Anhui University of Finance and Economics, Bengbu 233040, China; 120210096@aufe.edu.cn² College of Art Design, Nanjing Forestry University, Nanjing 210037, China

* Correspondence: jufei1986@njfu.edu.cn

Abstract

Northern Chinese rock art images are characterized by abstract structures, significant morphological variation, blurred carving traces, and complex background textures. These factors pose substantial challenges to automatic detection and classification. To address these issues, this study develops an improved object detection model, termed YOLOv8-ODConv, based on the YOLOv8 framework for rock art image recognition. The proposed model integrates Omni-Dimensional Dynamic Convolution (ODConv) into the key layers of the Backbone and Neck. This design enables convolution kernels to adapt dynamically across spatial, channel, and kernel dimensions, thereby enhancing the representation of complex carving structures, fine-grained local differences, and multi-scale features. A dedicated dataset of Northern Chinese rock art images is constructed, including three representative categories: Anthropomorphic_Face, Deer, and Human_Horse. To improve model robustness under challenging visual conditions, multiple data augmentation strategies are applied, including brightness variation, noise perturbation, and geometric transformations. Experimental results demonstrate that YOLOv8-ODConv outperforms SSD, YOLOv5, YOLOv8, and YOLOv8-Backbone-ODConv across multiple evaluation metrics. The proposed model achieves an F1-score of 95.0%, Recall of 89.3%, mAP@0.5 of 98.9%, mAP@0.5–0.95 of 84.6%, and an inference speed of 89.7 FPS. Confusion matrix analysis and visual detection results further confirm that the model effectively distinguishes typical rock art categories and maintains stable performance in complex backgrounds and fine-grained recognition tasks. These findings indicate that YOLOv8-ODConv provides an effective technical approach for the digital documentation, automatic recognition, and intelligent analysis of rock art cultural heritage.

Keywords: rock art image recognition; YOLOv8; omni-dimensional dynamic convolution (ODConv); fine-grained feature extraction; cultural heritage image analysis



Academic Editor: Ion Sandu

Received: 5 May 2026

Revised: 23 May 2026

Accepted: 26 May 2026

Published: 2 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Rock art is an important form of early human visual expression, widely distributed across the world. It records prehistoric human activities, social organization, belief systems, and perceptions of the natural environment. Compared with other archaeological remains, rock art preserves intuitive visual representations of animals, human figures, behavioral scenes, and symbolic motifs. It therefore holds irreplaceable value in archaeology, anthropology, art history, and cultural heritage studies. In recent years, research on rock art has gradually expanded from traditional image description, typological classification, and cultural interpretation to digital documentation, image analysis, and intelligent recognition.

Improving the efficiency of data organization and the depth of analysis through modern technologies has become a key research focus in this field [1–3].

China possesses abundant rock art resources, which are commonly categorized into Northern, Southwestern, and Southeastern regional systems. Northern Chinese rock art is mainly distributed in Inner Mongolia, Ningxia, Gansu, and Xinjiang. Its themes include animals, hunting, herding, anthropomorphic faces, and composite scenes, reflecting the relationship between humans, animals, and social activities in a pastoral economy [4–6]. In terms of visual representation, Northern rock art is typically created through carving and engraving, forming linear structures with abstract, symbolic, and stylized characteristics. For instance, the Deer category often exhibits relatively stable morphological patterns, whereas Anthropomorphic_Face images show greater structural variability and symbolic meaning [7,8].

Rock art images generally exhibit low contrast and incomplete structural information, which resembles challenges encountered in low-contrast object detection tasks. Previous studies have shown that weathering significantly affects the surface structure and spectral characteristics of stone cultural heritage. Different environmental conditions lead to varying degradation patterns, resulting in texture loss and weakened structural features [9]. Under such conditions, blurred boundaries and reduced feature saliency make it difficult for conventional methods to extract stable and discriminative representations, thereby degrading detection performance. In addition, relying on single-scale features is insufficient to capture the full extent of target information. Effective modeling requires the integration of local details and global contextual features to strengthen structural and boundary representation. Therefore, achieving multi-level feature fusion and enhancing key feature representation under low-contrast and complex background conditions remain fundamental challenges in computer vision [10,11].

These characteristics introduce significant difficulties for automatic detection and classification of Northern rock art images. On one hand, inter-class differences are subtle, and discriminative information is often localized in fine structural details. On the other hand, intra-class variation is high due to stylistic diversity and different preservation conditions. This problem is highly consistent with fine-grained visual recognition tasks, where small inter-class variation and large intra-class variation increase the complexity of feature modeling. Previous studies have shown that traditional methods struggle to capture such subtle differences, which significantly increases the difficulty of recognition [12].

In recent years, deep learning has achieved remarkable success in object detection and image classification. The YOLO (You Only Look Once) series, as a representative one-stage detection framework, offers a favorable balance between accuracy and efficiency and has been widely applied in industrial inspection, agricultural analysis, and remote sensing [13]. Studies have shown that incorporating attention mechanisms, improving feature fusion strategies, or optimizing detection heads can further enhance the performance of YOLOv8 in complex tasks [14]. However, for rock art images with abstract structures, discontinuous carving traces, and strong background interference, fixed convolution kernels are insufficient to adapt to diverse visual patterns across samples. Dynamic convolution provides a promising solution for modeling complex visual features. In particular, Omni-Dimensional Dynamic Convolution (ODConv) enables adaptive modulation of convolutional responses across multiple dimensions, including spatial, channel, and kernel dimensions. This allows the network to generate more task-specific feature representations based on input characteristics [15–17]. Previous studies have demonstrated the effectiveness of ODConv in tasks such as crack detection, crop disease recognition, and small object detection, especially in scenarios involving complex backgrounds, multi-scale targets, and fine-grained

variations [18]. These properties align well with the challenges of rock art images, which exhibit diverse carving structures, subtle local features, and complex background noise.

Based on the above analysis, this study focuses on Northern Chinese rock art images and selects three representative categories: Deer, Human_Horse, and Anthropomorphic_Face. The Deer category represents typical animal motifs, Human_Horse reflects composite relationships between humans and animals, and Anthropomorphic_Face embodies abstract symbolic expressions. These categories differ significantly in structural complexity, visual representation, and cultural meaning, making them suitable for evaluating key challenges in rock art recognition. To address these challenges, this study develops a YOLOv8-ODConv model by integrating ODConv into the YOLOv8 framework. Data augmentation strategies are also introduced to improve model robustness under diverse acquisition conditions and visual variations.

Unlike conventional visual tasks such as industrial inspection or natural image recognition, Northern rock art images exhibit highly abstract symbolic structures, discontinuous carving patterns, and significant weathering degradation. As a result, discriminative information often appears as weak local structural cues that do not consistently align with global semantic representations. Under such conditions, fixed convolution kernels fail to produce stable responses across different samples, limiting the ability to model key carving structures. Although dynamic convolution has been applied in tasks such as crack detection and small object recognition, these scenarios typically involve relatively continuous or regular textures. In contrast, rock art images represent highly abstract symbolic visual structures, where discriminative features are sparsely distributed and structurally complex. Therefore, developing adaptive feature modeling methods for such irregular visual patterns remains an open research problem.

The main contributions of this study are summarized as follows:

1. To address discontinuous carving lines, abstract symbols, and subtle inter-class differences in Northern rock art images, we propose a dynamic feature enhancement algorithm based on YOLOv8. Differing from prior dynamic convolution applications in regular, continuous textures, the method embeds Omni-Dimensional Dynamic Convolution (ODConv) into key Backbone and Neck layers. This enables adaptive operations across spatial, channel, and kernel dimensions, overcoming static convolution limitations and extending dynamic convolution to highly abstract symbolic visual structures.
2. A dedicated dataset for Northern rock art recognition was constructed, consisting of three categories: Deer, Human_Horse, and Anthropomorphic_Face. Data augmentation techniques such as brightness adjustment, noise simulation, and geometric transformations were applied to improve model robustness under varying illumination, image noise, and structural degradation.
3. Baseline models including SSD, YOLOv5, YOLOv8, and YOLOv8-Backbone-ODConv were used for comparison experiments. Evaluation metrics included F1-score, Recall, mean Average Precision, and FPS. Experimental results demonstrate that YOLOv8-ODConv achieves superior recognition accuracy, stability, and adaptability on complex cultural heritage images.

2. Related Work

2.1. Rock Art Research and Digital Development

Rock art research has long been grounded in archaeological and anthropological approaches. Early studies primarily focused on typological classification, regional distribution, and the interpretation of cultural meanings. For Northern Chinese rock art, previous work has systematically examined its typology and chronology, providing comprehensive

insights into its developmental stages and spatial patterns [4,5]. At the same time, research on specific themes has become increasingly detailed. Analyses of animal motifs, anthropomorphic faces, and composite scenes have not only revealed the evolving relationships between humans and animals but also reflected changes in social structures and belief systems across different historical periods [6]. At a more refined level, studies on stylized deer images and anthropomorphic representations further highlight the complexity and diversity of rock art in terms of morphological composition and symbolic expression [7,8].

With the advancement of research, traditional methods based on manual observation and expert judgment have shown clear limitations, particularly in large-scale data organization and fine-grained analysis. In recent years, international research has increasingly emphasized methodological transformation. On one hand, studies have pointed out that traditional archaeological approaches tend to be subjective and lack reproducibility when dealing with complex visual information [1]. On the other hand, the integration of physicochemical analysis and digital technologies has provided new pathways for the documentation, preservation, and analysis of rock art [2]. Building on these developments, some studies have begun to introduce machine learning techniques into rock art image recognition and classification, demonstrating the potential of intelligent analysis in this field [3].

Overall, rock art research is transitioning from a traditional paradigm dominated by qualitative analysis to a data-driven framework that integrates digital processing and intelligent analysis. This shift not only improves the efficiency of data organization and analysis but also lays an important foundation for the application of deep learning methods in automatic rock art recognition.

2.2. Cultural Heritage Image Recognition

With the rapid advancement of computer vision, deep learning methods have been increasingly applied to the recognition and analysis of cultural heritage images, providing new technical approaches for traditional image studies. In practical applications, existing research has incorporated attention mechanisms to model traditional print images, significantly improving the recognition of fine textures and key structural features [13]. In architectural heritage analysis, object detection methods that combine YOLOv8 with attention mechanisms have been used to automatically extract complex structural features, achieving high detection accuracy and robustness [14]. In addition, for traditional decorative image recognition tasks, object detection-based frameworks have also demonstrated effectiveness in both classification and localization of cultural images [19].

From the perspective of research objects, most existing methods are applied to cultural heritage images with relatively clear textures and regular structural patterns, such as prints, architectural components, and decorative motifs. These images typically exhibit well-defined boundaries and stable structures, which facilitate feature learning and category discrimination in deep learning models. From a visual analysis standpoint, cultural heritage images often present fine-grained characteristics, where inter-class differences are subtle, while intra-class variations arise from differences in scale, pose, and preservation conditions. Previous studies have pointed out that such images can be regarded as typical fine-grained visual objects, whose recognition relies on precise modeling of local structures and subtle variations [20].

In contrast, rock art images exhibit stronger abstraction and symbolic representation. Their carving structures are closely intertwined with natural rock-surface textures and are continuously affected by weathering processes. These characteristics introduce significant uncertainty in terms of contrast, structural completeness, and background complexity, thereby increasing the difficulty of automatic recognition. As a result, research on au-

tomatic recognition of rock art, as a special type of cultural heritage imagery, remains relatively limited. Existing methods still show deficiencies in handling complex backgrounds and modeling fine-grained structural features, indicating the need for further systematic investigation.

