applied sciences

MDPI

Article

An Improved YOLO11n Model for Multiclass Object Detection of Chariot-and-Horse Elements in Images of Han Pictorial Stones

Yue Hu [1], Hui Chen [1], Weifeng Hu [2,*] and Qiang Wang [1]

[1] School of Design, Jiangnan University, Wuxi 214122, China; 7230306004@stu.jiangnan.edu.cn (Y.H.); 7240310020@stu.jiangnan.edu.cn (H.C.); wangqiang@jiangnan.edu.cn (Q.W.)

[2] School of Digital Technology & Innovation Design, Jiangnan University, Wuxi 214122, China

* Correspondence: hwf2011@jiangnan.edu.cn

Abstract

Han pictorial stones constitute an important component of China's ancient cultural heritage, and their depictions of chariots, horses, and human figures provide valuable visual evidence for investigating travel activities in the Han dynasty. The identification and localization of chariot-and-horse elements provide a foundation for the retrieval and statistical analysis of relevant materials. However, current documentation practices rely largely on manual identification and transcription, limiting the efficiency of large-scale processing. In Han pictorial stone images, weak contours, textural interference, dense occlusion, and scale variation pose substantial challenges for reliable object detection. Accordingly, 539 images were collected and annotated with 3736 instances. Building on this dataset, an improved YOLO11n model was proposed, incorporating the RubbingDown multi-branch downsampling structure, C3k2GRM module, and task-adapted ASF multi-scale fusion pathway for information preservation, contextual representation, and cross-scale feature fusion. The proposed model achieved mAP@0.5, mAP@0.5:0.95, and recall of 88.0%, 54.5%, and 86.5%, respectively, representing improvements of 5.5, 2.8, and 9.9 percentage points, respectively, over the YOLO11n baseline. The model maintained 2.58 M parameters, while GFLOPs increased from 6.3 to 7.3. The detection outputs could support future workflows for documenting, retrieving, quantitatively analyzing, and manually verifying chariot-and-horse elements in Han pictorial stone images.

Keywords: Han pictorial stones; chariot-and-horse elements; cultural heritage; object detection; YOLO11; multi-scale feature fusion

Academic Editor: Keun-Chang Kwak

Received: 21 September 2026
Revised: 5 November 2026
Accepted: 7 November 2026
Published: 9 November 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) license.

1. Introduction

Han pictorial stones constitute an important part of China's ancient cultural heritage, preserving a rich body of visual materials from Han dynasty tombs and shrines. Among the various pictorial themes, scenes composed of human figures, chariots, and horses provide visual evidence of travel-related activities [1,2]. The meanings conveyed by these images are closely related to the tomb or shrine contexts in which they occur [3] , while their visual content provides important evidence for understanding the hierarchical order and religious beliefs of Han society [2,4]. Elements such as horses, wheels, chariot boxes, canopies, and chariot occupants form the basic units for documenting and comparing these images. However, relevant image materials are distributed across ar-

chaeological reports, pictorial stone catalogues, digital collections, and rubbing archives. When examining these images, researchers still rely primarily on manual identification and transcription to record the objects depicted. The retrieval and statistical analysis of these elements require information on their categories, quantities, and locations, while the need to process images individually limits the efficiency and scalability of such work.

Digital preservation provides a foundation for image access, but element retrieval and statistical analysis require the further extraction of category and spatial location information for individual elements from the images. Classification, retrieval, detection, and segmentation serve different analytical purposes in cultural heritage image analysis, and their practical effectiveness depends on factors such as data quality, scene variation, and the involvement of domain expertise [5–7]. Classification and retrieval facilitate image categorization and record matching; however, when the focus shifts to specific objects within an image, both their categories and spatial locations need to be identified. Object detection simultaneously provides category and location information, thereby offering a technical basis for the batch localization and subsequent documentation of chariot-and-horse elements.

Applying object detection to the organization of Han pictorial stone image materials containing chariot-and-horse elements requires both accurate recognition and comprehensive documentation of individual elements. Previous studies have shown that downsampling strategies, feature extraction methods, and bounding box processing mechanisms have been optimized to improve the detection of human figures in Han pictorial stone images [8], while adaptive convolution kernels have been employed to enhance object detection in rock art images [9]. Beyond object detection, semantic enhancement and text reorganization have been used to improve the correspondence between image content and textual descriptions of Han pictorial stones [10]. The integration of object detection with image–text semantic matching provides a potential means of automatically identifying pictorial elements and associating the detected elements with corresponding documentary records. Images containing chariot-and-horse elements typically include multiple categories, including horses, human figures, and vehicle components, which need to be identified and documented individually. Multiple instances of the same category may also occur within a single image. Therefore, multi-class object detection requires a unified category definition that enables individual instances to be distinguished and localized and the detected elements to be associated with their source images, thereby supporting subsequent retrieval, quantitative analysis, and manual verification.

However, accurate recognition and localization of horses, human figures, and vehicle components in Han pictorial stone images remain challenging due to weak fine-line contours, mutual occlusion among objects, and scale variation. Weathering and damage to the stone surface, together with noise introduced during rubbing or scanning, may further obscure object boundaries. Cracks and paper textures may resemble carved lines, making it necessary to preserve fine-line features while distinguishing genuine object contours from background interference. When horses, human figures, and vehicle components are densely distributed and mutually occluded within the same image, local visual features alone may be insufficient for reliable category identification; contextual information from surrounding structures is, therefore, needed to reduce missed detections and localization errors. Differences in image clarity and contrast across different types of source materials further complicate object recognition. In addition, chariot-and-horse elements vary considerably in size and aspect ratio, and instances of the same category may occupy substantially different proportions of the image, depending on whether a complete image or a cropped region is used. Effective feature extraction, therefore, re-

quires the preservation of local details, the incorporation of structural context, and cross-scale information exchange to accommodate variations in object representation.

To meet these documentation requirements and overcome the associated detection challenges, a dedicated dataset of chariot-and-horse elements was constructed primarily from Han pictorial stone images, with five categories uniformly annotated: horses, wheels, chariot boxes, canopies, and chariot occupants. Based on YOLO11n, a lightweight model in the You Only Look Once (YOLO) family, an improved detection model was developed while maintaining its lightweight parameter scale. RubbingDown was introduced to preserve fine-line features during downsampling, C3k2GRM was employed to integrate local and contextual information, and the multi-scale feature fusion strategy of Attentional Scale Sequence Fusion (ASF)–YOLO was adapted for the detection head. Under consistent experimental conditions, the effects of these three modifications on detection performance and computational cost were evaluated, recognition performance and error patterns across different object categories were analyzed, and the potential use of the detection results for organizing image materials, together with the interpretive limitations of such use, was examined.

The main contributions of this study are as follows:

(1) A dataset centered on Han pictorial stones was built for chariot-and-horse element detection, and a five-class object detection task was defined. The dataset is based on 539 original images, with a total of 3736 object instances annotated using bounding boxes, covering five categories, horse, wheel, chariot box, canopy, and chariot occupant, laying the data foundation for subsequent element retrieval and statistical analysis.

(2) An improved YOLO11n detection model is proposed to address weak contours, complex structural patterns, and scale variation in Han pictorial stone images. RubbingDown extends space-to-depth (SPD)-based downsampling [11] into a three-branch unit by adding a StripConv pathway for directional convolution and a Sobel-initialized learnable EdgePrior pathway, followed by spatial and channel re weighting. C3k2GRM replaces selected C3k2 bottlenecks with gated multi-receptive field convolution (GRConv) comprising two fixed-dilation branches, followed by the existing mixed local channel attention (MLCA) [12]. The task-adapted ASF multi-scale fusion pathway reuses Scale-Sequence Feature Fusion (SSFF) and the Triple Feature Encoder (TFE) [13], with feature connections and channel dimensions adapted to YOLO11n and additive fusion into P3. These modifications link fine-line preservation during downsampling, local-context selection within feature extraction, and high-resolution cross-scale fusion to the weak contours, occlusion, and scale variation of chariot-and-horse ele ments.

(3) Under consistent experimental settings, the proposed model was compared with five additional lightweight YOLO detectors, with YOLO11n retained as the direct baseline. Component-level and combinatorial ablation experiments, category-level performance metrics, threshold-based and error analyses, and representative visualizations were used to evaluate changes in overall performance, inter-category behavior, and typical detection scenarios. The results show higher recall and mean average precision values on the current dataset and indicate that the resulting detections could support subsequent retrieval, quantitative analysis, and manual verification of chariot-and-horse elements.

2. Literature Review

2.1. Cultural Heritage Image Analysis and Object Detection

Cultural heritage image analysis covers multiple levels of visual understanding, from image-level classification to cross-modal retrieval and object-level localization. At the image level, the classification method links visual features with chronological, the-

matic, or other content-related attributes. For example, color and shape descriptors have been combined with deep visual features to support the dating of Mogao Grottoes murals [14]; at the same time, image classification is also used to identify field photographs containing painted rock art, thus promoting large-scale screening of visual materials [15] . Cross-modal retrieval further expands the dimension of analysis by establishing a semantic correspondence between images and textual records. For Han pictorial stones, researchers used semantic enhancement and text reorganization to improve image–text retrieval by aligning image content with textual descriptions [10]. Similarly, in the study of ancient murals, style recognition and discriminative visual features are combined with text topic analysis, aiming to link visual features with the thematic organizational structure of cultural heritage materials [16]. Although these methods have their own emphasis, they usually take the whole image or image–text pair as the basic unit of analysis. Therefore, these methods are effective in image classification, content matching, and thematic organization, but they cannot directly identify and localize individual objects in the image.

Object detection addresses this limitation by simultaneously predicting the category and spatial location of individual object instances. Its performance in cultural heritage imagery, however, is closely related to the form of the source data, the characteristics of the target objects, and the complexity of the visual scene. For rock art, three-dimensional scan data have been transformed into enhanced two-dimensional representations to facilitate the localization and classification of motifs, including boats and anthropomorphic figures [17]. For in situ photographs of rock carvings, YOLO-based detection has been compared with template matching using the scale-invariant feature transform (SIFT) and random sample consensus (RANSAC) to evaluate different recognition strategies for field-acquired imagery [18]. Model-level adaptations have also been explored to account for complex object appearances and background interference. Attention mechanisms and modified convolutional structures have been incorporated into models for the recognition and detection of Yunnan Jiama prints and Terracotta Warriors [19,20]. Of particular relevance to the present study, human figure detection in Han pictorial stone images has been investigated using an improved YOLOv5 model incorporating SPD-Conv, Coordinate Attention, distance intersection over union-based non-maximum suppression (DIoU-NMS), and Alpha-IoU Loss. The resulting model improved the detection of Fuxi-Nuwa and music and dance figures on a self-constructed dataset, with Coordinate Attention contributing substantially to the reported performance gains [8]. Collectively, these studies suggest that object detection in cultural heritage imagery benefits from task-specific adaptation of the input representation, recognition strategy, and model architecture to the properties of the source images, target objects, and scene context.

Variation in image media introduces an additional source of complexity in cultural heritage analysis. The same visual content may appear very differently across photographs, rubbings, line drawings, and historical prints because of changes in texture, contrast, edge definition, and background characteristics. Accordingly, methods developed for one type of source material may not transfer directly to another. Historical illustration detection has used style transfer from photographs to historical illustrations together with pseudo-label training to reduce the appearance gap between training data and target images [21]. For oracle bone rubbings, local matching has been performed by combining character-level features with spatial structural information, enabling correspondences to be established across different rubbing versions derived from the same source [22]. In bronze inscription analysis, mixture of experts adapters have been incorporated into detection and recognition pipelines to accommodate photographs, rub-

bings, and line drawings, while also addressing low-frequency character categories [23]. These studies address media variation at different levels: through transformation of the training data, explicit modeling of local structural correspondence, and adaptation of the recognition model. Although the underlying tasks differ, they consistently show that performance under cross-media conditions depends on both the form of the input data and the definition of the target to be recognized.

When the recognition results are used for the organization and retrieval of documentary materials, the detected objects need to be associated with the corresponding textual, collection, or contextual records. For example, researchers have used ontological relationships to correlate the detected Egyptian shabti figurines and hieroglyphic names with cross-collection records in Europeana [24]. In oracle bone research, the OBIMD dataset realizes the alignment of rubbings, line drawings, and transcriptions, and provides structured annotations for character positions, sentence groups, and reading order [25]. The relevant research on historic buildings extends this association beyond textual records and enriches color and semantic information by combining design elements in images with laser point clouds [26]. Although these studies involve different types of cultural heritage data, they face a common need: in order to achieve a reliable association with external records, visual entities must be represented under a consistent annotation and documentation framework. Although object detection can provide category and location information for retrieval, counting, and record association, the relationships between objects and their deeper cultural meanings still need to be interpreted in combination with the scene context and other forms of evidence.

Overall, previous studies have addressed cultural heritage image analysis from multiple perspectives, including image-level classification, object detection, cross-media adaptation, and the linkage of visual content with documentary records. These studies also show that effective organization of cultural heritage imagery depends not only on reliable recognition but also on clearly defined object categories and consistent annotation and documentation schemes. Han pictorial stone images containing chariot-and-horse elements, however, present a more specific multi-object detection problem, in which human figures, animals, and vehicle components frequently coexist within complex visual scenes. This combination of object types motivates the examination of feature extraction and fusion strategies in the following section.