2.3. YOLOv8 and Dynamic Convolution Improvements

As one of the mainstream one-stage object detection frameworks, YOLOv8 achieves a favorable balance between detection accuracy and inference efficiency. It has been widely applied in various complex visual tasks. To meet different application requirements, researchers often enhance the model by introducing attention mechanisms, optimizing feature fusion structures, or improving detection head design.

In practical applications, related studies in road damage detection have addressed challenges such as elongated crack structures, low contrast, and complex background interference. By integrating ODConv and attention mechanisms into the YOLOv8 framework, along with improved feature extraction networks and loss functions, these approaches achieve precise crack detection. Experimental results show that such methods significantly improve detection performance in complex traffic scenarios, demonstrating the effectiveness of dynamic convolution in modeling fine-grained structures under challenging visual conditions [15,21].

In small object detection tasks, targets are typically small and weakly represented, making them highly susceptible to background noise. To address this issue, studies in infrared small object detection have introduced feature enhancement modules and region attention mechanisms. These techniques strengthen the response to weak features and suppress background interference, thereby improving detection accuracy and recall [22]. This challenge is similar to rock art images, where carving structures are small, low in contrast, and easily confused with background textures.

Dynamic convolution has also shown strong potential in agricultural vision tasks. For example, by incorporating ODConv into the YOLOv8 framework along with lightweight feature extraction networks and improved loss functions, researchers have achieved high-accuracy recognition of strawberry maturity stages, with notable improvements in mAP, Precision, and Recall [16,17,23]. Similarly, in industrial inspection tasks, ODConv combined with lightweight backbone networks and multi-scale feature enhancement has improved detection accuracy and recall for elongated structures such as anomalous pipelines, while maintaining real-time performance with reduced model complexity [18,24]. These studies indicate that dynamic convolution is well suited for tasks involving complex backgrounds, multi-scale structures, and fine-grained differences.

From a methodological perspective, dynamic convolution has emerged as an important approach for enhancing model representation capability. Unlike traditional static convolution, it adaptively adjusts kernel weights based on input features, enabling more targeted feature responses under varying visual conditions. This mechanism has been applied to diverse scenarios. For instance, in wood surface defect detection, dynamic convolution captures irregular texture patterns and improves recognition of subtle defects. In dynamic environments such as sports object detection, its adaptive nature enhances robustness under complex backgrounds. Moreover, in cultural heritage digitization and image analysis, related approaches have begun to incorporate dynamic convolution to improve modeling of complex structures and textures [25–27].

It is important to note that most existing applications of dynamic convolution focus on visual tasks with relatively continuous structures or stable texture patterns, such as industrial defect detection and agricultural recognition. In these scenarios, targets usually exhibit clear boundaries or repetitive structures. In contrast, rock art images are closer to symbolic

representations rather than natural textures. Their carving patterns are often discontinuous, partially missing, and uneven in scale. As a result, the applicability and effectiveness of dynamic convolution in such highly abstract visual objects remain insufficiently explored. This gap forms the main motivation of the present study.

2.4. Summary of Current Research and Problem Formulation

Although considerable progress has been made in rock art research and the digitalization of cultural heritage images in recent years, several key challenges remain in the automatic detection and classification of Northern Chinese rock art images.

From the perspective of research paradigms, existing studies are still largely grounded in archaeological and art historical approaches, with a strong emphasis on descriptive analysis. There is a lack of structured analytical frameworks and automated processing methods designed for large-scale image data. While some efforts have expanded into digital archiving, 3D reconstruction, and image processing, most remain limited to single visual dimensions or local feature analysis. A systematic methodology capable of supporting complex semantic modeling and category discrimination has yet to be established. In addition, traditional computer vision methods often rely on background modeling and foreground detection, using statistical features to separate targets. However, in scenarios with complex backgrounds, weak target features, and subtle structural differences, these approaches struggle to extract stable and discriminative representations, which limits their effectiveness [28].

From a technical application perspective, deep learning has been gradually introduced into the field of cultural heritage image analysis. However, its applicability to rock art, as a unique image type, remains insufficiently validated. Existing studies based on object detection models such as YOLO primarily focus on scenarios with clear structures and well-defined boundaries. In contrast, Northern rock art images typically exhibit subtle inter-class differences, high compositional similarity, and significant variation in preservation conditions. These factors pose challenges to the generalization ability and stability of models in such complex visual environments.

From a model design standpoint, most existing improvements focus on enhancing attention mechanisms or optimizing feature fusion structures, with an emphasis on general performance gains. There is a lack of task-driven design tailored to specific image characteristics. In the case of Northern rock art, visual representations are highly abstract and symbolic. Category discrimination often relies on local carving shapes and structural relationships. These features exhibit strong variability across samples, which further increases the difficulty of feature modeling.

At the application level, variations in data distribution and the presence of noisy samples also have a significant impact on detection performance. Previous studies have shown that traditional object detection models are typically developed under the assumption of static data distributions. Their performance tends to degrade when confronted with noisy samples, class imbalance, or dynamic data streams. In addition, label noise has become an important issue in image classification and recognition tasks. Existing methods attempt to address this problem by incorporating multi-view feature learning and graph-based propagation to correct noisy labels, thereby improving model robustness. However, these approaches mainly focus on data and label optimization, while offering limited improvements in the representation of complex visual features [29].

Based on the above analysis, there is a clear need for an automatic recognition method that can adapt to complex visual patterns and provide strong feature representation capabilities. To this end, this study focuses on Northern Chinese rock art images and proposes a targeted improvement to object detection models. By considering the inherent characteristics

of rock art—such as high abstraction, structural diversity, and complex backgrounds—the proposed approach aims to enhance recognition performance and applicability in complex cultural image scenarios.

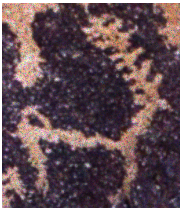
3. Materials and Methods

3.1. Dataset Construction and Preprocessing

3.1.1. Image Acquisition and Category Definition

Due to the lack of publicly available benchmark datasets for Northern Chinese rock art, this study constructs a dedicated dataset for rock art image detection and classification. The dataset contains 330 images in total, which are categorized into three representative classes based on visual themes: Anthropomorphic_Face, Deer, and Human_Horse. Table 1 provides representative example images for each category.

Table 1. Representative examples of Northern Chinese rock art image categories.

Rock Art Category	Example Images			
Anthropomorphic_Face				
Deer				
Human_Horse				

The images were primarily sourced from authoritative atlases and academic publications on Northern Chinese rock art, such as Complete Works of Chinese Art: Rock Art and Engraving Prints [30,31]. Research personnel with a background in arts and design manually extracted and categorized the images to ensure data validity. No additional transformations or modifications were applied, preserving the original appearance of the collected images. The selected images are highly representative in terms of subject matter, stylistic characteristics, and carving techniques, providing a comprehensive overview of the visual features and category distinctions of Northern Chinese rock art.

It is important to note that the acquisition of cultural heritage image data is subject to multiple practical constraints. On one hand, rock art is a non-renewable historical resource, and most sites are located in remote areas. Field data collection is therefore limited by geographical conditions, conservation regulations, and imaging constraints. On the other hand, publicly available resources are scarce, and images from different sources often vary significantly in resolution, viewpoint, and illumination conditions. As a result, it is difficult to construct a large-scale and standardized dataset. Consequently, rock art image studies generally face the challenge of limited sample size.

3.1.2. Image Preprocessing and Data Augmentation

Due to the limited availability of rock art samples and the inconsistency of acquisition conditions, the original dataset is insufficient in both scale and diversity for training deep learning models. To improve generalization and robustness, multiple data augmentation strategies are applied to expand the dataset. Data augmentation not only alleviates the problem of insufficient samples but also enhances the model's adaptability to complex environments by simulating various visual disturbances, such as illumination variation, noise perturbation, and structural degradation.

Previous studies have demonstrated that well-designed augmentation strategies and their combinations can significantly improve the performance of object detection models in complex scenarios [32]. In addition, data augmentation methods based on generative models can further enhance robustness under high-noise conditions, thereby narrowing the performance gap between deep learning models and human perception in challenging environments [33]. Under complex background conditions, combining image preprocessing with background suppression techniques can effectively reduce irrelevant texture interference and improve detection accuracy and stability [34].

To enhance the generalization capability of the dataset, four image augmentation methods—Image Darkening, Image Brightening, Salt-and-Pepper Noise, and Rotation—were applied to the collected images, resulting in a final dataset of 710 images. After augmentation, the numbers of samples in the Anthropomorphic_Face, Deer, and Human_Horse categories were expanded to 222, 196, and 292, respectively. The training, validation, and test sets were then randomly divided with a ratio of 7:2:1.

It should be emphasized that data augmentation is not only used to increase dataset size but also to simulate various visual disturbances that may occur during real-world acquisition and preservation of rock art images. These include illumination changes, noise interference, and structural degradation. By introducing diverse augmentation strategies, the model is encouraged to learn more robust and discriminative feature representations. The specific augmentation methods are described as follows.

1. Image Darkening

This operation reduces the overall brightness of the image to simulate low-light or underexposed conditions. It is implemented through a linear intensity transformation by scaling and shifting pixel values, thereby decreasing the overall grayscale intensity [35]. In practical scenarios, rock art images are often affected by environmental conditions, weathering, and differences in imaging devices, resulting in insufficient brightness or low contrast. Applying darkening augmentation helps improve the model's ability to recognize structural features under low-light conditions. Previous studies have shown that image enhancement techniques in low-illumination and low-contrast environments can effectively improve target visibility and strengthen feature representation, thereby enhancing detection performance [36].

The transformation is defined as:

$$I_{\text{darkened}} = \text{clip}(\alpha \cdot I_{\text{original}} + \beta) \quad (1)$$

Here, I_{original} denotes the pixel values of the input image, and I_{darkened} represents the pixel values after the darkening operation. The parameter α is the brightness scaling factor; when $\alpha < 1$, it linearly compresses the overall intensity and reduces image brightness. The parameter β is the brightness offset, which shifts pixel values globally and is typically set to a non-positive value to further enhance the darkening effect. The function clip constrains pixel values within a valid range (e.g., [0, 255]). In this study, $\alpha = 0.6$ and $\beta = 0$ are used to uniformly reduce image brightness while preserving structural information.

This augmentation effectively simulates low-light conditions and improves the model's robustness in recognizing carving structures and local details.

2. Image Brightening

This operation increases the overall brightness of the image to simulate overexposure or high-illumination conditions. It is also implemented through a linear intensity transformation by scaling and shifting pixel values, thereby increasing the overall intensity level. In practical data acquisition, rock art images are often affected by strong outdoor lighting, surface reflections, and camera parameter settings, which may lead to excessive brightness or local overexposure.

The transformation is defined as:

$$I_{\text{brightened}} = \text{clip}(\alpha \cdot I_{\text{original}} + \beta) \quad (2)$$

Here, I_{original} denotes the pixel values of the input image, and $I_{\text{brightened}}$ represents the pixel values after the brightening operation. The parameter α is the brightness scaling factor; when $\alpha > 1$, it amplifies pixel intensities and increases overall brightness. The parameter β is the brightness offset, which shifts pixel values globally and is typically set to a non-negative value to further enhance brightness. The function clip constrains pixel values within a valid range (e.g., $[0, 255]$). In this study, $\alpha = 1.5$ and $\beta = 0$ are used to simulate high-illumination or overexposed conditions while avoiding excessive distortion of structural information. This augmentation improves the model's stability and adaptability in recognizing rock art patterns and carving structures under bright lighting conditions.

3. Salt-and-Pepper Noise

This augmentation introduces random impulse noise to simulate pixel-level corruption during image acquisition, storage, or transmission. Salt-and-pepper noise is a typical form of discrete noise, characterized by randomly distributed high-intensity (salt) and low-intensity (pepper) pixels. In rock art images, such noise may arise due to surface weathering, uneven carving, and variations in imaging conditions, often appearing as isolated pixel artifacts in local regions. Incorporating this augmentation enhances the model's tolerance to local pixel anomalies and reduces the impact of noise on detection performance [37].

The transformation is defined as:

$$x(n, m) = \begin{cases} S & \text{with probability } p_s, \\ P & \text{with probability } p_p, \\ f(n, m) & \text{with probability } 1 - p_s - p_p \end{cases} \quad (3)$$

Here, $f(n, m)$ denotes the pixel value of the original image at location (n, m) , and $x(n, m)$ represents the pixel value after adding salt-and-pepper noise. The indices $n \in [1, N]$ and $m \in [1, M]$, where N and M denote the image dimensions. S represents the maximum pixel value corresponding to salt noise (typically white), and P represents the minimum pixel value corresponding to pepper noise (typically black). The parameters p_s and p_p denote the probabilities of salt noise and pepper noise, respectively. In this study, $p_s = 0.05$ and $p_p = 0.05$, meaning that approximately 5% of the pixels are randomly replaced with high-intensity values (white noise), and another 5% are replaced with low-intensity values (black noise). This augmentation effectively simulates local pixel anomalies caused by weathering damage, discontinuous carving, and imaging interference in rock art images, thereby improving the model's robustness to noise.

These augmentation strategies not only expand the dataset size but also simulate complex environmental conditions, such as varying illumination, local noise, and structural

degradation. This design helps the model to learn robust feature representations and mitigates the potential overfitting risks associated with the limited original dataset.

4. Rotation

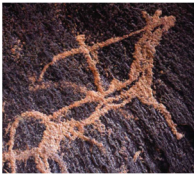
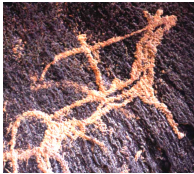
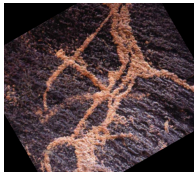
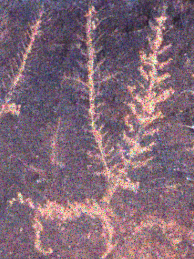
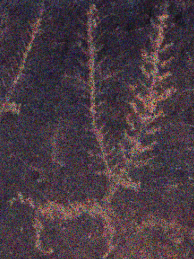
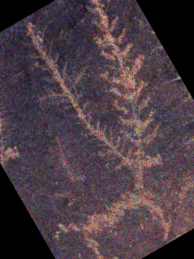
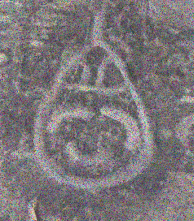
This augmentation applies a geometric transformation to rotate the image by a certain angle, simulating orientation deviations caused by camera tilt or improper device positioning during image acquisition or scanning. While preserving the original patterns and structural information, this operation changes only the spatial orientation of the image, thereby improving the model’s adaptability to different viewing angles. In the digital acquisition of rock art images, factors such as field conditions, equipment setup, and data archiving often introduce rotation deviations. To simulate such conditions, a two-dimensional rotation transformation centered at the image center is adopted.

The transformation is defined as:

$$\begin{cases} x' = (x - x_0)\cos \theta - (y - y_0)\sin \theta + x_0 \\ y' = (x - x_0)\sin \theta + (y - y_0)\cos \theta + y_0 \end{cases} \tag{4}$$

Here, (x,y) denotes the original pixel coordinates, and (x',y') represents the coordinates after rotation. (x₀,y₀) is the center of the image, and θ is the rotation angle. In this study, to preserve the main patterns and key carving lines, a random rotation angle θ ∈ [−15°, 15°] is applied. This setting simulates real-world conditions such as camera tilt during field acquisition and misalignment during scanning. By incorporating this augmentation, the model’s dependence on fixed orientations is reduced. As a result, it achieves more stable recognition of rock art shapes, patterns, and carving features under varying viewing angles, and its generalization ability is improved (Table 2).

Table 2. Data augmentation methods.

Original Image	Darkening	Brightening	Salt-and-Pepper Noise	Rotation
				
				
				

Under the above data constraints, data augmentation serves not only to increase sample size but also to enhance model generalization. By simulating illumination variation, noise perturbation, and geometric transformations, the model learns more robust feature

representations. This strategy helps compensate for the limited scale of the original dataset and has proven effective in small-sample visual tasks.

3.2. Key Techniques and Model Construction

3.2.1. Applicability of YOLOv8 in Rock Art Image Analysis

YOLOv8 is a high-performance one-stage object detection model proposed in recent years. It achieves a strong balance between detection accuracy and computational efficiency and has been widely validated in tasks such as industrial defect detection, agricultural recognition, and complex scene analysis. Its end-to-end feature learning capability, together with efficient multi-scale feature fusion, allows the model to extract discriminative semantic information directly from raw images without relying on handcrafted features. This makes YOLOv8 a suitable baseline model for Northern Chinese rock art recognition. Given the visual characteristics of rock art images—abstract structures, diverse forms, and complex backgrounds—YOLOv8 can extract multi-scale features ranging from low-level textures to high-level semantics through deep convolutional layers. This provides a solid foundation for the automatic recognition of different categories, including Anthropomorphic_Face, Deer, and Human_Horse. Without relying on prior rules, the model can learn carving structures, contour relationships, and local symbolic features directly from data, thereby improving the level of automation in the recognition process.

Structurally, YOLOv8 consists of three main components: the Backbone, the Neck, and the Head, as illustrated in Figure 1. The Backbone is built on CBS (Conv–BatchNorm–SiLU) modules and C2f modules, which enhance information flow through cross-layer connections and hierarchical feature extraction. This design also helps mitigate gradient degradation in deep networks. In addition, the SPPF (Spatial Pyramid Pooling Fast) module expands the receptive field, improving the model’s ability to capture targets at different scales. This is particularly important for rock art images, where object scales vary significantly.

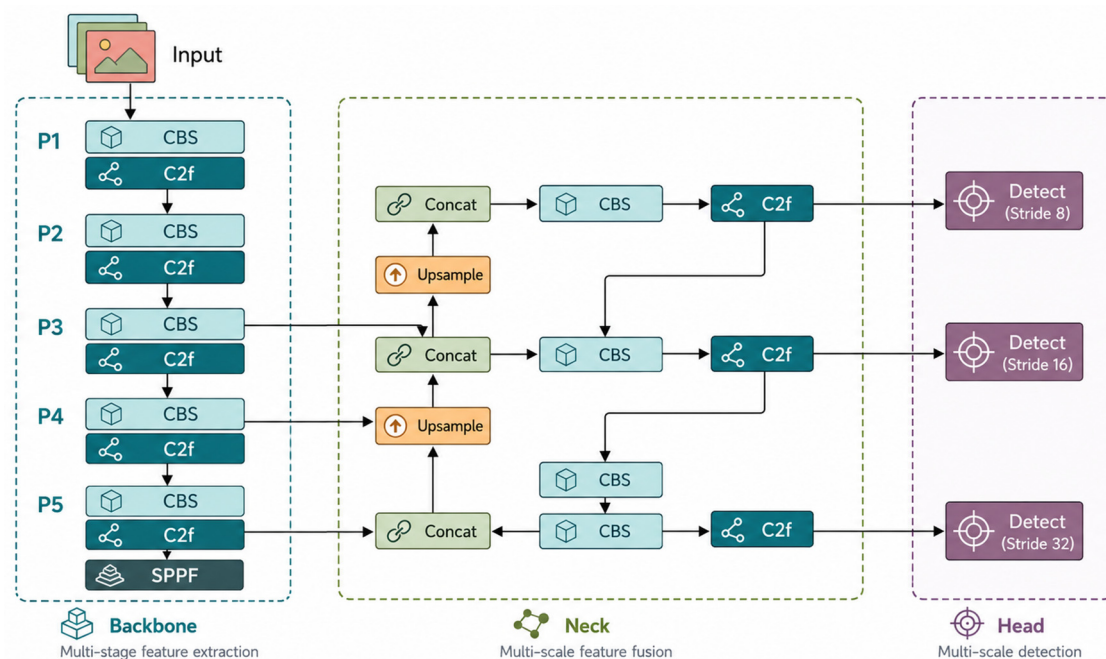


Figure 1. YOLOv8 network architecture model. Adapted from Ji et al. [13].

In the feature fusion stage, YOLOv8 adopts a feature pyramid-based multi-scale fusion strategy. Through upsampling and feature concatenation, features from different levels are effectively integrated. This enables the model to capture both fine-grained local

details and global semantic information. Such a mechanism is crucial for distinguishing rock art categories that share similar global structures but differ in subtle local features.

In the detection head, YOLOv8 employs a decoupled design that separates classification and bounding box regression tasks, reducing feature interference in multi-task learning. The model also adopts an anchor-free detection paradigm, combined with a Task-Aligned Assigner and Distribution Focal Loss, which further improves localization accuracy and training stability.

Despite its strong performance in many complex visual tasks, YOLOv8 still faces limitations when applied to Northern rock art images. On one hand, rock art images often contain blurred carvings, discontinuous boundaries, and strong background interference, making it difficult to extract stable features. On the other hand, different categories share similar overall structures, while discriminative information is concentrated in local regions. This places higher demands on feature representation. Under such conditions, feature extraction based on fixed convolution kernels struggles to adapt to highly variable visual patterns.

Based on this analysis, it is necessary to introduce a mechanism that can adaptively adjust feature responses within the YOLOv8 framework. To address this, the present study incorporates Omni-Dimensional Dynamic Convolution (ODConv) into the model. This enhancement improves the modeling of complex carving structures and fine-grained differences, thereby increasing overall detection performance and robustness in rock art image analysis.

3.2.2. Omni-Dimensional Dynamic Convolution (ODConv)

Omni-Dimensional Dynamic Convolution (ODConv) is a dynamic convolution method based on a multi-dimensional attention mechanism. Its core idea is to overcome the limitation of fixed convolution kernel parameters by introducing input-dependent weight modulation, enabling adaptive adjustment of convolution kernels across multiple dimensions. Unlike conventional dynamic convolution methods that focus only on the kernel-number dimension, ODConv jointly models the spatial dimension, input channel dimension, output channel dimension, and kernel dimension. This design allows the model to learn richer and more complementary feature representations. ODConv can be directly used as a replacement for standard convolution in various neural network architectures, significantly improving feature representation while maintaining computational efficiency [38].

Northern Chinese rock art images are characterized by subtle inter-class differences, significant variation in preservation conditions, and complex carving structures. Traditional static convolution, with fixed kernel weights, cannot effectively adapt to variations in pattern structure, carving morphology, and weathering conditions within a unified feature extraction framework. To enhance fine-grained feature modeling, this study incorporates ODConv into the YOLOv8 framework by replacing standard convolution layers in both the Backbone and Neck. This enables convolution kernels to dynamically adjust across spatial, channel, and kernel dimensions.

Compared with earlier dynamic convolution methods such as CondConv and DyConv, which mainly focus on kernel-level dynamic selection, ODConv employs a four-dimensional parallel attention mechanism. This allows convolutional layers to generate adaptive responses based on the input image content and contextual information, resulting in more discriminative feature representations in complex visual scenarios.