2.2. Feature Extraction and Multi-Scale Fusion for Object Detection

Heritage object detection is usually developed based on established object detection frameworks. The main difference between these frameworks lies in the representation of candidate objects and the prediction generation mechanism. The region-based convolutional neural network (R-CNN) and Faster R-CNN adopt region proposal-based detection strategies [27,28]. YOLO and the Single Shot MultiBox Detector (SSD) perform single-stage dense prediction [29,30]. The Detection Transformer (DETR) recasts object detection as a set prediction problem [31]. In addition, the training objective can also be adjusted independently of the detection paradigm; for example, Focal Loss reweights hard and easy samples during optimization to mitigate class imbalance [32]. Although these methods differ in prediction formulation and training strategies, they all rely on effective feature representation and multi-scale information integration to distinguish the appearance of objects and adapt to variations in object size.

Feature extraction needs to make a trade-off between detection accuracy and computational efficiency [33]. For low-resolution targets with limited appearance information, multi-scale representation and contextual modeling are particularly important to improve feature discriminability [34]. In rubbings, line drawings, and line-engraved im-

ages, fine lines and local contours contain key structural information, but cracks, paper texture, and uneven inking may also produce similar response characteristics. Therefore, existing studies have explored multiple ways to retain and enhance structural cues in the process of feature learning. In the oracle bone character detection task, researchers use pseudo-category labels to provide auxiliary structural supervision and introduce clustering strategies and loss constraints to optimize feature representation [35,36]. Another strategy focuses on the downsampling process itself: SPD-Conv rearranges spatial information into the channel dimension before performing stride-1 convolution; compared with conventional strided convolution, this method can retain more spatial details when processing low-resolution images and small objects [11]. The above methods work in different stages of feature extraction: the former introduces structure-aware supervision in the learning process, while the latter aims to reduce the loss of information during spatial downsampling. However, for line-dominated heritage imagery, it is not enough to retain only fine structural details; effective feature extraction must also be able to distinguish between meaningful object contours and visually similar background textures.

Preserving local detail alone is not sufficient for reliable discrimination because the significance of a local response often depends on its surrounding context and the overall object structure. Attention mechanisms address this issue by reweighting feature responses across channels, spatial locations, or branches, enabling the network to emphasize input-dependent information [37]. MLCA integrates local and global cues to adaptively modulate detection features [12]. Another relevant strategy is to make the convolution operation itself adaptive. In the rock art detection task, omni-dimensional dynamic convolution is introduced into the backbone and feature fusion stages of YOLOv8; this enables the convolution kernels to be adjusted according to the input so that feature discrimination is improved under complex backgrounds and subtle inter-class variation [9]. Although the working principles of attention-based modulation and dynamic convolution are different, both aim to enhance the selectivity of feature responses by integrating information beyond isolated local patterns. Local features are crucial to describing fine morphology, while broader contextual cues help to eliminate ambiguity and capture the structural configuration of objects. However, the relative contribution of these two types of features often depends on object visibility, occlusion, and scene complexity. Feature extraction for heritage imagery should, therefore, balance fine-detail preservation with context-aware response selection rather than treating them as independent design objectives.

Beyond feature selection within a single resolution level, robust detection of objects with varying sizes also depends on effective information exchange across feature scales. The Feature Pyramid Network (FPN) propagates high-level semantic information to higher-resolution feature maps through a top-down pathway with lateral connections [38]. In heritage element detection, a bidirectional feature pyramid has been combined with a dynamic detection head for the analysis of Jiangnan private garden images, enabling richer interactions among feature levels [39]. More elaborate fusion strategies have also been proposed to aggregate multi-scale information adaptively. ASF-YOLO, originally developed for cell instance segmentation, combines features from different scales through SSFF and TFE, followed by channel- and position-aware feature weighting [13]. Although these approaches all exploit complementary information across feature levels, they differ in how features are propagated, aggregated, and reweighted after fusion. Their effectiveness, therefore, depends on the scale distribution of the target objects and the requirements of the specific task. For chariot-and-horse element detection, the multi-scale fusion strategy of ASF-YOLO cannot be assumed to transfer directly and must be adapted to the feature hierarchy and object-scale characteristics of Han pictorial

stone images.

Taken together, the studies reviewed above provide useful foundations for object localization, fine-detail preservation, contextual feature modeling, and multi-scale information fusion. However, because they were conducted using different image sources, category systems, and evaluation protocols, their results do not reveal how these feature processing strategies interact when combined within a unified chariot-and-horse detection task. Such a task requires clearly defined object categories and annotation boundaries, as well as systematic evaluation of how fine-detail preservation, local-context integration, and cross-scale fusion influence the detection of different object types. Category imbalance and differences in object-scale distribution may further affect class-specific performance [40], making direct performance comparisons across studies an unreliable basis for judging the effectiveness of individual feature processing methods. Accordingly, this study formulates a unified five-class detection task for horses, wheels, chariot boxes, canopies, and chariot occupants, using consistent category definitions, data partitions, and evaluation criteria to examine the individual and combined contributions of detail preservation, contextual representation, and cross-scale feature fusion.

3. Materials and Methods

3.1. Research Objects and Category Definitions

The detection task covers five object categories: horses, wheels, chariot boxes, canopies, and chariot occupants. The category definitions were formulated according to the compositional characteristics of chariot-and-horse scenes in Han pictorial stones, together with visual distinguishability, annotation consistency, and their potential use in subsequent retrieval and statistical analysis. Detailed category definitions and annotation rules are provided in Section 3.2. The model predicts category labels, bounding boxes, and confidence scores, which can support subsequent retrieval, quantitative analysis, and manual verification of chariot-and-horse elements.

3.2. Data Sources, Category System, and Annotation

The materials involved include regional research on Han pictorial stones, archaeology and art history catalogs, scanned editions of publications, digital archives maintained by research institutions, online museum collections, and public image repositories. In terms of geographical distribution, these materials cover the main distribution areas of Han pictorial stones such as Shandong, Henan, Jiangsu, Anhui, Zhejiang, Shaanxi, Shanxi, and Sichuan. Among the 539 original images finally included, 346 were taken from catalogs, book plates, and related special publications, and 193 came from various digital resources, including the Fu Ssu-nien Library of the Institute of History and Philology, the digital archives of other institutions, online museum collections, and public image repositories.

The candidate images have been screened by automatic quality inspection and manual review one by one. The automatic check records indicators such as image size, contrast, and edge response. During manual image review, each candidate was required to contain a reliably identifiable chariot body, wheel, or other clear chariot-and-horse structural feature. Images with severe blur, large-area damage, too small subjects, a large number of irrelevant areas, lack of stable components, and images with duplicate content or high similarity and poor quality are eliminated, and the reasons for elimination are recorded for reference. The file-level check did not find any completely duplicate images. In the source-level organization stage, different crops, resized versions, rubbings, or photographs from the same original plate are classified into the same source group. These 539 images correspond to a total of 388 source groups, of which 74 groups contain

multiple visual versions. In terms of data management, the final 539 images were divided into two categories: 502 core Han pictorial stone images and 37 images of pictorial bricks, tomb murals, or objects whose supporting medium remains to be verified. The reason for retaining these 37 supplementary images is that although the published plates and digital records often cannot clearly determine the supporting medium, the chariot-and-horse content contained in them is still valuable to the established detection task. These images are annotated and partitioned according to the same rules as the core images. These images together constitute a chariot-and-horse element dataset centered on Han pictorial stones. The samples vary in resolution, grayscale, background color, rubbing quality, crop scale, and degree of damage, reflecting differences introduced during collection, publication, and digitization.

After screening, the dataset contains 539 original images. Using the X-AnyLabeling annotation tool, a total of 3736 bounding boxes were annotated. Based on object visibility, the consistency of the annotation, and the value of the follow-up research, the category system comprises five classes: horse, wheel, chariot box, canopy, and chariot occupant. In this study, "chariot occupant" refers to a human figure depicted as a driver or passenger of a chariot. Nearby figures are not included merely because of their spatial proximity or compositional association with the chariot. This definition was applied consistently throughout the annotation process. For occluded or damaged objects, as long as their visible parts are sufficient to support reliable category determination and comply with unified boundary rules, they are included in the dataset.

As summarized in Table 1, the 539 original images were divided into 431 training images, 54 validation images, and 54 test images, containing 2966, 374, and 396 annotated objects, respectively. Across the complete dataset, chariot occupants and horses are the two largest categories, with 1216 and 905 instances; wheels and canopies contain 762 and 694 instances, whereas chariot boxes contain 159. This distribution reflects the characteristics of chariot-and-horse scenes: a chariot is commonly associated with multiple horses and human figures, whereas the chariot box may be only partly visible because of occlusion, damage, or local cropping. Figure 1 shows the representative annotated objects in various categories and visually presents the differences within the same category in terms of contour completeness, pose, scale, and rubbing quality.

Table 1. Dataset partitions and numbers of objects in the five categories.

Dataset Subset	No. of Images	Horse	Wheel	Chariot Box	Canopy	Chariot Occupant	Total Objects
Original training images	431	725	602	124	551	964	2966
Validation set	54	89	77	16	74	118	374
Test set	54	91	83	19	69	134	396
Total	539	905	762	159	694	1216	3736

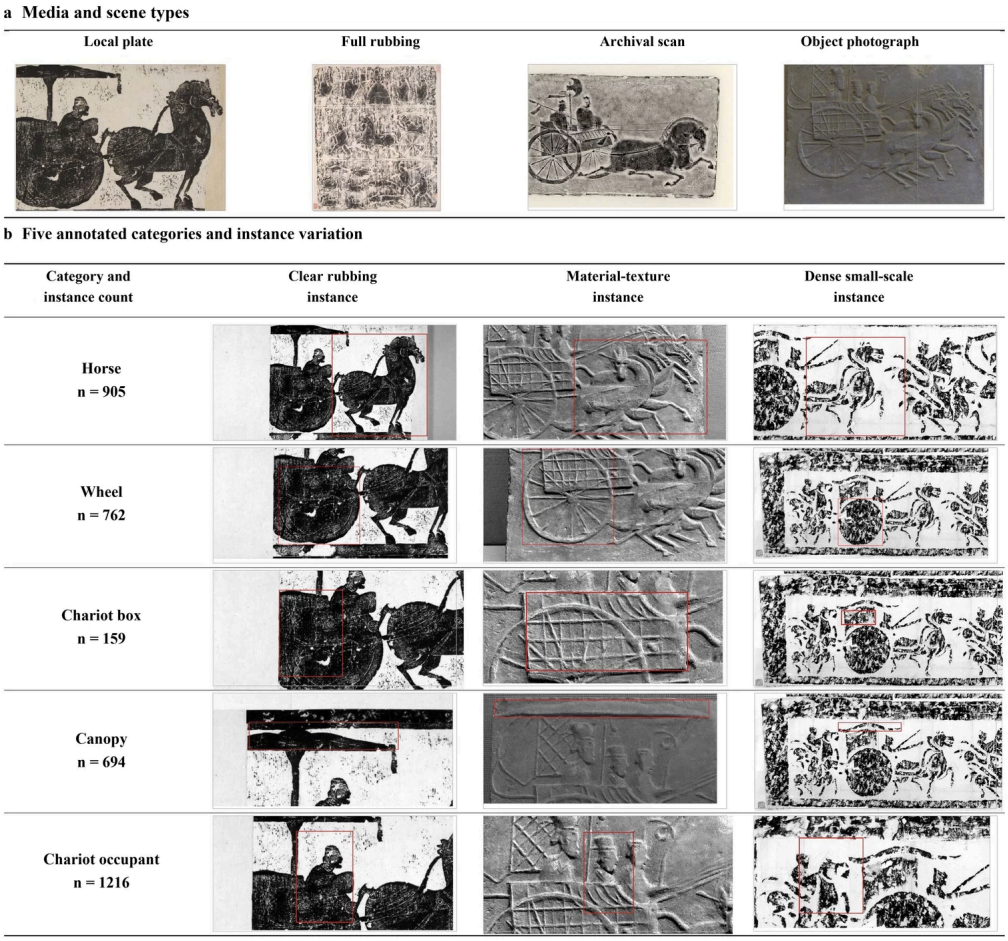


Figure 1. Visual composition of the chariot-and-horse element dataset centered on Han pictorial stones. (**a**) Representative original images from different media and scene types, including cropped plates, full-view rubbings, archival scans, and photographs of physical objects; (**b**) instance counts and representative local samples for the five annotated categories: horse, wheel, chariot box, canopy, and chariot occupant. The red solid boxes delineate representative instances of the corresponding annotated categories.

Each object was annotated using a bounding box fitted as closely as possible to its visible, recognizable extent. For horses, human figures, and vehicle components that cover each other, as long as a single instance retains visible structural features sufficient to determine the category, a separate bounding box is created for it. Multiple horses, wheels, or human figures in the same image are annotated one by one, instead of combining them into a single large bounding box. Although the wheels, chariot boxes, and canopies are the same vehicle components, they are labeled independently to support subsequent component-level retrieval and statistical analysis.