The core of ODConv lies in introducing four types of complementary attention to the convolution kernel. All four attention branches are jointly optimized during end-to-end training through backpropagation, enabling the model to adaptively learn content-

dependent feature responses without manual weight assignment. The mechanisms are described as follows:

1. Spatial Attention (α_{si})

Spatial attention assigns adaptive weights to the spatial positions of the convolution kernel, enhancing the model's ability to focus on critical local structures. Let a single-branch convolution kernel be denoted as $w_i^m \in \mathbb{R}^{k \times k \times c_{in}}$, and the spatial attention weights as $\alpha_{si} \in \mathbb{R}^{k \times k \times 1}$, which are dynamically learned from the input features, as shown in Figure 2a. Through a broadcasting mechanism, the attention weights are multiplied element-wise with the convolution kernel to obtain the spatially weighted kernel:

$$W_i^{(s)} = \alpha_{si} \odot w_i^m \quad (5)$$

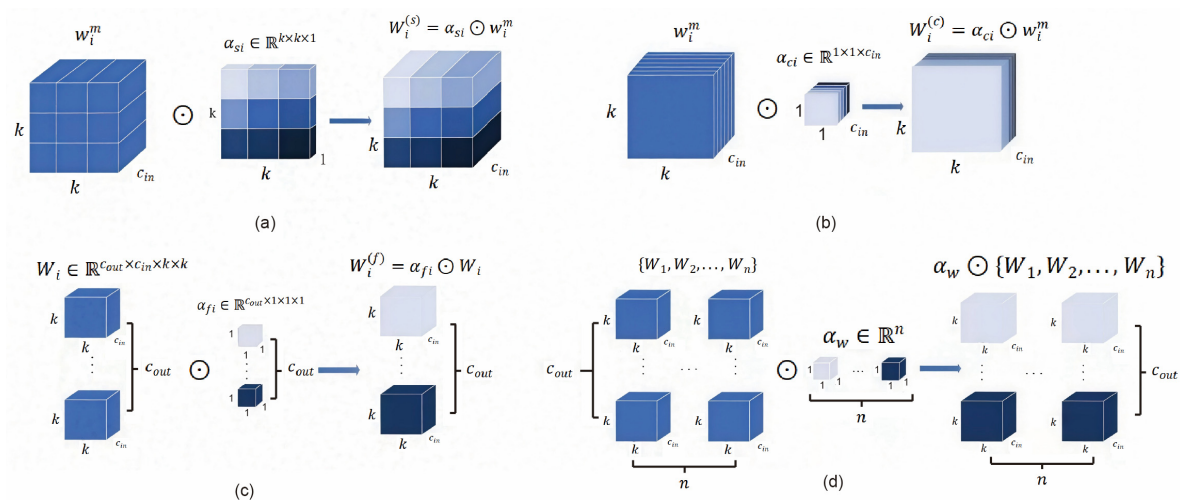


Figure 2. Schematic diagram of convolutional kernel attention modulation across different dimensions: (a) spatial-dimension attention modulation; (b) input-channel attention modulation; (c) output-channel attention modulation; and (d) kernel-number attention modulation. Adapted from Li, C et al. [38].

After the operation, the feature dimension remains $\mathbb{R}^{k \times k \times c_{in}}$. This enables the convolution to focus more on informative regions while suppressing responses from irrelevant spatial locations. As a result, the model can adaptively select appropriate convolutional kernel combinations based on the overall content of rock art images, thereby enhancing feature representation without increasing network depth.

2. Input Channel Attention (α_{ci})

Input channel attention assigns adaptive weights to the input channel dimension of the convolution kernel. This mechanism strengthens informative feature channels while suppressing redundant or noisy ones. Let a single-branch convolution kernel be denoted as $w_i^m \in \mathbb{R}^{k \times k \times c_{in}}$, and the input channel attention weights as $\alpha_{ci} \in \mathbb{R}^{1 \times 1 \times c_{in}}$, which are dynamically generated from the input features, as illustrated in Figure 2b. Through a broadcasting mechanism, the attention weights are multiplied element-wise with the convolution kernel to obtain the channel-weighted kernel:

$$W_i^{(c)} = \alpha_{ci} \odot w_i^m \quad (6)$$

After the operation, the feature dimension remains $\mathbb{R}^{k \times k \times c_{in}}$. This improves the model's ability to select and represent informative feature channels. For rock art images

with severe color degradation, this mechanism can adaptively enhance channels that retain weak color information while suppressing channels dominated by noise.

3. Output Channel Attention (Filter Attention, α_{fi})

Output channel attention assigns adaptive weights to the output channel (filter) dimension of the convolution kernel, thereby improving the model's ability to select high-level semantic features. Let the full convolution kernel be denoted as $W_i \in \mathbb{R}^{c_{out} \times c_{in} \times k \times k}$, and the output channel attention weights as $\alpha_{fi} \in \mathbb{R}^{c_{out} \times 1 \times 1 \times 1}$, which are dynamically learned from the input features, as shown in Figure 2c. Through a broadcasting mechanism, the attention weights are multiplied element-wise with the convolution kernel to obtain the filter-weighted kernel:

$$W_i^{(f)} = \alpha_{fi} \odot W_i \quad (7)$$

After the operation, the dimension remains $\mathbb{R}^{c_{out} \times c_{in} \times k \times k}$, which enhances the response of discriminative output channels. This mechanism helps distinguish between shared low-level textures across rock art categories (e.g., weathering patterns) and high-level semantic features that are more discriminative (e.g., deer antler contours or facial ornaments).

4. Kernel Attention (α_w)

Kernel attention performs global adaptive weighting over multiple parallel expert convolution kernels to obtain an optimal kernel combination for different inputs. Assume there are n expert kernels $\{W_1, W_2, \dots, W_n\}$. The kernel attention weights $\alpha_w \in \mathbb{R}^n$ are learned through global average pooling followed by fully connected layers and normalized using a Softmax function, as illustrated in Figure 2d. By computing the weighted sum of these expert kernels, the final dynamic convolution kernel is obtained:

$$K = \sum_{i=1}^n \alpha_{w,i} \cdot W_i \quad (8)$$

This mechanism expands the feature representation space without increasing network depth and improves the model's adaptability to complex inputs. It is particularly effective for capturing the local continuity of discontinuous carving lines in rock art images. It strengthens the response along carving edges while suppressing irrelevant features from smooth rock surface regions.

The four types of attention are broadcast to match the dimensions of the expert convolution kernels. They are then combined through element-wise multiplication and weighted summation to produce a content-adaptive dynamic convolution kernel:

$$K = \sum_{i=1}^n \alpha_{w,i} \cdot (\alpha_{fi} \odot \alpha_{ci} \odot \alpha_{si} \odot W_i) \quad (9)$$

Compared with conventional dynamic convolution methods that mainly enhance features in continuous textures or regular structures, ODConv introduces attention modulation across multiple dimensions. This enables more flexible feature responses under complex visual conditions. This property is particularly important for rock art images. On one hand, carving patterns often appear as discontinuous linear structures, which require strong local adaptability in the spatial domain. On the other hand, discriminative cues between categories rely on subtle structural differences, such as angle variations and branching patterns. This requires fine-grained feature selection at the channel and kernel levels. By adaptively modulating feature responses across multiple dimensions, ODConv is well suited for rock art images, which exhibit highly abstract and structurally unstable visual characteristics.

3.2.3. YOLOv8-ODConv Network Architecture and Integration Strategy

Based on the original YOLOv8 architecture and considering the visual characteristics of Northern rock art images—abstract structures, diverse forms, and complex backgrounds—this study develops an improved model termed YOLOv8-ODConv (see Figure 3). The overall architecture consists of three components: Backbone, Neck, and Head. ODConv modules are introduced at key stages of feature extraction and fusion to enhance the modeling of complex carving structures and fine-grained feature differences.

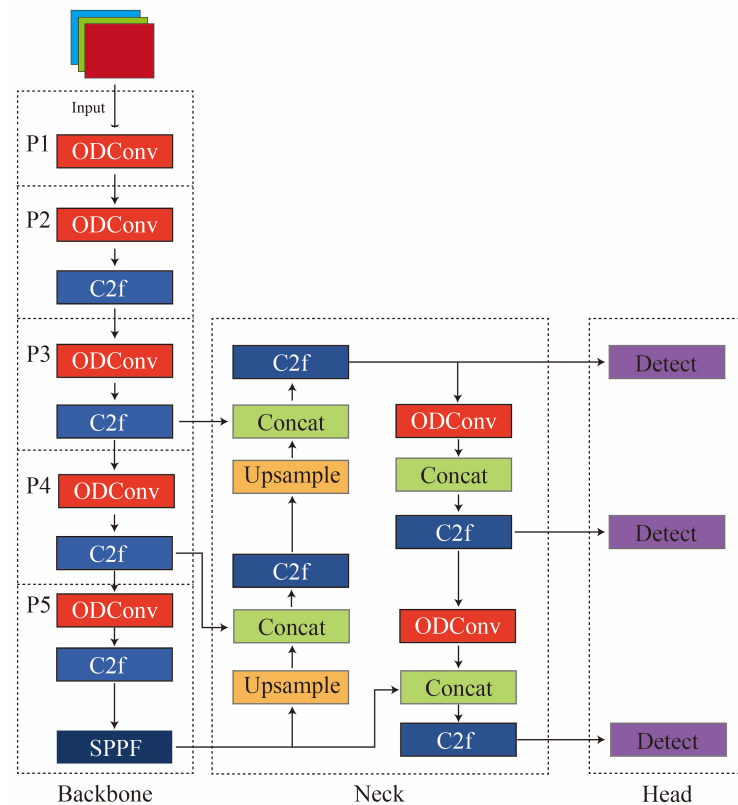


Figure 3. YOLOv8-ODConv network architecture model. Adapted from Bai, L et al. [23].

In the Backbone stage, the model follows the hierarchical feature extraction design of YOLOv8. It extracts multi-scale features, ranging from low-level textures to high-level semantics, through a feature pyramid structure. Unlike the original YOLOv8, which relies on standard convolution, ODConv modules are embedded into selected convolution layers from stages P1 to P5 (as highlighted in red in Figure 3), and are combined with C2f blocks. This design preserves the lightweight nature of the network while introducing input-dependent dynamic convolution. As a result, convolution kernel weights are adaptively adjusted according to feature content, improving the representation of diverse carving patterns.

In the Neck stage, the model adopts a combination of Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) to achieve multi-scale feature fusion and information propagation. First, high-level semantic features are fused with low-level detail features through upsampling and concatenation, enhancing the representation of small-scale targets. Then, further feature integration is performed along the downsampling path to strengthen information exchange across different scales. During this process, ODConv modules are inserted at key fusion nodes to dynamically model the fused features. This allows the network to adaptively adjust spatial and channel responses within multi-scale representations, thereby improving feature discrimination under complex backgrounds.

This strategy is particularly effective in reducing the interference caused by the overlap between carving structures and rock surface textures.

In the Head stage, the model adopts the decoupled detection head of YOLOv8, which performs classification and bounding box regression separately for feature maps at different scales. With the enhanced feature representations provided by ODConv, the model can effectively detect and localize rock art objects across multiple scales, including Anthropomorphic_Face, Deer, and Human_Horse.

The placement of ODConv modules was determined according to the functional characteristics of different network stages. In this study, ODConv was used to replace standard convolution (Conv) layers. In the Backbone, ODConv was primarily embedded into shallow and intermediate convolutional layers to enhance the extraction of local structural details and discontinuous carving patterns commonly observed in rock art images. In the Neck, ODConv was introduced into key feature fusion layers to improve the adaptive aggregation of multi-scale features under complex background conditions. This design preserves the original multi-scale detection architecture of YOLOv8 while balancing feature representation capability and computational efficiency.

In addition, this study did not conduct a comprehensive sensitivity analysis for all possible ODConv insertion positions. Future work will further investigate the influence of different insertion strategies on model performance and computational efficiency.