These five types of labels record visible objects separately, and the relative positions of different types of bounding boxes can provide clues for sorting out the connection between vehicle components. In a complete scene, the combination of horses, human figures, and vehicle components can be checked with the original image; and if the image is damaged or locally cropped, the surviving components can still be recorded. The task directly outputs the object category and location information. The vehicle affiliation relationship between objects and the driving or riding relationship needs to be confirmed in the subsequent material organization stage.

The annotation file is stored in YOLO format. For each object instance, the class in-

dex and the normalized center coordinates, width, and height of the bounding box are recorded. After the completion of initial annotation, according to the unified category definition and bounding box rules, it is verified whether there are omissions, incorrect labels, category inconsistencies, and improper boundaries to ensure that images from different media (including rubbings, scanned plates, and photographs of physical objects) are annotated according to consistent standards. The final numbers of images and category instances in the dataset are shown in Figure 1b.

3.3. Data Partitioning and Augmentation Strategies

The original data is partitioned into the training set, validation set, and test set in a ratio of about 8:1:1. The training set contains 431 original images, and the validation set and test set each contain 54 images. The partitioning operation is based on the source-group level: different crops, resized versions, rubbings, or photographs from the same original plate are classified into the same subset to avoid source-related overlap between the training set, validation set, and test set. Only the training set is processed using offline augmentation. Six variants are generated for each original training image: horizontal flipping, vertical flipping, combined horizontal and vertical flipping, random rotation, random scaling, and saturation adjustment. These variants, together with the original images, make up 3017 training files. Each augmented image is stored together with its corresponding detection labels for training. During label post-processing, the bounding boxes that exceed the usable image region or no longer meet the validity checks after geometric transformation are eliminated. In the end, 91 of the expected 20,762 bounding box instances were eliminated, so that the augmented training set contained 20,671 valid bounding boxes. Vertical and combined horizontal–vertical flips were included to increase orientation diversity in the training data. Image coordinates and bounding boxes were transformed synchronously, while the validation and test sets remained unaugmented. Their effect is evaluated in Section 4.10.

To further characterize the inputs actually received by the model during training, Figure 2 summarizes the category, center position, and scale distributions of 20,671 annotated boxes across the 3017 training images after offline augmentation. The five categories comprise 5066 horses, 4208 wheels, 867 chariot boxes, 3810 canopies, and 6720 chariot occupants. The normalized center coordinates cover the principal regions of the image, indicating variation in target positions. Bounding box widths and heights are concentrated primarily in smaller ranges while retaining a limited number of larger instances, reflecting the diversity of target positions and scales in the training samples.

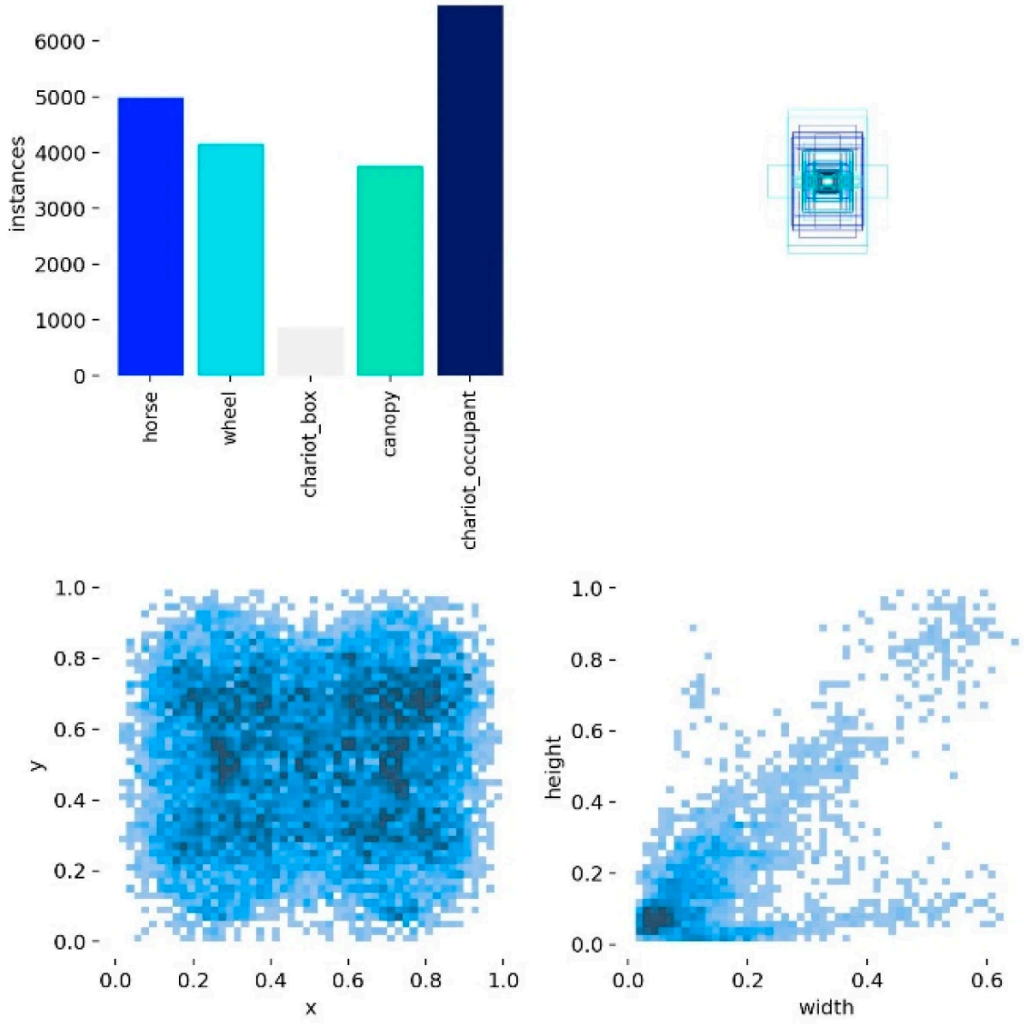


Figure 2. Distribution of annotated boxes in the training set after offline augmentation. The upper-left panel shows category-wise instance counts; the upper-right panel visualizes the relative sizes and aspect ratios of bounding boxes by superimposing them at a common center; the lower-left panel shows the distribution of normalized center coordinates; and the lower-right panel shows the distribution of normalized box widths and heights. The statistics are based on 20,671 valid object instances in 3017 training images, including 5066 horses, 4208 wheels, 867 chariot boxes, 3810 canopies, and 6720 chariot occupants. The colors of the bars and bounding-box outlines distinguish the five annotated categories, whereas darker shading in the density plots indicates a higher concentration of instances.

Geometric and color transformations can increase the visual variability of training samples, but transformation intensity and label processing must be determined according to task semantics [41]. Online augmentation was additionally applied during training, including hue, saturation, and value (HSV) perturbation, translation, scaling, horizontal flipping, and Mosaic. Offline augmentation provides reproducible media and geometric variations, whereas online augmentation further diversifies local combinations within each training batch, thereby reducing the model's dependence on fixed layouts, a single level of rubbing contrast, and specific target locations. Figure 3 presents three groups of chariot-and-horse rubbings with complete compositions and clear visual features, showing representative correspondences between the original images and versions produced by horizontal flipping, random rotation, random scaling, and saturation adjustment. These examples emphasize the weak contours, rubbing textures, and dense

linework addressed by the proposed method.

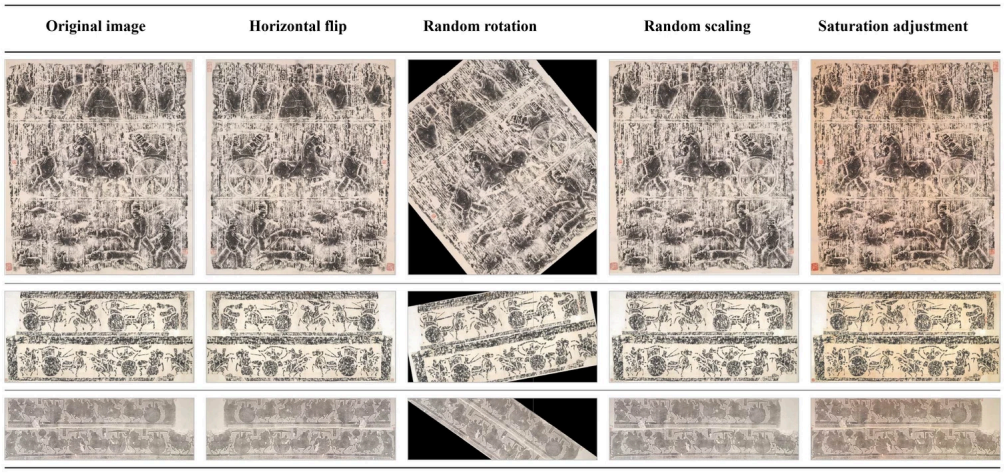


Figure 3. Representative results of offline data augmentation for chariot-and-horse rubbing samples. Each row shows the original rubbing followed by the corresponding horizontally flipped, randomly rotated, randomly scaled, and saturation-adjusted versions.

3.4. YOLO11n Baseline and Overall Framework

Ultralytics YOLO11n was selected as the baseline detector [42]. The model uses a hierarchical convolutional backbone to extract multi-scale features, expands the receptive field and strengthens high-level representations through Spatial Pyramid Pooling-Fast (SPPF) and C2 Position-Sensitive Attention (C2PSA), performs top-down and bottom-up feature propagation in the neck, and finally conducts classification and bounding box regression at the P3, P4, and P5 scales. The n-scale model was selected to evaluate the effect of structural modifications on cultural heritage images under a relatively small parameter count and computational cost. The baseline model contains approximately 2.58 M parameters and requires approximately 6.3 giga floating point operations (GFLOPs).

The proposed improved YOLO11n model introduces structural modifications at three locations in YOLO11n. First, conventional strided convolutions before P3/8 and P4/16 in the backbone are replaced with RubbingDown. Second, selected C3k2 modules at multiple stages of the backbone and in the P3 detection branch are replaced with C3k2GRM. Third, the task-adapted ASF multi-scale fusion pathway is introduced into the detection head. In the figure, TFE is denoted by Zoom_cat, and SSFF is denoted by ScalSeq. The fused result is added to the P3 detection branch through Add. Detection outputs are still produced at the P3, P4, and P5 scales.

Across the complete information flow, the input image first passes through two strided convolutions to reach the P2/4 feature stage. It then proceeds through C3k2GRM and RubbingDown to form the P3/8 and P4/16 features, followed by deeper C3k2GRM, strided convolution, C3k2, SPPF, and C2PSA modules to obtain the P5/32 high-level semantic representation. In the top-down pathway of the detection head, features associated with P5, P4, and P3 undergo channel adjustment and TFE (Zoom_cat)-based scale alignment, enabling cross-level concatenation at the P4 and P3 resolutions. The P3 fusion node further uses C3k2GRM to strengthen the representation of small-scale objects. The bottom-up pathway then applies two rounds of downsampling, concatenation, and C3k2 aggregation to form the P4 and P5 detection features. In parallel, SSFF (ScalSeq) receives P3, P4, and P5 backbone features, unifies their channel dimensions, aligns their spatial scales, aggregates the scale sequence, and supplements the P3 detection branch through Add. The enhanced P3 feature, together with the P4 and P5 features, is input to the

Detect layer to produce three-scale predictions for the five object categories. Figure 4 illustrates the proposed model's backbone, multi-scale feature fusion, and three-scale detection information flow.

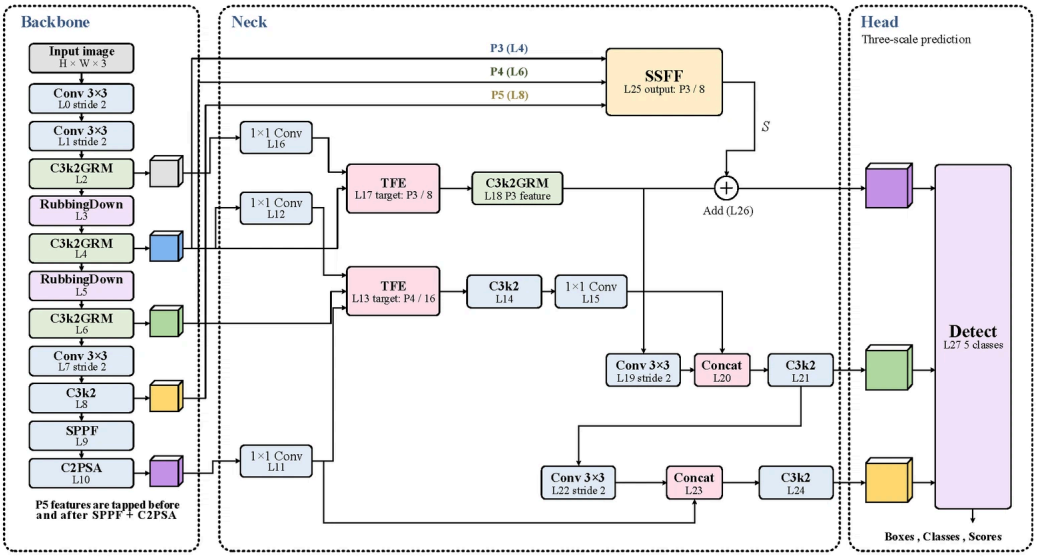


Figure 4. Overall architecture of the proposed model. The backbone employs RubbingDown when generating the P3/8 and P4/16 features and deploys C3k2GRM at multiple levels. In the multi-scale feature fusion component, TFE (Zoom_cat) and SSFF (ScalSeq) perform top-down fusion and scale-sequence feature aggregation, and the fused result is added to the P3 detection branch through Add. The detection module predicts bounding boxes, categories, and confidence scores for the five object classes at the P3, P4, and P5 scales. Arrows indicate the direction of feature propagation, while the color coding distinguishes the standard YOLO11n components, RubbingDown, C3k2GRM, and the task-adapted ASF pathway, as specified in the legend.