From the perspective of the overall integration strategy, this study does not perform a large-scale redesign of the YOLOv8 architecture. Instead, a “local replacement and hierarchical embedding” approach is adopted, in which ODConv is introduced as a substitute for standard convolution at key locations in the Backbone and Neck. This design preserves the inherent advantages of YOLOv8 in multi-scale detection while enhancing the adaptability of feature extraction through dynamic convolution. As a result, a favorable balance is achieved between model performance and computational efficiency. Previous studies have shown that incorporating ODConv into object detection frameworks can effectively improve recognition accuracy in complex scenarios, particularly in tasks involving multi-scale targets and fine-grained feature modeling.

Based on this design, the YOLOv8-ODConv model introduces dynamic convolution into both feature extraction and fusion stages. This enables the network to adaptively model the complex and variable carving structures and abstract representations in Northern rock art images, thereby providing more stable and discriminative feature representations for subsequent detection and classification tasks.

3.3. Experimental Environment and Training Settings

The proposed model is implemented based on the Ultralytics YOLOv8 framework. The experimental environment consists of Python 3.8, PyTorch 1.10, and CUDA 11.0. All experiments are conducted on an NVIDIA RTX 4060 Ti GPU with 12 GB of memory.

For training, the number of epochs is set to 200, the batch size is 4, and all input images are resized to 640×640 pixels. These settings ensure stable training while maintaining a balance between computational efficiency and memory usage. In addition, to avoid potential data leakage under small-sample conditions, strict data independence is maintained among the training, validation, and test sets. This ensures the objectivity and reliability of the evaluation results.

For all experiments, the models were trained using the same experimental settings to ensure fair comparison. Specifically, SSD, YOLOv5, YOLOv8, YOLOv8-Backbone-ODConv, and YOLOv8-ODConv adopted the same dataset split, training epochs, batch size, and input image resolution. The optimizer used in this study was SGD, with an initial learning

rate of 0.01. Pretrained weights provided by the official implementations were used for model initialization.

3.4. Evaluation Metrics

To comprehensively evaluate model performance, several standard metrics in object detection are adopted, including Precision, Recall, Average Precision (*AP*), *F*₁-score, and mean Average Precision (*mAP*).

In detection evaluation, a prediction is considered a True Positive (*TP*) when the Intersection over Union (IoU) between the predicted bounding box and the ground truth exceeds a predefined threshold. Otherwise, it is classified as a False Positive (*FP*) or False Negative (*FN*), depending on the prediction outcome. Based on these definitions, the evaluation metrics are computed as follows.

Precision (*P*) measures the proportion of correctly predicted targets among all predicted targets. It is defined as:

$$P = \frac{TP}{TP + FP} \quad (10)$$

Here, *TP* (True Positive) denotes the number of samples correctly predicted as the target class, i.e., cases where both the prediction and the ground truth are positive. *FP* (False Positive) denotes the number of samples incorrectly predicted as the target class, where the prediction is positive but the ground truth is negative. Precision measures the reliability of the model's predictions. A higher Precision indicates fewer false detections and more accurate prediction results.

Recall (*R*) represents the proportion of actual targets that are correctly identified by the model. It is defined as:

$$R = \frac{TP}{TP + FN} \quad (11)$$

Here, *TP* (True Positive) denotes the number of samples correctly predicted as the target class, and *FN* (False Negative) denotes the number of missed target samples, i.e., cases where the ground truth is positive but the model fails to detect them. Recall measures the model's ability to cover actual target instances. A higher Recall indicates a lower miss rate.

Average Precision (*AP*) evaluates detection performance across different recall levels and is defined as the area under the Precision–Recall (*PR*) curve. It reflects the overall trade-off between Precision and Recall under varying thresholds. A higher *AP* indicates that the model can maintain high prediction accuracy while achieving strong recall. The *F*₁-score provides a comprehensive evaluation of Precision and Recall by computing their harmonic mean. It is defined as:

$$F_1 = \frac{2PR}{P + R} \quad (12)$$

Here, *P* denotes Precision and *R* denotes Recall. The *F*₁-score balances these two metrics and reaches a high value only when both Precision and Recall are high. Therefore, it effectively avoids bias caused by focusing on a single metric and is widely used to evaluate the overall performance of classification and object detection models.

Mean Average Precision (*mAP*) is used to evaluate the overall performance of a model in multi-class object detection tasks. It is defined as the arithmetic mean of the Average Precision (*AP*) across all categories. Here, *AP* represents the detection performance of a single class and is computed as the area under the Precision–Recall (*PR*) curve. The *mAP* is calculated as follows:

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (13)$$

Here, *N* denotes the total number of target categories. In this study, two evaluation metrics are used: *mAP*@0.5 and *mAP*@0.5–0.95. The *mAP*@0.5 is computed at an

Intersection over Union (IoU) threshold of 0.5. The mAP@0.5–0.95 follows the COCO evaluation protocol, where IoU thresholds range from 0.5 to 0.95 with a step size of 0.05, and the corresponding AP values are averaged to provide a more comprehensive performance assessment.

4. Results

4.1. Training Process and Detection Performance of YOLOv8-ODConv

To evaluate the performance of the proposed YOLOv8-ODConv model in Northern rock art image recognition, the variations in loss functions during training and the detection metrics are analyzed. The results are shown in Figure 4.

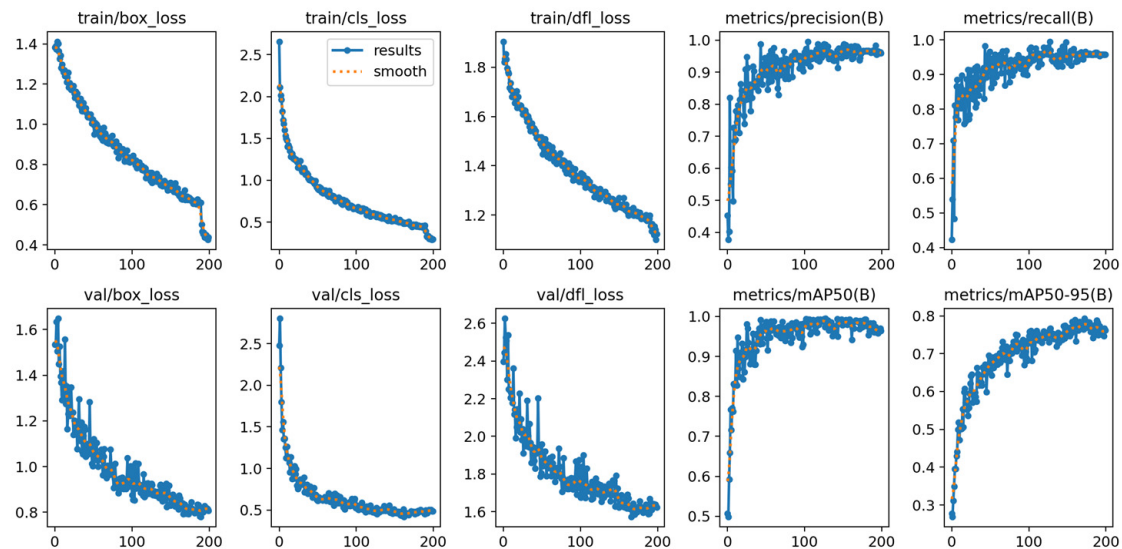


Figure 4. Performance Metrics of the YOLOv8-ODConv Model.

From the training process, all loss functions exhibit stable convergence. Specifically, train/box_loss, train/cls_loss, and train/dfl_loss decrease rapidly in the early stages of training, followed by a gradual stabilization. The validation losses (val/box_loss, val/cls_loss, and val/dfl_loss) show similar trends to those of the training set. Although slight fluctuations are observed in the early stage, the overall trend is smooth and decreasing. The consistency between training and validation curves, without noticeable divergence, indicates good convergence and generalization ability, with no evident overfitting.

In terms of detection performance, both Precision and Recall increase steadily during training and reach a stable stage after approximately 100–150 epochs. Precision approaches 1.0 in the later stages, indicating a low false detection rate. Recall remains at a relatively high level, demonstrating strong coverage of target instances. Meanwhile, mAP@0.5 stabilizes at approximately 0.98, and mAP@0.5–0.95 converges to around 0.78. It can be observed that mAP@0.5 converges faster, whereas mAP@0.5–0.95 improves more gradually, suggesting that there is still room for improvement under stricter localization requirements.

Overall, all metrics show minimal fluctuations during convergence, and the training process remains stable. These results indicate that the proposed YOLOv8-ODConv model can achieve effective convergence in complex rock art image scenarios while maintaining strong detection performance and robustness. Even under limited sample conditions, the model demonstrates stable convergence and high performance, highlighting its suitability for small-sample, complex visual tasks.

To further analyze the model's performance across different categories, a normalized confusion matrix is constructed based on the test set, as shown in Figure 5. From the

diagonal entries, all categories achieve relatively high classification accuracy. Among them, the Human_Horse category performs the best, with an accuracy of approximately 0.97. The Anthropomorphic_Face category achieves around 0.94, while the Deer category shows a relatively lower accuracy of about 0.86. Overall, the high values along the diagonal indicate that the model predictions are consistent with ground truth labels for most samples, demonstrating strong classification capability.

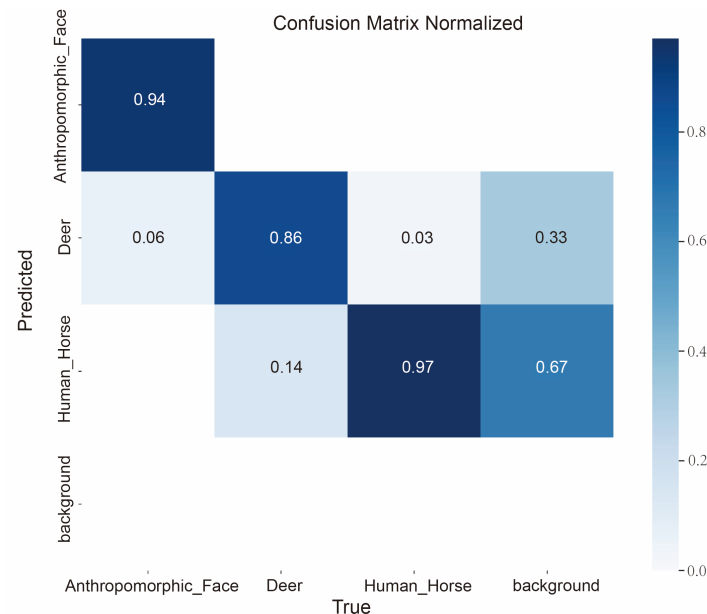


Figure 5. Normalized Confusion Matrix of the Northern China Rock Art Dataset.

From the off-diagonal distribution, a certain degree of class confusion is still observed. Specifically, about 0.06 of Anthropomorphic_Face samples are misclassified as Deer. Approximately 0.14 of Deer samples are misclassified as Human_Horse, and a small portion of Deer samples are predicted as background. In addition, some background regions are incorrectly detected as target categories. Among these, the proportions predicted as Deer and Human_Horse are approximately 0.33 and 0.67, respectively. This indicates that false detections still occur in regions with complex textures.

Overall, the model shows strong discriminative capability for the Human_Horse category, while the Deer category remains more challenging due to its structural similarity to other classes. This confusion can be attributed to partial morphological similarities between Deer and Human_Horse, such as linear branching structures and antler-like contours, as well as the weaker continuity of carving patterns in Deer images. These factors increase the difficulty of feature discrimination during the extraction stage. Although ODConv improves the modeling of local structural differences, some feature overlap still occurs when dealing with highly similar abstract symbolic structures.

To analyze the overall performance of the model under different confidence thresholds, the F1–Confidence curve is plotted, as shown in Figure 6. From the overall trend, the F1 score for all categories follows a pattern of first increasing and then decreasing as the confidence threshold rises, reaching its peak within a moderate confidence range. This indicates that at low thresholds, although recall is high, the number of false detections is also large; at high thresholds, precision improves, but recall drops significantly, leading to a decrease in the F1 score.