3.5. RubbingDown Downsampling for Preserving Fine-Line Structures in Han Pictorial Stone Images

Object contours in Han pictorial stone chariot-and-horse images are often fine and faint and are accompanied by uneven inking, paper texture interference, and textural degradation. When reducing spatial resolution, strided convolution may weaken directional fine-line responses such as wheel spokes, chariot shafts, axles, and incomplete contours. To address the loss of fine-line information during shallow downsampling, this study constructs the RubbingDown downsampling mechanism. Given an input feature $X \in R^{(B\times C_1\times H\times W)}$, the mechanism compresses the spatial dimensions to $H/2 \times W/2$ while preserving shallow information relevant to locating the five object categories from three perspectives: pixel organization, directional structure, and edge priors.

RubbingDown consists of three parallel pathways—SPD, StripConv, and EdgePrior—and a collaborative fusion unit. The SPD pathway uses a space-to-depth transformation to avoid directly discarding adjacent pixels. The StripConv pathway describes elongated structures with pronounced directionality through horizontal, vertical, and local square convolutions. The EdgePrior pathway initializes learnable depthwise convolutions with Sobel operators, explicitly introducing priors for rubbing lines and contours. Each branch produces a feature map at half the input height and width before fusion. The outputs are concatenated along the channel dimension, reweighted by spatial and channel attention, and passed through a 1×1 channel projection. This projection retains the spatial resolution of the branch outputs. Figure 5 shows how these operations replace a single stride-2 convolution.

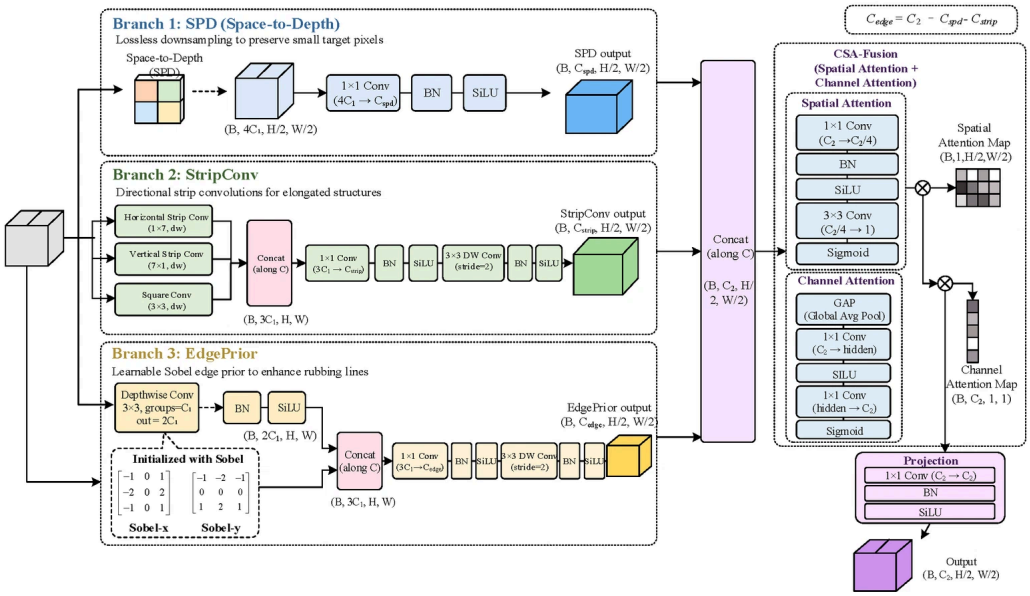


Figure 5. Complete structure of the RubbingDown detail-preserving downsampling module. The SPD, StripConv, and EdgePrior pathways encode pixel rearrangement, directional structure, and edge responses, respectively. After the features from the three pathways are concatenated, spatial and channel attention jointly select informative responses, followed by a 1 × 1 projection to produce the downsampled output. Arrows indicate feature propagation; blue, green, and yellow identify the SPD, StripConv, and EdgePrior branches, respectively, pink marks within-branch concatenation, and purple highlights cross-branch fusion and projection.

SPD-Conv [11] couples spatial rearrangement with a non-strided convolution. RubbingDown retains spatial rearrangement as its SPD pathway and introduces parallel StripConv and EdgePrior pathways, followed by attention-based fusion and projection. Together, these components form a multi-branch downsampling unit that jointly models pixel organization, directional structures, and edge responses. The SPD pathway draws on the space-to-depth transformation [11], partitioning the input feature according to odd and even rows and columns into four complementary sub-features, which are then concatenated along the channel dimension. A 1 × 1 convolution subsequently compresses the channels to produce the SPD branch output. In the equations below, each convolution operator is implemented as a sequence comprising a convolutional layer, batch normalization (BN), and Sigmoid Linear Unit (SiLU) activation.

$$X_{spd} = \varphi_{1\times1}(\mathrm{Cat}(X_{00}, X_{10}, X_{01}, X_{11})) \quad (1)$$

Here, X_{00}, X_{10}, X_{01}, and X_{11} correspond to the four fixed sampling positions within each adjacent 2 × 2 region, and $\varphi_1{\times}_1$ denotes a 1 × 1 convolutional mapping. Spatial rearrangement first preserves the responses at four adjacent positions along the channel dimension, after which convolution performs channel compression, retaining information for local structural representation after downsampling. The StripConv pathway further applies 1 × 7, 7 × 1, and 3 × 3 depthwise convolutions to extract horizontal, vertical, and local structures, respectively. Let D(k × k, s) denote a depthwise convolution with a k × k kernel and stride s. After the three directional branches are concatenated, a 1 × 1 convolution first compresses the channels, followed by downsampling:

$$X_{strip} = D_{3\times3,s=2}(\varphi_{1\times1}(\mathrm{Cat}(D_{1\times7}(X), D_{7\times1}(X), D_{3\times3}(X)))) \quad (2)$$

The horizontal and vertical strip convolutions expand the effective coverage in a single direction, while the local square convolution supplements corners and neighbor-

hood textures. Their parallel combination enables the network to represent directional cues such as chariot shafts, axles, horse body contours, and wheel spokes while reducing the sensitivity of a single fixed convolution kernel to pose variation.

The EdgePrior pathway uses learnable grouped depthwise convolution kernels K_x and K_Y, initialized with the horizontal Sobel operator S_x and the vertical Sobel operator S_Y, respectively, to explicitly introduce edge priors. The Sobel kernels determine only the starting point of training; thereafter, K_x and K_Y are updated jointly with the other network parameters to adapt to the actual distributions of rubbing lines, paper texture, and stone-surface noise. The computation and initial conditions are defined as follows:

$$E_x = K_x * X, \quad E_y = K_y * X, \quad X_{edge} = D_{3\times3,s=2}(\varphi_{1\times1}(\mathrm{Cat}(X,E_x,E_y))) \tag{3}$$

$$K_x^{(0)} = S_x = (-1,0,1;-2,0,2;-1,0,1), \quad K_y^{(0)} = S_y = (-1,-2,-1;0,0,0;1,2,1) \tag{3a}$$

SPD is allocated approximately half of the target output channels to control the output size, while the remaining channels are divided as evenly as possible between StripConv and EdgePrior; the total number of channels across the three pathways is maintained at c_2. Their outputs are concatenated along the channel dimension to form Z. A spatial attention branch and a channel attention branch based on global average pooling (GAP) then generate complementary weights:

$$Z = \mathrm{Cat}(X_{spd},X_{strip},X_{edge}), \quad A_s = \sigma(f_s(Z)), \quad A_c = \sigma(f_c(\mathrm{GAP}(Z))) \tag{4}$$

$$Y = \varphi_{proj}(Z \odot A_s \odot A_c) \tag{5}$$

Here, Cat denotes concatenation along the channel dimension, ⊙ denotes element-wise multiplication, and φ_proj denotes a projection layer consisting of 1 × 1 convolution, Batch Normalization, and SiLU. Spatial attention highlights regions in the rubbing foreground that may contain valid object structures, while channel attention dynamically adjusts the relative contributions of pixel rearrangement, directional convolution, and edge pathways, enabling pixel details, directional structures, and edge information to form complementary representations. In line-dominated rubbings, cracks and paper texture can generate edges as well as object contours. Combining the learned edge responses with directional and rearranged features allows their contribution to be conditioned on the full fused representation. This design targets the retention of weak canopy supports, wheel spokes, and fragmented vehicle outlines while learning to reduce responses to background texture.

3.6. C3k2GRM Receptive Field Selection for Representing Scale Variation

Han pictorial stone chariot-and-horse images contain objects and local structures at various scales, exhibiting significant scale variation. Elements such as wheels, chariot boxes, and relatively complete vehicle contours require broader structural context, while fine-grained regions such as spokes, axle joints, contour corners, and damaged edges depend more strongly on local texture. Fixed receptive field convolution cannot readily meet these two representational requirements at the same time. To address this conflict, a C3k2GRM receptive field selection mechanism is constructed within the C3k2 framework: this mechanism uses the Bottleneck_GRM module to replace selected conventional bottlenecks and combines GRConv with the existing MLCA mechanism. Figure 6 shows three key components: GRConv, MLCA, and Bottleneck_GRM.

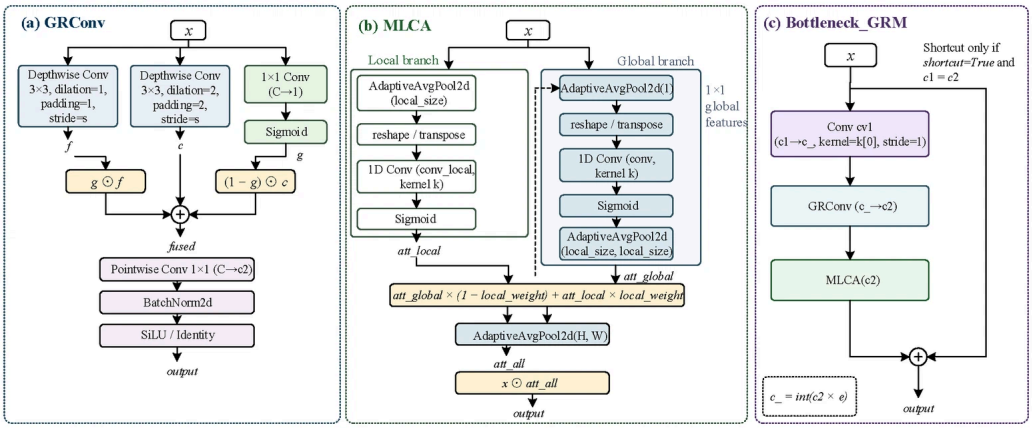


Figure 6. Key components of C3k2GRM. (**a**) GRConv adaptively fuses different receptive fields through fine-grained, coarse-grained, and gating branches; (**b**) MLCA jointly models local and global channel attention; (**c**) Bottleneck_GRM sequentially combines channel transformation, GRConv, and MLCA and applies a residual connection when the input and output channels are identical and the shortcut condition is satisfied. Arrows indicate feature propagation, while colors visually distinguish the computational branches and operations within each component.

GRConv replaces the conventional 3 × 3 convolution in the Bottleneck with two 3 × 3 depthwise convolution branches. The fine-grained branch uses a dilation rate of d = 1 and emphasizes local edges and textures. The coarse-grained branch uses a dilation rate of d = 2, forming an effective 5 × 5 receptive field without substantially increasing the parameter count and capturing broader object contours and structural context. Given an input X, a 1 × 1 convolution and a Sigmoid function generate a scalar gating weight G at each spatial location:

$$G = \sigma(\mathrm{Conv}_{1\times1}(X)) \tag{6}$$

The two branches form effective receptive fields of 3 × 3 and 5 × 5, respectively, and their outputs are expressed as follows:

$$F_f = \mathrm{DWConv}_{3\times3,d=1}(X), \quad F_c = \mathrm{DWConv}_{3\times3,d=2}(X) \tag{7}$$

$$F = G \odot F_f + (1 - G) \odot F_c, \quad Y_{GR} = \mathrm{SiLU}(\mathrm{BN}(\mathrm{Conv}_{1\times1}(F))) \tag{8}$$

The weights and biases of the 1 × 1 convolution in the gating branch are initialized to 0; consequently, G ≈ 0.5 at the beginning of training, and the two receptive fields contribute to the fusion in approximately equal proportions. As training proceeds, the gating weights are learned from the input features, allowing different spatial locations to use different proportions of local and contextual features. Compared with direct branch concatenation, location-wise gating does not expand the output channels and allows different regions within the same image to adopt different receptive field combinations.

The gating weights in Equation (8) are shared across channels at the same spatial location. A larger gate value assigns greater fusion weight to the fine-grained branch, whereas a smaller gate value gives greater weight to the coarse-grained branch. Both convolutional branches participate in the forward computation, and their dilation rates are fixed at 1 and 2. Adaptivity is realized through a continuously varying weighting ratio across spatial locations. After weighted fusion, pointwise convolution further mixes the channel information, so that spatial selection among receptive field branches is followed by channel transformation.

MLCA attention [12] is applied after GRConv to further coordinate local structure

and global semantics. The output size of local adaptive average pooling is set to 5 × 5, while the global branch uses global average pooling. R denotes sequence rearrangement of the local grid. The local and global branches use separate one-dimensional convolution parameters, yielding the following two channel weights:

$$A_l = \sigma(C_{1D}{}^{loc}(R(P_l(Y_{GR})))), \quad A_g = \sigma(C_{1D}{}^{glo}(P_g(P_l(Y_{GR})))) \tag{9}$$

MLCA expands the global weights to the local grid and fuses the local and global responses using $\alpha = 0.5$, after which adaptive average pooling resizes the result to the input spatial dimensions. This process allows the features extracted by multi-granularity convolution to emphasize discriminative local structures along the channel dimension while retaining semantic responses associated with the complete vehicle and surrounding objects:

$$A = R_{H,W}(\alpha\, A_l + (1\text{-}\alpha)A_g), \quad Y_{MLCA} = Y_{GR} \odot A, \quad \alpha = 0.5 \tag{10}$$

GRConv and MLCA act on the branch fusion ratio and the fused feature responses, respectively. At each location, GRConv assigns channel-shared branch weights to the two receptive fields. MLCA combines local-grid and global statistics to generate attention weights that vary across channels and grid locations. The mixing coefficient α in MLCA is fixed at 0.5, while the one-dimensional convolution parameters in the local and global branches are learned; therefore, even with a fixed mixing ratio, the final attention weights vary with the input features. When the two modules are connected sequentially, the network first combines features from different receptive fields and then refines the combined representation.

Bottleneck_GRM sequentially performs channel transformation, GRConv, and MLCA. When shortcut = True and the input and output channels are identical, residual addition is applied; otherwise, the transformed result is output directly. C3k2GRM follows the split, progressive stacking, and channel concatenation structure of C2f, selecting C3kGRM or Bottleneck_GRM as its internal unit according to the c3k switch. In C3kGRM, conventional C3 bottlenecks are replaced with Bottleneck_GRM. This structure is used at selected C3k2 positions in the YOLO11n backbone and P3 detection branch while preserving the overall feature hierarchy.

C3k2GRM retains the local-grid/global pooling and channel-weighting formulation of MLCA [12]. The structural change is the Bottleneck_GRM sequence embedded within C3k2: spatial gating first mixes the two fixed-dilation convolution branches, after which MLCA reweights their fused features. This separates receptive field selection from attention-based feature refinement. For an occluded chariot box or chariot occupant, surviving contour fragments require local detail, whereas the surrounding wheel, canopy, and vehicle structure provide contextual cues for recognition. Applying spatially varying branch weights before MLCA is intended to accommodate these coexisting requirements within the same feature map.

3.7. Task-Adapted ASF Multi-Scale Fusion Pathway

ASF-YOLO was originally developed for cell instance segmentation. The model aims to use spatial and scale information to improve the segmentation of cells of different sizes while maintaining computational efficiency [13]. It contains two key mechanisms: SSFF and TFE. SSFF aggregates features from different network levels to enhance scale-sequence representations, so as to provide richer multi-scale information when objects vary markedly in size; TFE sequentially applies convolution, Batch Normalization, and activation to three input features of different sizes, aligns their spatial dimensions through upsampling or downsampling, and then uses concatenation and related fusion

operations to construct a combined feature representation that contains both contextual information and local detail.

Han pictorial stone chariot-and-horse images contain complete scenes, vehicle bodies, and local components, requiring the integration of features at different resolutions. The TFE and SSFF operations are adopted from ASF-YOLO [13]; the changes concern their connections and channel matching within YOLO11n. Whereas ASF-YOLO integrates these paths using channel and position attention, the present implementation uses the Add module to feed the scale-sequence output into P3. The implementation comprises TFE (Zoom_cat), SSFF (ScalSeq), and Add. TFE (Zoom_cat) aligns and concatenates high-, intermediate-, and low-resolution features; SSFF (ScalSeq) aggregates P3, P4, and P5 information along an explicit scale dimension; and Add supplements the P3 detection branch with the fused result. This adaptation does not alter YOLO11n's three-scale detection outputs. The connection positions of the three structures in the proposed model are shown in Figure 4, and their feature processing procedures are described below.

$$Z_m = \mathrm{Cat}(P_{max}{}^m(X_l) + P_{avg}{}^m(X_l), X_m, U_m(X_s)) \tag{11}$$

In Equation (11), the three inputs denote a large-sized (high-resolution) feature map, an intermediate-sized feature map, and a small-sized (low-resolution) feature map, respectively. Adaptive max pooling and average pooling produce the intermediate target size, nearest neighbor upsampling enlarges the small feature map, and Cat denotes concatenation along the channel dimension. Max and average pooling align shallow high-resolution features to the intermediate scale while preserving salient responses and overall statistical information; nearest neighbor upsampling aligns deep low-resolution features to the intermediate scale. TFE (Zoom_cat) is used at two fusion nodes in the detection head, allowing shallow localization information, current-level representations, and deep semantic information to interact explicitly before entering the subsequent C3k2 or C3k2GRM module.

SSFF (ScalSeq) receives the P3, P4, and P5 features from the backbone and first uses 1 × 1 convolution to unify their channel counts. The configured 256 channels are scaled to 64 channels by the YOLO11n width factor of 0.25. P4 and P5 are then resized to the P3 resolution using nearest neighbor interpolation. Let R(i→3) denote the scale alignment operation from level i to the P3 spatial dimensions:

$$P_i^a = R_{i\to 3}(\mathrm{Conv}_{1\times 1}(P_i)), \quad i = 3,4,5 \tag{12}$$

The three aligned scales are stacked along a new scale dimension and then sequentially processed by a 1 × 1 × 1 three-dimensional convolution, Batch Normalization, leaky rectified linear unit (LReLU), and 3 × 1 × 1 max pooling. The 1 × 1 × 1 convolution performs channel transformation at each scale position without directly mixing adjacent scales, while 3 × 1 × 1 max pooling aggregates responses along the scale dimension to produce a single fused feature:

$$T = \mathrm{Stack}(P_3^a, P_4^a, P_5^a), \quad S = \mathrm{Squeeze}(\mathrm{Pool}_{3\times 1\times 1}(\mathrm{LReLU}(\mathrm{BN}(\mathrm{Conv}_{1\times 1\times 1}(T))))) \tag{13}$$

The fused feature S is added element-wise to the output of the P3 detection branch through Add, supplementing the small-scale detection feature with semantic information from deeper levels and cross-scale associations:

$$F_{P3}{}^{out} = F_{P3} + S \tag{14}$$

This pathway performs feature alignment, scale sequence aggregation, and feature addition through TFE (Zoom_cat), SSFF (ScalSeq), and Add. The adaptation involves

their connection positions in YOLO11n, the channel configuration, and the supplementation of the P3 branch, while preserving the P3, P4, and P5 three-scale detection outputs. The two TFE (Zoom_cat) modules are located at the P4 and P3 top-down fusion nodes. As a parallel pathway, SSFF (ScalSeq) maps the backbone P3, P4, and P5 features to 64 channels and the P3 spatial resolution, aggregates them across scales, and adds the result element-wise to the P3 branch feature. The P4 and P5 detection features continue to be generated by the bottom-up pathway and are input to the Detect layer together with the enhanced P3 feature. The three prediction scales are retained, with their input features generated by the modified fusion pathway. The emphasis on P3 addresses small or partly visible elements whose localization depends on higher-resolution features, while the P4 and P5 inputs supply broader scene context. Retaining all three prediction scales also accommodates the large differences in the apparent sizes of horses, chariot occupants, and vehicle components across full-view rubbings and cropped plates.

3.8. Loss Functions and Inference Output

The model retains the classification loss, bounding box regression loss, and Distribution Focal Loss (DFL) of the YOLO11 detection framework. The total training loss is expressed as follows:

$$L = \lambda_{box} L_{box} + \lambda_{cls} L_{cls} + \lambda_{dfl} L_{dfl} \quad (15)$$

Here, λbox = 7.5, λcls = 0.5, and λdfl = 1.5. The classification branch uses binary cross-entropy with logits. Bounding box regression uses Complete Intersection over Union (CIoU), which constrains geometric factors including overlap, center distance, and aspect ratio [43]. DFL models the four boundary distances—left, top, right, and bottom—as discrete distributions [44], following the implementation of the YOLO11 framework [42]. The three losses constrain category discrimination, geometric matching of bounding boxes, and boundary distance estimation, respectively, and their formulations remain unchanged. During inference, candidate boxes at the P3, P4, and P5 scales undergo confidence filtering and non-maximum suppression, after which the category, bounding box, and confidence score are output.

4. Experimental Design and Results

4.1. Experimental Setup

The experiments were conducted in Ubuntu 20.04 using Python 3.8, PyTorch 2.0.0, the Ultralytics detection framework (v8.3.0), and CUDA 11.8. Training used one NVIDIA GeForce RTX 3090 GPU with 24 GB of memory (NVIDIA Corporation, Santa Clara, CA, USA). The system also included an Intel Xeon Gold 6330 CPU (Intel Corporation, Santa Clara, CA, USA). The proposed model used the stochastic gradient descent (SGD) optimizer and was trained for 200 epochs with a batch size of 32, an initial learning rate of 0.01, a momentum of 0.937, a weight decay of 0.0005, and automatic mixed precision enabled. The effective input size was 672 × 672 pixels after the framework applied network stride adjustment. This effective input size was used consistently for the baseline, all ablation models, and all comparison models, and the reported GFLOPs were calculated under the same setting. Mosaic augmentation was disabled during the final 10 epochs so that the model could fit the original image distribution more stably during the late stage of training. The baseline, all ablation models, and the comparison models used the same data partitions, input settings, number of training epochs, data augmentation strategies, random seed, and hardware environment to ensure comparability across experiments.

4.2. Evaluation Metrics

Comparative research on object detection metrics has shown that the intersection over union (IoU) threshold, the method used to integrate the precision–recall (PR) curve, and implementation details all affect the interpretation of results [45]. This study evaluates detection performance using precision, recall, average precision at an IoU threshold of 0.5 (AP@0.5), mean average precision at 0.5 (mAP@0.5), and mean average precision across IoU thresholds from 0.5 to 0.95 (mAP@0.5:0.95) and reports the parameter count and GFLOPs to quantify model complexity. First, the intersection over union between a predicted box B_p and a ground-truth box B_g is defined as follows:

$$\mathrm{IoU}(B_p, B_g) = |B_p \cap B_g| / |B_p \cup B_g| \tag{16}$$

At a given IoU threshold and confidence threshold, precision and recall are defined as follows:

$$\mathrm{Precision} = TP/(TP + FP), \quad \mathrm{Recall} = TP/(TP + FN) \tag{17}$$

Here, TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively. The confidence value corresponding to the maximum of the smoothed F1 curve on the validation set is used as the threshold, and this threshold is fixed during test set evaluation to calculate precision and recall. For category c and IoU threshold τ, the implementation first constructs a monotonic precision envelope and then interpolates it at 101 uniformly spaced recall points. Let r_j and p_j denote the j-th recall sample and the corresponding interpolated envelope precision, respectively; AP is then calculated using the trapezoidal rule:

$$AP_c(\tau) = \sum_{j=0}^{99} [(p_j + p_{j+1})/2](r_{j+1} - r_j), \quad r_j = j/100 \tag{18}$$

For C = 5 categories, the mean average precision at a fixed IoU threshold τ is defined as follows:

$$mAP@\tau = C^{-1} \sum_{c=1}^{C} AP_c(\tau) \tag{19}$$

Common Objects in Context (COCO)-style mAP@0.5:0.95 [45,46] further averages over 10 IoU thresholds from 0.50 to 0.95 in increments of 0.05:

$$mAP@0.5{:}0.95 = (10C)^{-1} \sum_{k=0}^{9} \sum_{c=1}^{C} AP_c(0.50 + 0.05k) \tag{20}$$

Thus, mAP@0.5 is the mean AP across the five categories at an IoU threshold of 0.5, whereas mAP@0.5:0.95 places greater emphasis on bounding box localization quality under different overlap requirements. In addition to overall mAP, this study reports AP@0.5 for horse, wheel, chariot box, canopy, and chariot occupant to characterize detection performance across object categories.