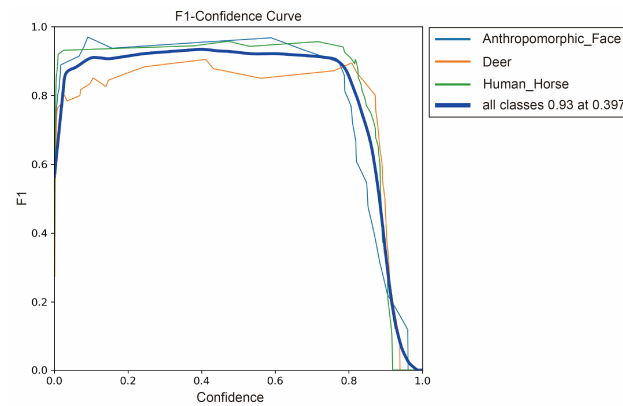


Figure 6. F1–Confidence Curve.

In terms of category-wise performance, the Human_Horse category maintains a high F1 score across most confidence intervals, demonstrating strong stability. The Deer category shows a relatively smooth curve and maintains balanced performance within the medium to high confidence range. In contrast, the Anthropomorphic_Face category exhibits a more pronounced decline at higher confidence thresholds, suggesting a certain degree of missed detections under stricter filtering conditions. Overall, the model achieves optimal performance at a confidence threshold of approximately 0.40, where the F1 score reaches around 0.93. As the threshold increases further, the F1 score decreases rapidly, indicating that the model's recall becomes limited under high-confidence constraints.

In summary, the F1–Confidence curve demonstrates that the proposed model achieves a good balance between precision and recall within the moderate confidence range, while also revealing differences in feature representation and classification difficulty across categories.

The Precision–Confidence curve illustrates how detection precision varies with different confidence thresholds, as shown in Figure 7. As the confidence threshold increases, the precision for all categories shows an overall upward trend and gradually stabilizes at higher confidence levels. In terms of category-wise performance, the Human_Horse category maintains a high precision across the entire confidence range. The Anthropomorphic_Face category exhibits a relatively stable curve and gradually approaches 1.0 in the medium to high confidence range. The Deer category shows lower precision at low confidence thresholds, but its performance improves significantly as the threshold increases, eventually approaching 1.0 at higher confidence levels. Overall, the model achieves a precision of 1.00 at a confidence threshold of approximately 0.92. Within a broad confidence interval (approximately 0.1–0.8), the precision remains above 0.9, indicating consistently high prediction reliability.

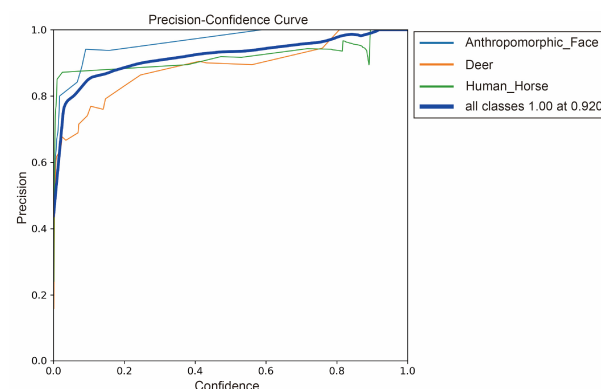


Figure 7. Precision–Confidence Curve.

The Recall–Confidence curve shows how recall varies with different confidence thresholds, as illustrated in Figure 8. At low confidence levels (close to 0), the recall for all categories approaches 1.00, indicating that the model is able to detect nearly all target instances. As the confidence threshold increases, recall gradually decreases and exhibits a noticeable decline in the high-confidence range. From the category-wise perspective, the Human_Horse category maintains a high recall level in the low to medium confidence range. The Deer category also preserves relatively high recall in the medium confidence interval, with a more gradual decline overall. In contrast, the Anthropomorphic_Face category shows a more pronounced drop in the high-confidence region, with a rapid decrease occurring after the confidence threshold exceeds approximately 0.75. Overall, the model maintains a high recall level within the moderate confidence range (approximately 0.5–0.7). However, when the confidence threshold exceeds 0.8, recall decreases sharply, showing a clear attenuation trend.

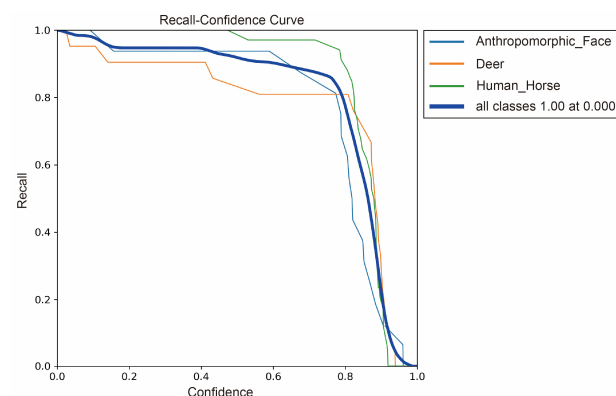


Figure 8. Recall–Confidence Curve.

The Precision–Recall (PR) curve is used to evaluate the model’s precision performance under different recall levels, as shown in Figure 9. The PR curves for all categories are mainly distributed in the upper-left region of the coordinate space, indicating strong detection performance. From the category-wise results, the Average Precision (AP) for Anthropomorphic_Face is approximately 0.991, while the AP for Human_Horse is about 0.966, and that for Deer is around 0.961. Notably, all categories maintain relatively high precision even at higher recall levels, demonstrating the model’s robustness. Overall, the model achieves an mAP@0.5 value of approximately 0.973 on the validation set.

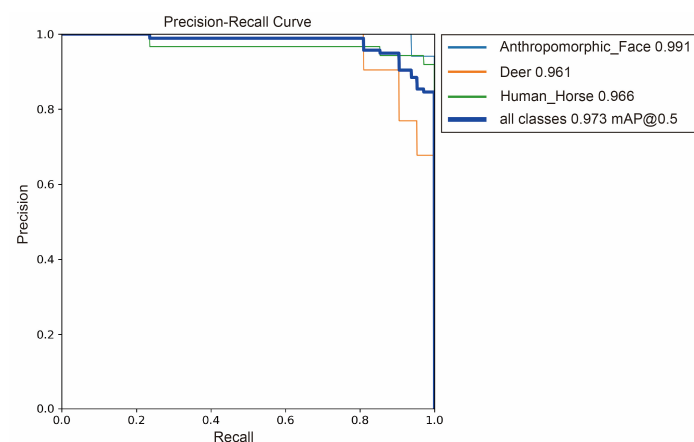


Figure 9. Precision–Recall (PR) curves.

It should be noted that the mAP@0.5 derived from the PR curve is calculated through precise curve integration, which may show slight differences compared to the mAP values

reported in Table 3, obtained during the validation stage. However, the overall trends remain consistent.

Table 3. Experimental Comparison Results of Different Models on the Validation Set.

Modules	F1	Recall	mAP@0.5	mAP@0.5–0.95	FPS	FLOPs/G
SSD	82.5%	79.8%	85.6%	70.8%	102.5	23.7
YOLOv5	92.7%	86.7%	95.8%	82.1%	94.3	10.5
YOLOv8	93.0%	88.5%	97.3%	83.9%	91.4	8.6
YOLOv8–Backbone–ODConv	93.8%	88.4%	97.6%	83.7%	90.2	8.1
YOLOv8–Neck–ODConv	94.1%	88.6%	98.1%	83.8%	90.0	7.9
YOLOv8-ODConv	95.0%	89.3%	98.9%	84.6%	89.7	7.2

4.2. Comparison with Different Models and Ablation Analysis

To evaluate the performance of the proposed YOLOv8-ODConv model, SSD, YOLOv5, YOLOv8, and YOLOv8–Backbone–ODConv are selected as baseline models for comparison. All models were evaluated using the same dataset and under identical experimental conditions. The results are presented in Table 3. Among them, YOLOv8–Backbone–ODConv is used to analyze the effect of introducing ODConv only in the Backbone, while YOLOv8-ODConv represents the complete model with ODConv embedded in both the Backbone and Neck.

From the overall results, SSD shows relatively lower performance across all metrics, with F1, Recall, mAP@0.5, and mAP@0.5–0.95 values of 82.5%, 79.8%, 85.6%, and 70.8%, respectively. YOLOv5 demonstrates a significant improvement over SSD, achieving an F1 score of 92.7% and mAP@0.5 of 95.8%. YOLOv8 further improves detection performance, with an F1 score of 93.0%, Recall of 88.5%, mAP@0.5 of 97.3%, and mAP@0.5–0.95 of 83.9%.

When ODConv is introduced only in the Backbone, YOLOv8–Backbone–ODConv achieves an F1 score of 93.8%, mAP@0.5 of 97.6%, and mAP@0.5–0.95 of 83.7%. However, its Recall (88.4%) is slightly lower than that of YOLOv8 (88.5%). When ODConv is further integrated into both the Backbone and Neck, the YOLOv8-ODConv model achieves the best overall performance, with F1, Recall, mAP@0.5, and mAP@0.5–0.95 reaching 95.0%, 89.3%, 98.9%, and 84.6%, respectively. In terms of detection speed, SSD achieves the highest FPS at 102.5. The FPS values for YOLOv5, YOLOv8, YOLOv8–Backbone–ODConv, and YOLOv8-ODConv are 94.3, 91.4, 90.2, and 89.7, respectively. As model complexity increases, FPS shows a slight decrease. However, YOLOv8-ODConv still maintains a relatively high inference speed. In terms of computational complexity, the proposed YOLOv8-ODConv achieves 7.2 G FLOPs, which is lower than the original YOLOv8 model (8.6 G). Although the FPS decreases slightly from 91.4 to 89.7, the proposed model achieves notable improvements in F1-score and mAP metrics while maintaining relatively low computational cost. These results indicate that YOLOv8-ODConv achieves a favorable balance between detection accuracy and computational efficiency.

Overall, YOLOv8-ODConv achieves the best performance across all key metrics, including F1, Recall, mAP@0.5, and mAP@0.5–0.95. The performance difference between YOLOv8–Backbone–ODConv and YOLOv8-ODConv also indicates that the embedding position of ODConv has a significant impact on model performance, with the joint integration in both Backbone and Neck yielding the most effective results.

4.3. Visualization of Detection Results

To provide a more intuitive comparison between YOLOv8 and YOLOv8-ODConv in rock art image recognition, representative samples are selected for visualization anal-

ysis. The results are presented in Table 4. The visualization covers categories including Human_Horse, Anthropomorphic_Face, and Deer, highlighting the differences in object localization and classification performance between the two models. The images were annotated using the Labelling tool by research personnel with a background in arts and design. During annotation, care was taken to ensure object completeness. Due to the dense distribution of rock art, adjacent carvings might partially fall into bounding boxes, which is normal given natural carving interruptions. No additional transformations were applied, preserving the original appearance of the images.

Table 4. Performance Comparison between YOLOv8 and YOLOv8-ODConv.