4.3. Training Convergence

The proposed model exhibited a stable optimization process over 200 epochs. The training set box loss, classification loss, and DFL loss generally decreased throughout training, while their validation set counterparts decreased rapidly at the beginning and then gradually stabilized. Precision, recall, mAP@0.5, and mAP@0.5:0.95 increased as training progressed and remained at relatively high levels during the later stages. Figure 7 presents the complete trajectories of the losses and core evaluation metrics.

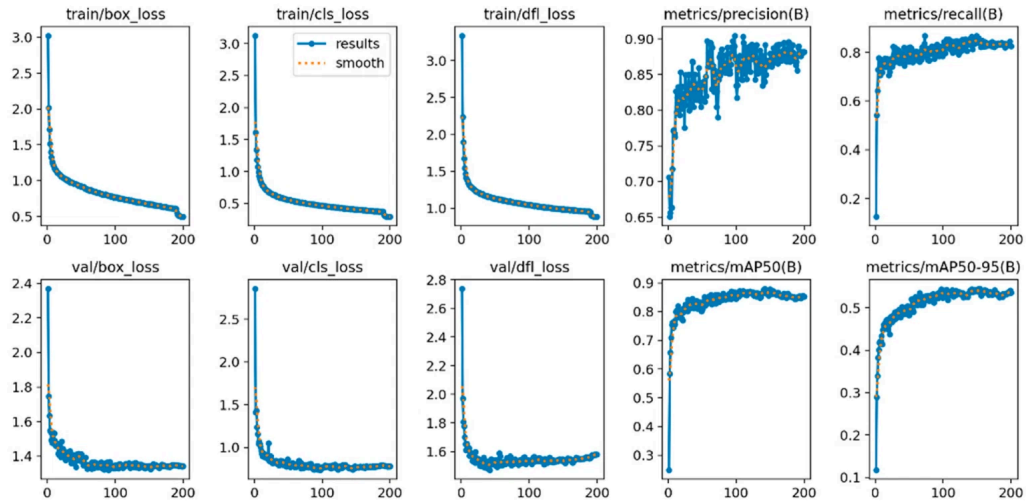


Figure 7. Training losses and validation metrics of the proposed model. The top row shows, from left to right, training box regression loss, classification loss, DFL loss, precision, and recall. The bottom row shows validation box regression loss, classification loss, DFL loss, mAP@0.5, and mAP@0.5:0.95.

4.4. Ablation Experiments

To analyze the relationship between the three architectural modifications and changes in detection performance, eight configurations were evaluated under identical experimental settings: the YOLO11n baseline; RubbingDown, C3k2GRM, and the task-adapted ASF pathway introduced independently; all three pairwise combinations; and the complete model containing all three components. Table 2 reports precision, recall, both mAP metrics, and model complexity for each configuration.

Table 2. Component-level and combinatorial ablation results for the improved YOLO11n model.

Model	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	Parameters/M	GFLOPs
YOLO11n	86.2	76.6	82.5	51.7	2.58	6.3
+RubbingDown	85.2	79.8	84.5	53.3	2.51	6.7
+C3k2GRM	85.4	79.9	83.7	54.0	2.56	6.2
+ASF	86.0	76.9	84.1	52.9	2.67	7.1
+RubbingDown + C3k2GRM	89.5	79.9	86.0	54.7	2.49	6.5
+RubbingDown + ASF	88.8	81.6	86.1	54.0	2.60	7.4
+C3k2GRM + ASF	85.8	79.4	85.7	52.4	2.65	6.9
+RubbingDown + C3k2GRM + ASF	84.0	86.5	88.0	54.5	2.58	7.3

After RubbingDown was added, mAP@0.5 increased from 82.5% to 84.5%, mAP@0.5:0.95 increased from 51.7% to 53.3%, and recall increased by 3.2 percentage points. These changes are consistent with the design objective of preserving pixel organi-

zation, directional structures, and edge responses during shallow downsampling. They also indicate that this composite module yields a performance gain with only a small change in complexity: the parameter count decreases from 2.58 M to 2.51 M, while computation increases by only 0.4 GFLOPs. When introduced independently, C3k2GRM produced an mAP@0.5 of 83.7% and an mAP@0.5:0.95 of 54.0%, while the task-adapted ASF pathway produced corresponding values of 84.1% and 52.9%. Thus, all three individual components improved both mAP metrics relative to the baseline under the same evaluation conditions.

Adding C3k2GRM on top of RubbingDown further increased mAP@0.5 to 86.0% and mAP@0.5:0.95 to 54.7%, while precision reached 89.5%. The results at this stage are consistent with the design objective of using dual-receptive field branches, spatial gating, and MLCA to enhance representations for object discrimination and localization. Among the additional pairwise configurations, RubbingDown + ASF achieved a recall of 81.6% and an mAP@0.5 of 86.1%, while C3k2GRM + ASF also exceeded the baseline in recall and both mAP metrics. After the task-adapted ASF multi-scale fusion pathway was further added, recall increased from 79.9% to 86.5%, and mAP@0.5 increased from 86.0% to 88.0%, whereas precision decreased from 89.5% to 84.0% and mAP@0.5:0.95 decreased slightly from 54.7% to 54.5%. These results indicate that the ASF pathway shifts the performance balance toward higher recall and mAP@0.5, consistent with its role in supplementing cross-scale information and improving object coverage. Relative to the YOLO11n baseline, the complete model improved recall, mAP@0.5, and mAP@0.5:0.95 by 9.9, 5.5, and 2.8 percentage points, respectively, while retaining a parameter count of 2.58 M.

To further identify the contribution of the internal components of RubbingDown, branch-level ablations were conducted by removing SPD, StripConv, EdgePrior, spatial attention (SA), and channel attention (CA) individually while keeping the remaining training settings unchanged. The complete RubbingDown configuration and the YOLO11n baseline are included as references in Table 3.

Table 3. Branch-level ablation results for the internal components of RubbingDown.

Model	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	Parameters/M	GFLOPs
YOLO11n	86.2	76.6	82.5	51.7	2.58	6.3
RubbingDown	85.2	79.8	84.5	53.3	2.51	6.7
RubbingDown w/o SPD	77.3	80.7	83.5	53.0	2.56	6.9
RubbingDown w/o StripConv	84.2	81.9	84.2	53.2	2.54	6.4
RubbingDown w/o EdgePrior	82.0	78.5	82.5	50.6	2.54	6.5
RubbingDown w/o SA	81.1	79.5	84.0	51.5	2.51	6.6
RubbingDown w/o CA	85.7	76.4	84.1	51.6	2.51	6.7

The complete RubbingDown configuration achieved the highest mAP@0.5 (84.5%) and mAP@0.5:0.95 (53.3%) among the branch-level variants. Removing EdgePrior produced the largest decrease, reducing mAP@0.5 and mAP@0.5:0.95 by 2.0 and 2.7 percent-

age points, respectively. Removing spatial attention or channel attention reduced mAP@0.5:0.95 by 1.8 and 1.7 percentage points, respectively. Excluding SPD reduced precision from 85.2% to 77.3% and mAP@0.5 by 1.0 percentage point, whereas removing StripConv produced smaller changes of 0.3 and 0.1 percentage points in the two mAP metrics. These results indicate that EdgePrior and the two attention branches provide the largest measurable contributions, while SPD and StripConv provide complementary information within the complete module.

4.5. Comparison with Mainstream Lightweight Object Detectors

The proposed model was benchmarked against five mainstream lightweight detectors—YOLOv5n, YOLOv8n, YOLOv10n, YOLO12n, and YOLO26n—with YOLO11n retained as the direct baseline. All models used the same data partitions, input settings, number of training epochs, data augmentation strategies, random seed, and hardware environment and were evaluated using a unified procedure. Table 4 reports precision, recall, mAP@0.5, mAP@0.5:0.95, parameter count, and GFLOPs.

Table 4. Performance comparison of mainstream lightweight detection models.

Model	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	Parameters/M	GFLOPs
YOLOv5n	83.2	79.1	81.9	50.5	2.50	7.1
YOLOv8n	81.4	80.4	82.4	51.3	3.01	8.1
YOLOv10n	84.1	76.3	81.1	48.3	2.27	6.5
YOLO12n	85.1	75.2	81.6	49.8	2.56	6.3
YOLO26n	82.6	77.7	81.3	51.4	2.38	5.2
YOLO11n	86.2	76.6	82.5	51.7	2.58	6.3
Proposed model	84.0	86.5	88.0	54.5	2.58	7.3

As shown in Table 4, the proposed model achieves a recall of 86.5%, an mAP@0.5 of 88.0%, and an mAP@0.5:0.95 of 54.5%, all of which are the highest among the compared models. Relative to the best results among the five mainstream lightweight models, these three metrics improve by 6.1, 5.6, and 3.1 percentage points, respectively. Compared with the direct YOLO11n baseline, mAP@0.5 and mAP@0.5:0.95 increase by 5.5 and 2.8 percentage points, respectively. The proposed model contains 2.58 M parameters, which is comparable to the other lightweight models. These results demonstrate that the proposed model expands object coverage and achieves higher overall detection accuracy without increasing the parameter count relative to YOLO11n. These accuracy gains are accompanied by an increase in GFLOPs from 6.3 to 7.3.

4.6. Category-Wise AP for the Five Object Classes

Category-wise AP reveals how individual objects contribute to the overall mAP. As shown in Table 5, the model incorporating all three modifications outperforms the baseline for wheel, chariot box, canopy, and chariot occupant, while remaining broadly unchanged for horse. Specifically, AP increases from 0.689 to 0.805 for chariot box, from 0.828 to 0.910 for canopy, and from 0.822 to 0.874 for chariot occupant, demonstrating the model's enhanced capability for chariot boxes, canopies, and chariot occupants.

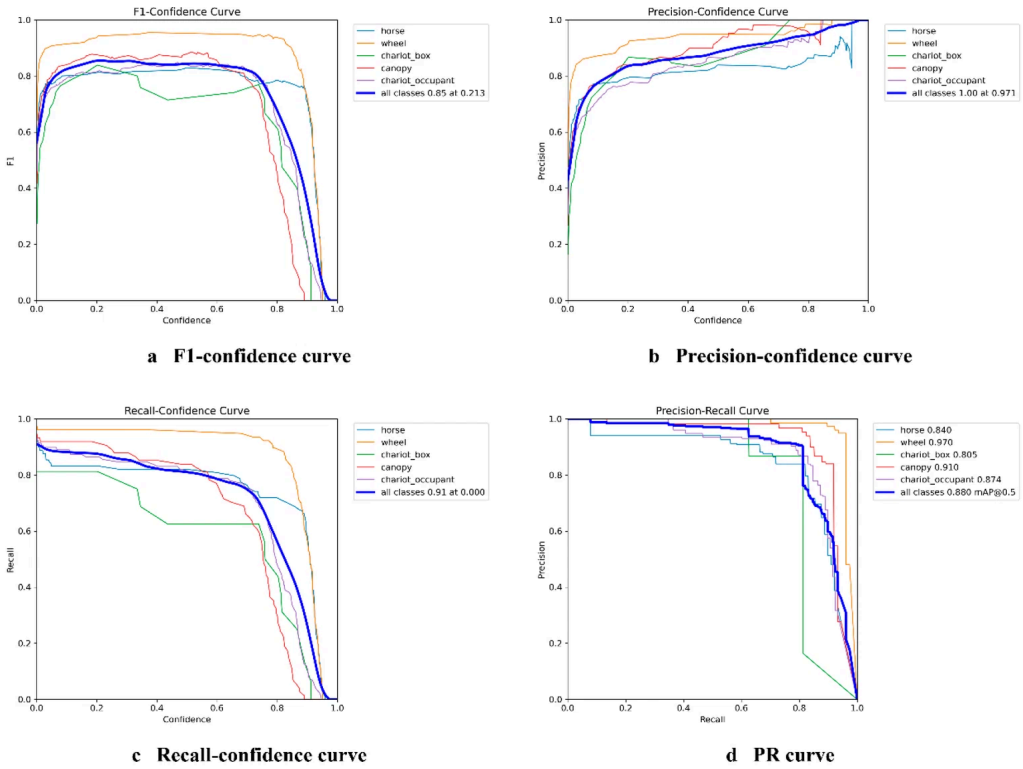
Table 5. AP@0.5 for the five categories across the cumulative ablation models.

Model	Horse	Wheel	Chariot Box	Canopy	Chariot Occupant	mAP@0.5
YOLO11n	0.841	0.947	0.689	0.828	0.822	0.825
+RubbingDown	0.843	0.954	0.727	0.903	0.799	0.845
+RubbingDown + C3k2GRM	0.871	0.970	0.759	0.864	0.835	0.860
+RubbingDown + C3k2GRM + Task-Adapted ASF Pathway	0.840	0.970	0.805	0.910	0.874	0.880

Wheel achieves a high AP of 0.947 with the baseline model, which further increases to 0.970 after the three modifications are incorporated. AP for horse changes from 0.841 to 0.840 and thus remains essentially stable. RubbingDown increases canopy AP by 7.5 percentage points; at the C3k2GRM stage, horse, wheel, chariot box, and chariot occupant reach comparatively high values for that stage; and the task-adapted ASF pathway further improves chariot box, canopy, and chariot occupant. These category-level patterns may reflect differences in object appearance and context, as discussed in Section 5.2 .