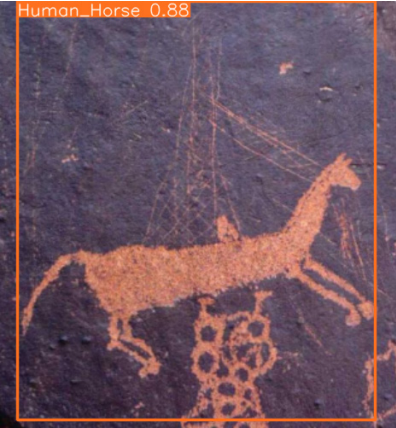
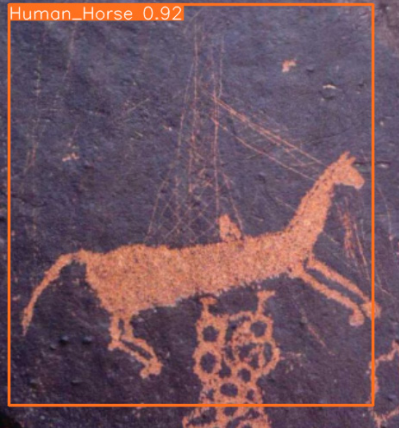
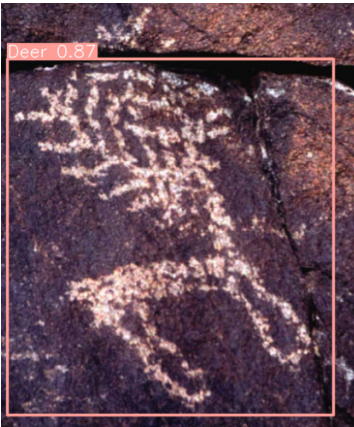
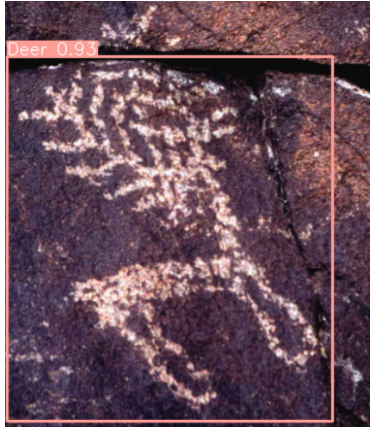
YOLOv8	YOLOv8-ODConv
	
	
	

Table 4. Cont.

YOLOv8	YOLOv8-ODConv
	
	
	

As shown in Table 4, under the same sample conditions, YOLOv8-ODConv achieves higher prediction confidence across multiple categories compared to the original YOLOv8. For instance, in Human_Horse samples, YOLOv8 produces confidence scores of 0.88 and 0.90, whereas YOLOv8-ODConv improves these to 0.92. For Anthropomorphic_Face, the confidence increases from 0.86 to 0.87 to 0.90–0.91. Similarly, for Deer samples, the confidence rises from 0.87 to 0.93.

In terms of bounding box localization, both models are able to detect the primary target regions. However, YOLOv8-ODConv demonstrates more stable alignment between predicted boxes and object contours in certain cases. These visualization results are consistent with the quantitative findings in Table 3 and further confirm the effectiveness of YOLOv8-ODConv in detecting Northern rock art images.

Some failure cases were still observed in regions with blurred carving structures and complex background textures. In particular, partial confusion between Deer and Human_Horse categories occurred when local contour patterns and discontinuous lin-

ear structures were highly similar. These results indicate that highly abstract symbolic structures and weak local features remain challenging for the current model.

5. Discussion

5.1. Mechanism of Performance Improvement in YOLOv8-ODConv

The experimental results demonstrate that the proposed YOLOv8-ODConv model achieves consistent improvements over baseline models such as YOLOv5 and YOLOv8 across multiple metrics, including F1, mAP@0.5, and mAP@0.5–0.95. As shown in the comparative experiments in Section 4, model performance improves progressively from YOLOv8 to YOLOv8–Backbone–ODConv and further to YOLOv8-ODConv. This trend indicates that ODConv contributes effectively at different levels of the network. The performance gain mainly stems from the multi-dimensional dynamic modeling capability introduced by ODConv during feature representation. Unlike conventional convolution with fixed kernels, ODConv incorporates dynamic weight allocation across spatial, channel, and kernel dimensions. This allows the convolution operation to adapt to input features, thereby enhancing the response to discriminative information.

Further analysis reveals that this mechanism is particularly effective in complex rock art scenarios. For example, confusion matrix results show high accuracy for Human_Horse and Deer, while F1–Confidence and Recall–Confidence curves indicate stable performance for these categories across a wide confidence range. In contrast, Anthropomorphic_Face shows performance degradation at higher confidence thresholds, suggesting challenges in cases with weak local features or minimal structural differences. From a feature perspective, Northern rock art images often exhibit blurred carvings, discontinuous structures, and complex background textures, which lead to feature overlap and redundancy. ODConv mitigates these issues by adaptively modulating convolution responses, enhancing local structure extraction and suppressing background interference. This improves detection stability under complex conditions. Additionally, curve analysis shows that the model achieves optimal performance at moderate confidence levels (approximately 0.6–0.7), while recall declines at higher thresholds. This reflects an inherent trade-off between precision and recall.

In summary, the performance improvement of YOLOv8-ODConv primarily arises from the adaptive modeling capability introduced by dynamic convolution, which is well suited for complex and structurally diverse rock art images.

5.2. Performance Differences Across Categories and Visual Interpretation

Although the overall detection performance is strong, differences remain across categories. Based on confusion matrix and Recall–Confidence results, Human_Horse and Deer maintain relatively high performance, whereas Anthropomorphic_Face shows lower performance, particularly in terms of recall at high confidence levels. These differences are closely related to visual characteristics. Human_Horse and Deer typically exhibit clearer global structures, such as body proportions, limb configurations, and posture patterns. These features are relatively consistent across samples, making them easier for the model to learn. As a result, their detection performance remains stable across confidence levels. In contrast, Anthropomorphic_Face exhibits higher abstraction and greater variability. Discriminative features are often localized, such as facial contours or symbolic carvings. These features vary significantly across samples and are more susceptible to weathering, blurred carvings, and background interference. This leads to unstable feature representation, reflected in reduced recall at high confidence thresholds and increased misclassification in the confusion matrix.

From a methodological perspective, although ODConv enhances feature representation through dynamic convolution, its effectiveness is still constrained by the intrinsic

separability of features. Highly abstract categories that rely on subtle local differences remain challenging. This highlights the limitations of convolution-based feature modeling when dealing with weak structural constraints and high semantic ambiguity.

Overall, these findings reflect a common issue in cultural heritage image recognition: semantic ambiguity, where categories lack stable and consistent structural discriminators. This increases learning difficulty and places higher demands on feature representation methods.

5.3. Role of Dynamic Convolution in Complex Visual Environments

Northern rock art images are characterized by dense carvings, overlapping textures, and significant background interference. Under such conditions, conventional convolution with fixed kernels lacks adaptability to input variations and may introduce redundant responses, thereby weakening discriminative feature extraction. To address this limitation, ODConv introduces dynamic weight allocation across spatial, channel, and kernel dimensions. This enables the convolution operation to adaptively adjust feature responses according to input content. As a result, target-related structural information is enhanced, while irrelevant background patterns are effectively suppressed during multi-scale feature extraction and fusion.

The effectiveness of this mechanism is supported by experimental results. After integrating ODConv, the model exhibits improved stability across multiple evaluation metrics. Specifically, the Precision–Confidence curve maintains high precision in the medium-to-high confidence range, while the Recall–Confidence curve shows a more gradual decline in recall. In addition, the confusion matrix reveals a reduction in off-diagonal misclassifications. These observations indicate enhanced feature discrimination under complex background conditions.

From a feature modeling perspective, ODConv enables adaptive selection of discriminative representations under varying visual patterns. This allows the model to maintain stable feature extraction even in challenging cases, such as blurred carvings, discontinuous line structures, and texture interference. This capability is particularly important for rock art images, where category discrimination relies heavily on subtle local structural cues.

Similar advantages of dynamic convolution have been reported in other complex vision tasks. For example, in robotic grasp detection, ODConv improves the representation of key regions and enhances the separation between targets and background when combined with feature fusion strategies. These findings demonstrate that dynamic convolution can improve detection accuracy while preserving computational efficiency. Such behavior is consistent with the visual characteristics of rock art images, where carving structures are tightly coupled with background textures, providing cross-domain support for the proposed approach [39].

Overall, by improving feature selection and suppressing irrelevant responses, dynamic convolution enhances the adaptability of the model to diverse visual patterns, leading to improved robustness in complex cultural image recognition tasks.

5.4. Limitations and Future Work

Despite its strong performance, the YOLOv8-ODConv model has several limitations that warrant further investigation.

First, the dataset size is relatively small, and some categories have limited samples, which may influence the generalization of the model. Although the original dataset is relatively small, the combination of data augmentation and independent subset splitting provides sufficient diversity for model training and evaluation. Cross-validation could further improve reliability, but due to computational resource constraints, we adopted

fixed splits and evaluated model performance comprehensively using F1-score, Recall, mAP, and FPS metrics. Future work will focus on expanding the dataset and improving cross-regional coverage.

Second, although YOLOv8 supports multi-object detection, the current dataset is based on single-class annotations. This does not fully capture the “scene-based” nature of rock art, where multiple elements (e.g., Human_Horse and Deer) often coexist within a single image. These compositions involve complex spatial relationships and require stronger multi-object detection and semantic understanding. Future studies should construct multi-object annotated datasets for scene-level analysis.

Third, while ODConv improves feature representation, it introduces additional computational overhead, leading to a slight reduction in inference speed. For real-time applications, such as mobile devices or field deployment, lightweight optimization strategies should be explored to balance accuracy and efficiency.

Finally, this study focuses on visual features and does not incorporate higher-level cultural semantics, such as symbolic meanings, behavioral relationships, or historical context. Rock art contains rich cultural information that cannot be fully captured by visual cues alone. Future research could integrate multimodal data (e.g., textual annotations, archaeological records) and leverage global modeling approaches such as Transformers to enhance semantic understanding.

It should be noted that the current experimental results are based on a single random dataset split and single-run training process. Therefore, the reported performance should not be interpreted as definitive evidence of statistical stability or robustness. Future work will further investigate model stability through repeated experiments with multiple random seeds and cross-validation strategies.

6. Conclusions

This study addresses the challenges of Northern Chinese rock art images, including abstract structures, diverse forms, and complex backgrounds, by proposing an improved object detection model based on YOLOv8, termed YOLOv8-ODConv. By introducing Omni-Dimensional Dynamic Convolution (ODConv) into key layers of the Backbone and Neck, the model enables adaptive adjustment of convolution kernels across spatial, channel, and kernel dimensions. This enhances the representation of complex carving structures and fine-grained feature differences. Unlike previous applications of dynamic convolution in regular texture domains, this work demonstrates that ODConv is effective for modeling highly abstract symbolic visual structures, providing a new perspective for applying deep learning to cultural heritage image analysis.

Based on the constructed rock art image dataset, extensive experiments are conducted to evaluate the proposed model. The results show that YOLOv8-ODConv outperforms SSD, YOLOv5, and YOLOv8 across key metrics such as F1-score, Recall, and mAP. Specifically, the model achieves an mAP@0.5 of 98.9% and an F1-score of 95.0%, while maintaining strong real-time performance. Further analysis indicates that the model performs more consistently on categories with relatively stable structural features (e.g., Human_Horse and Deer), whereas performance decreases for the more abstract Anthropomorphic_Face category. This suggests that feature separability plays a critical role in detection performance for complex cultural images. Considering the inherent difficulty in acquiring large-scale cultural heritage datasets, the proposed method demonstrates strong robustness under limited data conditions, highlighting its practical value for real-world heritage preservation applications.

Overall, the YOLOv8-ODConv model achieves high accuracy and stability in the automatic recognition of Northern rock art images. It provides an effective solution for

intelligent analysis of complex cultural images and offers technical support for the digital documentation and study of rock art.

Future work will focus on three directions. First, expanding the dataset and incorporating cross-regional rock art data to improve generalization. Second, integrating multimodal information, such as textual annotations and archaeological context, to enhance semantic understanding. Third, exploring lightweight model designs to support deployment in mobile and field environments.