4.7. Confidence Curves and PR Curves

Curves of F1, precision, and recall against the confidence threshold characterize detection results under different thresholds, while PR curves provide an integrated view of each category's detection capability as the threshold changes. Figure 8 shows that recall for all five categories gradually decreases as the threshold increases, and the overall F1 curve attains a favorable balance at an intermediate confidence level. The wheel PR curve encloses a relatively large area, whereas the chariot box curve exhibits greater variation across thresholds. In conjunction with Table 2, these results show that adding the task-adapted ASF multi-scale fusion pathway increases recall from 79.9% to 86.5% under the current experimental settings.



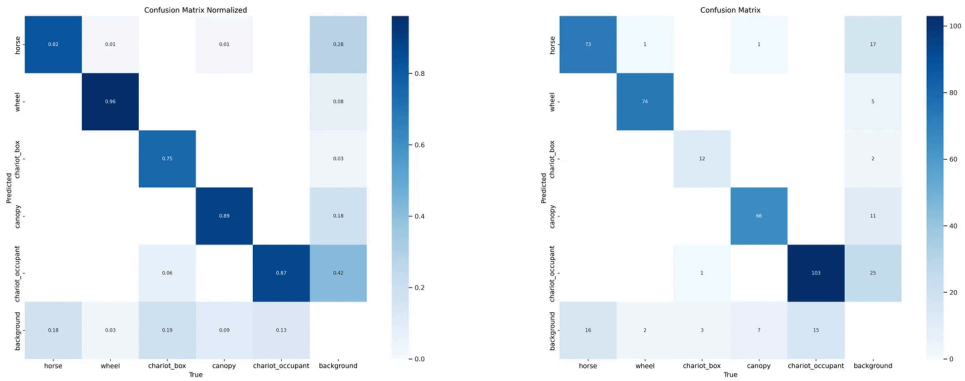
a F1-confidence curve **b Precision-confidence curve**

c Recall-confidence curve **d PR curve**

Figure 8. Threshold curves and PR curves of the proposed model. (**a**) F1 confidence curve; (**b**) precision confidence curve; (**c**) recall confidence curve; (**d**) PR curve.

4.8. Confusion Matrix Analysis

Figure 9 presents the normalized and raw-count confusion matrices used to evaluate class-wise recognition and misclassification patterns. The dominant diagonal entries indicate that the model effectively captures the distinguishing visual characteristics of the five object categories. Wheels have relatively distinct category boundaries because of their stable geometric contours, whereas chariot boxes, canopies, and chariot occupants depend more strongly on neighboring objects and the overall vehicle structure. The background row and column reflect false positives and false negatives under complex textures, providing evidence for further interpretation of threshold selection and scene difficulty.



a Normalized confusion matrix **b Raw-count confusion matrix**

Figure 9. Confusion matrices of the proposed model. (**a**) Confusion matrix normalized by ground-truth category; (**b**) raw count confusion matrix.

4.9. Qualitative Detection Results

To qualitatively compare the two models, Figure 10 presents five groups of test set samples, including dense rubbings, collection rubbings, multi-register rubbings, archival scans, and multi-register narrative rubbings, alongside predictions from the YOLO11n baseline and the proposed model. Compared with the baseline model, the proposed model covers more corresponding objects in these samples and recovers some previously missed horses, wheels, and other relevant objects. This result is consistent with the improvement in overall recall and selected category-wise AP values and intuitively shows how the three-stage feature enhancement pathway improves detection in complex chariot-and-horse scenes.

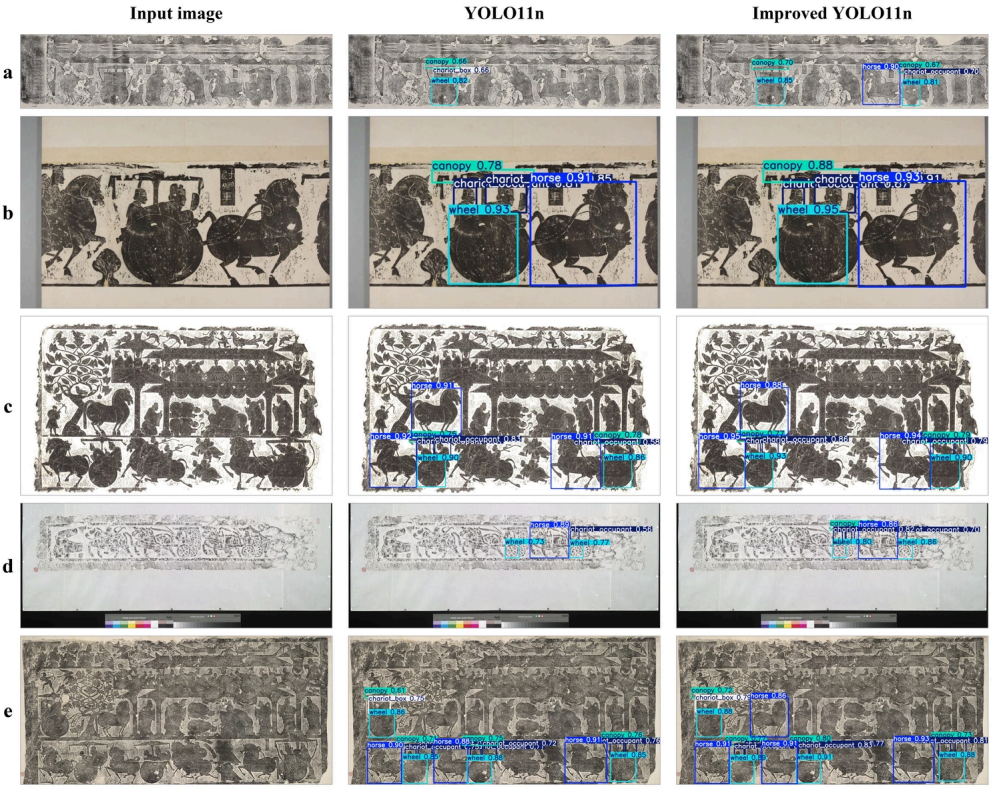


Figure 10. Qualitative comparison of the detection results on representative samples. The left column is the original input image, the middle column is the prediction result of the YOLO11n baseline, and the right column is the prediction result of the proposed model. Panels (**a**–**e**) show dense rubbings, collection rubbings, multi-register rubbings, archival scans, and multi-register narrative rubbings, respectively. Bounding-box colors distinguish the predicted object categories, while the accompanying labels indicate the category names and confidence scores.

To further examine model behavior under complex acquisition conditions, the supplementary qualitative results are presented in Figure 11. Across the illustrated samples, the proposed model detects more annotated horses, wheels, canopies, and chariot occupants than YOLO11n and produces fewer false positives in non-target regions such as mounting edges, black backings, and color charts. These qualitative observations are consistent with the higher overall recall reported in Table 2; however, because they are based on selected examples, they should not be interpreted as evidence of general cross-domain robustness. Quantitative performance from the independent cross-source evaluation is reported separately in Section 4.10..

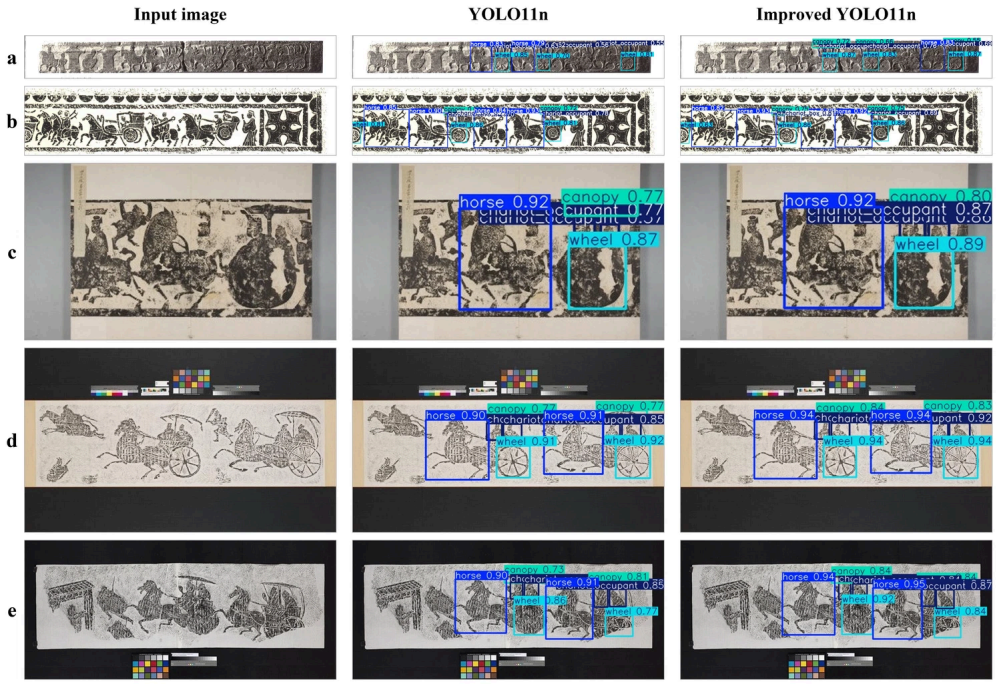


Figure 11. Supplementary qualitative detection results under complex acquisition conditions. The left column shows the original input images, the middle column shows predictions from the YOLO11n baseline, and the right column shows predictions from the proposed model. Panels (**a**–**e**) present a small-scale continuous chariot-and-horse scene, a dense multi-chariot scene, a clear chariot-and-horse rubbing with a mounting background, a multi-chariot archival image, and a collection scan containing a color chart, respectively. The figure illustrates detection behavior under variations in aspect ratio, target density, non-target blank areas, and acquisition media. Bounding-box colors distinguish the predicted object categories, while the accompanying labels indicate the category names and confidence scores.

4.10. Data Augmentation Ablation and Cross-Source Evaluation

To assess the vertical flip augmentations, the proposed model was trained under identical settings using either the full augmented training set (3017 images) or a no_vertical set obtained by removing 862 vertically flipped and combined horizontal–vertical-flipped samples (2155 images). The validation and test sets were unchanged. Table 6 reports the results of an independent controlled run conducted for the data augmentation ablation.

Table 6. Ablation results for vertical flip-related data augmentation in the proposed model.

Model/Configuration	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	Parameters/M	GFLOPs
Proposed model (no_vertical)	84.3	75.4	81.1	51.0	2.58	7.3
Proposed model (full)	84.8	79.9	83.7	51.1	2.58	7.3

With the full augmentation strategy, recall increased from 75.4% to 79.9%, and mAP@0.5 increased from 81.1% to 83.7% relative to no_vertical; precision increased from 84.3% to 84.8%, while mAP@0.5:0.95 changed from 51.0% to 51.1%. The full augmentation strategy, therefore, improved recall and mAP@0.5 under the controlled augmenta-

tion comparison.

For cross-source evaluation, 93 images from an independent data source were kept separate from the training, validation, and internal test sets and used exclusively as an external test set. The held-out source comprised original images from the digital archive of the Fu Ssu-nien Library of the Institute of History and Philology. YOLO11n and the proposed model were trained under identical settings and evaluated on this external test set. Table 7 reports the cross-source results.

Table 7. Cross-source evaluation results on the independent external test set.

Model/Configuration	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	Parameters/M	GFLOPs
YOLO11n	80.6	63.7	72.6	40.1	2.58	6.3
Proposed model	81.3	79.8	84.5	50.7	2.58	7.3

On the independent external test set, the proposed model achieved a precision of 81.3%, a recall of 79.8%, an mAP@0.5 of 84.5%, and an mAP@0.5:0.95 of 50.7%, compared with 80.6%, 63.7%, 72.6%, and 40.1% for YOLO11n. The corresponding gains in recall, mAP@0.5, and mAP@0.5:0.95 were 16.1, 11.9, and 10.6 percentage points, respectively, showing that the proposed model retained its performance advantage under cross-source evaluation.

5. Discussion

5.1. Task-Driven Feature Enhancement and Fusion

The central finding of this study is that the overall detection performance of YOLO11n changes after its downsampling, feature extraction, and multi-scale fusion components are modified in stages to address fine-line structures, background texture interference, and scale variation in Han pictorial stone chariot-and-horse images. The three modifications perform distinct architectural functions: RubbingDown targets information preservation during spatial compression, C3k2GRM coordinates local and contextual representations, and the task-adapted ASF pathway supplements the P3 branch with cross-scale information. The component-level results show that each modification independently increases both mAP metrics, while the pairwise configurations exhibit complementary but non-uniform effects across precision, recall, and localization accuracy. In the cumulative ablation sequence, the ASF stage shifted the performance balance: recall increased from 79.9% to 86.5% and mAP@0.5 from 86.0% to 88.0%, whereas precision decreased from 89.5% to 84.0% and mAP@0.5:0.95 from 54.7% to 54.5%. Thus, the final-stage modification favored object coverage under the present evaluation conditions, with gains and declines across metrics.