Author Contributions: Conceptualization, L.G.; methodology, L.G. and F.J.; software, L.G.; validation, L.G. and F.J.; formal analysis, L.G. and F.J.; investigation, L.G.; resources, L.G.; data curation, L.G.; writing—original draft preparation, L.G.; writing—review and editing, L.G. and F.J.; visualization, L.G.; supervision, L.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset related to this study is available at: <https://www.kaggle.com/datasets/mawenzhe/northern-chinese-rock-art-image-dataset> (accessed on 28 April 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bednarik, R.G. The Main Problems in Rock Art Research. *Man India* **2008**, *88*, 199–213.
2. Ilmi, M.M.; Maryanti, E.; Nurdini, N.; Setiawan, P.; Kadja, G.T.M.; Ismunandar. A review of radiometric dating and pigment characterizations of rock art in Indonesia. *Archaeol. Anthropol. Sci.* **2021**, *13*, 120. [CrossRef]
3. Jalandoni, A.; Zhang, Y.S.; Zaidi, N.A. On the use of Machine Learning methods in rock art research with application to automatic painted rock art identification. *J. Archaeol. Sci.* **2022**, *144*, 105629. [CrossRef]
4. Zhang, W.J. Types and Distribution of Yinshan Rock Art. *Res. Front. Archaeol.* **2012**, 265–273. Available online: <https://d.wanfangdata.com.cn/thesis/J0066231> (accessed on 25 May 2026).
5. Wei, J.Z.Y. Types and Chronology of Rock Art in the Great Bend Region of the Yellow River from an Archaeological Perspective. *Qinghai Natl. Stud.* **2024**, 127–133. [CrossRef]
6. Liu, Y. Image Modeling and Cultural Implications of Regional Rock Art in Inner Mongolia. Ph.D. Thesis, Fudan University, Shanghai, China, 2012. Available online: https://kns.cnki.net/kcms2/article/abstract?v=A2Z-m-A1gckbeDxy4NRd7sjyho4plkb_uqlRx_C4lPFnWUieEzi1c1uOCNRYkxi3efrzaUnOBjldQ195cqPNWeLHzHZvmudlawZ_IYh-MbsDn10Z4IYjpb5e2JLSpO4rSZ-zC3HCQRcXSPpTK8-yxLhWDCLZAWUTHo05bFxuxkh2NjwnTV4qqKfRu8K5ooy&uniplatform=NZKPT&language=CHS (accessed on 1 May 2026).
7. Ning, Y. A Study of Stylized Deer Rock Art in Northwest China. *Res. Front. Archaeol.* **2024**, 338–352. Available online: https://kns.cnki.net/kcms2/article/abstract?v=CP5zkKmU_7eAnGDn6J5TmZPlpkP04Sbd94PbZ7qfzS9DPOJ8oq5_uFvUnbhp-UcGqfcw2bV5R9QsdWbo-kDK27yRLUVPndosWkxGFvFAMxN1IpI-YbFwrHemTznt0SrPMd6LPLF3D1CBXpep7mIB-0YU2_0rMZEe934GgNg7RTfoZsHsZT2zdKpy2VS7l4XG&uniplatform=NZKPT&captchaId=d5ef201c-10e5-4704-812a-340e80281f1e (accessed on 25 May 2026).
8. Zhang, L.M. The Inspiration of “Human Face Images” in Northern Rock Art for Contemporary Visual Design. Master’s Thesis, Lanzhou University of Finance and Economics, Lanzhou, China, 2024. Available online: <https://link.cnki.net/doi/10.27732/d.cnki.gnzsx.2024.000145> (accessed on 25 May 2026).
9. Wang, X.; Cheng, Y.; Zhang, R.Y.; Fan, Y.; Huang, J.Z.; Zhang, Y.; Yan, H.B. Classification of Weathering Environments for Stone Cultural Heritage Based on Hyperspectral Imaging Technology. *Laser Optoelectron. Prog.* **2025**, *62*, 1037005. [CrossRef]
10. Guo, T.; Xu, X. Salient object detection from low contrast images based on local contrast enhancing and non-local feature learning. *Vis. Comput.* **2021**, *37*, 2069–2081. [CrossRef]
11. Sultan, W.; Anjum, N.; Stansfield, M.; Ramzan, N. Hybrid Local and Global Deep-Learning Architecture for Salient-Object Detection. *Appl. Sci.* **2020**, *10*, 8754. [CrossRef]
12. Wei, X.-S.; Song, Y.-Z.; Mac Aodha, O.; Wu, J.; Peng, Y.; Tang, J.; Yang, J.; Belongie, S. Fine-grained image analysis with deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 8927–8948. [CrossRef] [PubMed]
13. Ji, N.; Ju, F.; Wang, Q. An Application Study on Digital Image Classification and Recognition of Yunnan Jiaoma Based on a YOLO-GAM Deep Learning Framework. *Appl. Sci.* **2026**, *16*, 1551. [CrossRef]

14. Gao, C.; Zhao, G.F.; Gao, S.; Du, S.X.; Kim, E.; Shen, T. Advancing architectural heritage: Precision decoding of East Asian timber structures from Tang dynasty to traditional Japan. *Herit. Sci.* **2024**, *12*, 219. [\[CrossRef\]](#)
15. Zhang, S.Z.; Liu, Z.H.; Wang, K.P.; Huang, W.W.; Li, P. OBC-YOLOv8: An improved road damage detection model based on YOLOv8. *PeerJ Comput. Sci.* **2025**, *11*, e2593. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Xu, S.X.; Huang, W.J.; Wang, D.C.; Zhang, B.Y.; Sun, H.; Yan, J.Y.; Ding, J.L.; Wang, J.J.; Yang, Q.L.; Huang, T.C.; et al. Automatic pine wilt disease detection based on improved YOLOv8 UAV multispectral imagery. *Ecol. Inform.* **2024**, *84*, 102846. [\[CrossRef\]](#)
17. Huang, Z.B.; Li, X.H.; Fan, S.T.; Liu, Y.; Zou, H.; He, X.C.; Xu, S.; Zhao, J.H.; Li, W.F. ORD-YOLO: A Ripeness Recognition Method for Citrus Fruits in Complex Environments. *Agriculture* **2025**, *15*, 1711. [\[CrossRef\]](#)
18. Zhang, Y.J.; Gao, G.F.; Chen, Y.D.; Yang, Z.J. ODD-YOLOv8: An algorithm for small object detection in UAV imagery. *J. Supercomput.* **2025**, *81*, 202, Correction in *J. Supercomput.* **2025**, *81*, 384. <https://doi.org/10.1007/s11227-024-06829-9>. [\[CrossRef\]](#)
19. Li, Y.; Zhao, M.; Mao, J.; Chen, Y.; Zheng, L.; Yan, L. Detection and recognition of Chinese porcelain inlay images of traditional Lingnan architectural decoration based on YOLOv4 technology. *Herit. Sci.* **2024**, *12*, 137. [\[CrossRef\]](#)
20. Prasomphan, S. Toward Fine-grained Image Retrieval with Adaptive Deep Learning for Cultural Heritage Image. *Comput. Syst. Sci. Eng.* **2023**, *44*, 1295–1307. [\[CrossRef\]](#)
21. Wu, H.Y.; Kong, L.Y.; Liu, D.H. Crack Detection on Road Surfaces Based on Improved YOLOv8. *IEEE Access* **2024**, *12*, 190850–190864. [\[CrossRef\]](#)
22. Xi, X.; Wang, J.; Li, F.; Li, D. IRSDet: Infrared small-object detection network based on sparse-skip connection and guide maps. *Electronics* **2022**, *11*, 2154. [\[CrossRef\]](#)
23. Bai, L.; Xia, C.; Liu, F.; Yang, X.; Zhang, T. Full-dimensional dynamic convolution and progressive learning strategy for strawberry recognition based on YOLOv8. *Front. Plant Sci.* **2025**, *16*, 1541365. [\[CrossRef\]](#)
24. Mao, Q.; Wang, Y.; Xue, X.; Zhou, T.; Su, Y. Intelligent Identification Method for Abnormal Pipelines in Coal Mine Roadways Based on Improved YOLOv8n. *Ind. Mine Autom.* **2026**, *52*, 63–72.
25. Chen, Z.X.; Feng, J.J.; Zhu, X.Y.; Wang, B. YOLOv8-OCHD: A Lightweight Wood Surface Defect Detection Method Based on Improved YOLOv8. *IEEE Access* **2025**, *13*, 84435–84450. [\[CrossRef\]](#)
26. Ning, T.; Fu, M.; Wang, Y.Z.; Duan, X.D.; Abedin, M.Z. Application of deep learning for automatic detection of table tennis balls from an intelligent serving machine. *Appl. Soft Comput.* **2024**, *167*, 112329. [\[CrossRef\]](#)
27. Jalandoni, A.; Taçon, P.S.C. A New Recording and Interpretation of the Rock Art of Angono, Rizal, Philippines. *Rock Art Res.* **2018**, *35*, 47–61.
28. Li, L.; Huang, W.; Gu, I.Y.-H.; Tian, Q. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Trans. Image Process.* **2004**, *13*, 1459–1472. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Chen, J.; Yin, G.; Sun, K.; Dong, Y. Multi-view robust discriminative feature learning for remote sensing image with noisy labels. *Mob. Netw. Appl.* **2022**, *27*, 2487–2505. [\[CrossRef\]](#)
30. Jin, W.N.; Zhang, Y.S.; Xing, J. *Complete Works of Chinese Art Rock Art and Engraving Prints*; Huangshan Publishing House: Hefei, China, 2010.
31. Editorial Committee of Complete Collection of Chinese Rock Art. *Complete Collection of Chinese Rock Art Northern Rock Art*; Liaoning Fine Arts Publishing House: Shenyang, China, 2006.
32. Choi, Y.; Seo, S.; Jang, H.; Yoon, D. Improving the Performance of Deep-Learning-Based Ground-Penetrating Radar Cavity Detection Model using Data Augmentation and Ensemble Techniques. *Geophys. Geophys. Explor.* **2023**, *26*, 211–228.
33. Lerch, L.; Huber, L.S.; Kamath, A.; Pöllinger, A.; Pahud de Mortanges, A.; Obmann, V.C.; Dammann, F.; Senn, W.; Reyes, M. DreamOn: A data augmentation strategy to narrow the robustness gap between expert radiologists and deep learning classifiers. *Front. Radiol.* **2024**, *4*, 1420545. [\[CrossRef\]](#)
34. Tran, T.L.C.; Huang, Z.-C.; Tseng, K.-H.; Chou, P.-H. Detection of bottle marine debris using unmanned aerial vehicles and machine learning techniques. *Drones* **2022**, *6*, 401. [\[CrossRef\]](#)
35. Hao, S.; Han, X.; Guo, Y.; Wang, M. Decoupled low-light image enhancement. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2022**, *18*, 1–19. [\[CrossRef\]](#)
36. Singh, S.; Kumari, R.; Pallavi, P.; Saurabh, P. A systematic review of deep learning methods for low-light image enhancement and object detection. *Discov. Appl. Sci.* **2026**, *8*, 241. [\[CrossRef\]](#)
37. Djurović, I. BM3D filter in salt-and-pepper noise removal. *EURASIP J. Image Video Process.* **2016**, *2016*, 13. [\[CrossRef\]](#)
38. Li, C.; Zhou, A.; Yao, A. Omni-dimensional dynamic convolution. *arXiv* **2022**, arXiv:2209.07947. [\[CrossRef\]](#)
39. Kuang, X.; Tao, B. ODGNet: Robotic grasp detection network based on omni-dimensional dynamic convolution. *Appl. Sci.* **2024**, *14*, 4653. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.