With all three components incorporated, the model achieves an mAP@0.5 of 88.0% and a recall of 86.5%, representing gains of 5.5 and 9.9 percentage points, respectively, over the YOLO11n baseline. The expanded component-level and combinatorial ablations clarify the individual and paired contributions of RubbingDown, C3k2GRM, and ASF. Among all evaluated configurations, the complete model achieves the highest recall and mAP@0.5, while its mAP@0.5:0.95 remains close to the best observed value. The parameter count remains 2.58 M, identical to that of the baseline, while GFLOPs increase from 6.3 to 7.3, corresponding to an increase of approximately 15.9%. The observed metric gains are, therefore, accompanied by a moderate increase in computational complexity without additional model parameters.

Across the five additional light-weight detectors, the proposed model achieved the highest recall, mAP@0.5, and mAP@0.5:0.95 under the common experimental protocol, exceeding the best comparison values by 6.1, 5.6, and 3.1 percentage points, respectively.

These results demonstrate the effectiveness of the proposed model for detecting chariot-and-horse elements in Han pictorial stone images under the evaluated experimental setting.

5.2. Detection Performance and Visual Characteristics Across Categories

The category-wise AP values show that the magnitude of performance changes varies across the five object classes. Wheels have relatively stable closed contours, and the baseline already achieves a high AP, leaving limited room for improvement. Chariot boxes are often occluded by human figures, wheels, and canopies and, therefore, retain fewer visible boundaries; the arched tops and slender supports of canopies depend strongly on directional edges; and the recognition of chariot occupants is affected by scale variation, pose differences, and overlap between human figures and vehicles. On the current dataset, the observed AP gains are larger for chariot boxes, canopies, and chariot occupants than for wheels, indicating category-dependent performance changes.

The category-wise results are broadly consistent with this interpretation. After RubbingDown is introduced, canopy AP increases from 0.828 to 0.903. Following the addition of C3k2GRM, horse and wheel AP reach 0.871 and 0.970, respectively. With all three modifications, AP for chariot box, canopy, and chariot occupant reaches 0.805, 0.910, and 0.874, respectively. These changes suggest that preserving directional structures and edge information may facilitate canopy contour recognition; coordinating local and contextual features may improve representations of horses and wheels; and supplementing cross-scale information may be beneficial for categories characterized by incomplete structures, pronounced scale variation, and complex compositional relationships. Overall, the complete model raises recall and mAP@0.5 to 86.5% and 88.0%, respectively, demonstrating improved object coverage and overall detection performance.

Research on object detection for cultural heritage has shown that adapting detection categories and feature processing architectures to specific material forms and recognition targets is a common means of improving detector adaptability to complex heritage images [8,39,47]. The findings of this study are consistent with this view. However, chariot-and-horse images contain five interrelated object categories with marked scale differences, requiring the model not only to recognize individual contours but also to exploit local and global relationships among vehicle components and human figures. Component-level and combinatorial ablation, category-wise AP, and qualitative results jointly demonstrate that the modifications designed to preserve line-engraved structures, coordinate context, and supplement cross-scale information improve detection of chariot boxes, canopies, chariot occupants, and related categories on the current dataset. These findings support the task-specific adaptation of YOLO11n for this dataset. The proposed explanations of feature behavior remain interpretations requiring direct mechanistic verification.

5.3. Supporting Value of Chariot-and-Horse Element Detection for Archaeological Image Research

The preceding category-wise and qualitative results show that the proposed model can separately locate horses, wheels, chariot boxes, canopies, and chariot occupants within the same chariot-and-horse scene while retaining the spatial position of each element in the original image. Compared with whole-image thematic classification, category and bounding box information refine the unit of analysis from the entire image to specific internal elements, providing a comparable descriptive basis for element counts, co-occurrence relationships, and spatial configurations across images. In the study of Han pictorial stone chariot-and-horse images, such information can provide cues for

sample-based comparisons of vehicle composition, figure arrangement, and pictorial organization. At the same time, research in digital archaeology emphasizes that digital results must be interpreted in relation to the circumstances of data creation, recording, and processing and within their specific interpretive contexts [48].

Model outputs provide category and spatial location information for visible elements, offering verifiable, structured cues for further research on vehicle types, figure identities, ritual ranks, and historical meanings. Linking the detection results with source image provenance, catalogue records, excavation information, and manual verification records can further extend their use in image retrieval, quantitative analysis, and cross-sample comparison. The category and bounding box outputs extend Han pictorial stone analysis from whole-image thematic recognition to the classification and localization of internal pictorial elements. They also provide structured information that can support future workflows in image retrieval, quantitative analysis, cross-sample comparison, and record linkage.

5.4. Scope and Future Directions

Future research can extend this work to three complementary directions. First, broader cross-source evaluation across additional collections, regions, historical periods, and media types would provide further evidence of model robustness under domain shift [49]. Second, controlled comparisons using fixed data partitions and training settings would enable the assessment of new strategies for fine-detail preservation, contextual modeling, multi-scale feature fusion, and computational efficiency, supported by component-level ablations. YOLO26, as documented in [50], offers a current reference through its end-to-end inference, lightweight detection head, and updated training strategies. YOLO26-based architectural adaptations can be evaluated under the same protocol, while subsequent models, such as YOLO27, can be incorporated once their architectures, weights, and evaluation protocols are publicly documented. Third, integrating uncertainty estimation and linking detection outputs with source records and expert verification would support the development and evaluation of downstream research workflows.

6. Conclusions

This study developed a five-class dataset centered on Han pictorial stones for the detection of chariot-and-horse elements, comprising 539 original images and 3736 annotated object instances. An improved YOLO11n model was subsequently developed by incorporating RubbingDown, C3k2GRM, and a task-adapted ASF multi-scale fusion pathway to enhance fine-detail preservation during downsampling, contextual feature representation, and cross-scale feature fusion, respectively.

The experimental results show that the proposed model achieves an mAP@0.5 of 88.0%, an mAP@0.5:0.95 of 54.5%, and a recall of 86.5%, representing improvements of 5.5, 2.8, and 9.9 percentage points, respectively, over the YOLO11n baseline, with the parameter count remaining unchanged at 2.58 M. Performance gains are also observed for individual categories, particularly chariot box, canopy, and chariot occupant. In comparison with five mainstream lightweight YOLO models, the proposed model achieves higher recall and mAP values. GFLOPs increase from 6.3 to 7.3, indicating that the improvements in detection performance are accompanied by a moderate increase in computational complexity without an increase in parameter count. In the independent cross-source evaluation, the proposed model also increased recall, mAP@0.5, and mAP@0.5:0.95 by 16.1, 11.9, and 10.6 percentage points, respectively, over YOLO11n.

The resulting category and spatial location information can support future work-

flows in image retrieval, quantitative analysis, co-occurrence analysis, and manual verification. Future studies may further evaluate these applications across data sources and link detection outputs with relevant source records. The role of this model is limited to identifying and localizing visually observable elements; the interpretation of vehicle types, figure identities, ritual ranks, and symbolic meanings requires in-depth analysis by domain experts based on archaeological contexts, historical texts, and iconographic evidence.

Author Contributions: Conceptualization, Y.H. and W.H.; methodology, Y.H.; validation, Y.H. and H.C.; formal analysis, Y.H.; investigation, Y.H. and H.C.; data curation, Y.H.; writing—original draft preparation, Y.H.; writing—review and editing, W.H. and Q.W.; visualization, Y.H. and H.C.; supervision, W.H. and Q.W.; funding acquisition, W.H. and Q.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research and its publication were funded by the Philosophy and Social Science Major Project of Jiangsu Higher Education Institutions (2022SJZD118)—Research on Ancient Chinese Vehicle Design, and the General Art Project of the National Social Science Fund of China (23BG126)—Shaping the Chinese Product Design Style in the Context of "Going Global".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

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

Abbreviations

The following abbreviations are used in this manuscript:

YOLO	**You Only Look Once**
SPD	Space to Depth
MLCA	Mixed Local Channel Attention
ASF	Attentional Scale Sequence Fusion
TFE	Triple Feature Encoder
SSFF	Scale-Sequence Feature Fusion
HSV	Hue, Saturation, and Value
SGD	Stochastic Gradient Descent
AP	Average Precision
mAP	Mean Average Precision
IoU	Intersection over Union
GFLOPs	Giga Floating-Point Operations

References

1. Xin, L. A study of chariot procession images in Han pictorial art. *Southeast Cult.* **1999**, *1*, 47–63. (In Chinese) [Google Scholar]
2. Wu, M.M. The imperial imagination in depictions of journeys by horse and carriage in Han Dynasty stone reliefs. *Chin. Stud.* **2010**, *28*, 1–39. [Google Scholar] [CrossRef]
3. Shi, J. Rolling between burial and shrine: A tale of two chariot processions at Chulan Tomb 2 in Eastern Han China (171 C.E.). *J. Am. Orient. Soc.* **2015**, *135*, 433–452. [Google Scholar] [CrossRef]
4. Liu, Q. The Han-dynasty cosmological spatial view reflected in procession images on pictorial stones from southern Shandong. *Lit. Hist. Philos.* **2018**, *3*, 69–77. (In Chinese) [Google Scholar]

5. Fiorucci, M.; Khoroshiltseva, M.; Pontil, M.; Traviglia, A.; Del Bue, A.; James, S. Machine learning for cultural heritage: A survey. *Pattern Recognit. Lett.* **2020**, *133*, 102–108. [Google Scholar] [CrossRef]
6. Mishra, M.; Lourenço, P.B. Artificial intelligence-assisted visual inspection for cultural heritage: State-of-the-art review. *J. Cult. Herit.* **2024**, *66*, 536–550. [Google Scholar] [CrossRef]
7. Colace, F.; Gaeta, R.; Lorusso, A.; Pellegrino, M.; Santaniello, D. New AI challenges for cultural heritage protection: A general overview. *J. Cult. Herit.* **2025**, *75*, 168–193. [Google Scholar] [CrossRef]
8. Zhang, J.; Zhang, Y.; Liu, J.; Lan, Y.; Zhang, T. Human figure detection in Han portrait stone images via enhanced YOLO-v5. *Herit. Sci.* **2024**, *12*, 119. [Google Scholar] [CrossRef]
9. Guo, L.; Ju, F. Object detection of Northern Chinese rock art images using YOLOv8 with omni-dimensional dynamic convolution. *Appl. Sci.* **2026**, *16*, 5522. [Google Scholar] [CrossRef]
10. Jiang, K.; Qian, W.H.; Liu, P. Image-text retrieval for Han portrait stones through the synergy of semantic enhancement and text reorganization. *J. Image Graph.* **2026**, *31*, 3030–3044. (In Chinese) [Google Scholar] [CrossRef]
11. Sunkara, R.; Luo, T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. In *Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2022*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2023; Volume 13715, pp. 443–459. [Google Scholar] [CrossRef]
12. Wan, D.; Lu, R.; Shen, S.; Xu, T.; Lang, X.; Ren, Z. Mixed local channel attention for object detection. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106442. [Google Scholar] [CrossRef]
13. Kang, M.; Ting, C.M.; Ting, F.F.; Phan, R.C.W. ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell instance segmentation. *Image Vis. Comput.* **2024**, *147*, 105057. [Google Scholar] [CrossRef]
14. Li, Q.; Zou, Q.; Ma, D.; Wang, Q.; Wang, S. Dating ancient paintings of Mogao Grottoes using deeply learnt visual codes. *Sci. China Inf. Sci.* **2018**, *61*, 092105. [Google Scholar] [CrossRef]
15. Jalandoni, A.; Zhang, Y.; 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. [Google Scholar] [CrossRef]
16. Sun, S.; Li, T.; Li, Q. Automated detection and topic mining of ancient murals across different styles. *npj Herit. Sci.* **2026**, *14*, 112. [Google Scholar] [CrossRef]
17. Horn, C.; Ivarsson, O.; Lindhé, C.; Potter, R.; Green, A.; Ling, J. Artificial intelligence, 3D documentation, and rock art: Approaching and reflecting on the automation of identification and classification of rock art images. *J. Archaeol. Method Theory* **2022**, *29*, 188–213. [Google Scholar] [CrossRef]
18. Tsigkas, G.; Sfikas, G.; Pasialis, A.; Vlachopoulos, A.; Nikou, C. Markerless detection of ancient rock carvings in the wild: Rock art in Vathy, Astypalaia. *Pattern Recognit. Lett.* **2020**, *135*, 337–345. [Google Scholar] [CrossRef]
19. Ji, N.; Ju, F.; Wang, Q. An application study on digital image classification and recognition of Yunnan Jiama based on a YOLO-GAM deep learning framework. *Appl. Sci.* **2026**, *16*, 1551. [Google Scholar] [CrossRef]
20. Li, Y.; Wang, L.; Zhang, X.; Dong, S.; Zhu, X. YOLOv10-TWD: An improved YOLOv10n for Terracotta Warrior recognition. *Appl. Sci.* **2026**, *16*, 2616. [Google Scholar] [CrossRef]
21. Kim, Y.; Im, C.; Mandl, T. Object detection in historical images: Transfer learning and pseudo labelling. *J. Comput. Cult. Herit.* **2024**, *17*, 1–15. [Google Scholar] [CrossRef]
22. Li, B.; Ding, Z.; Zhang, Y.; Zhang, H.; Yang, J.; Fang, J.; Jin, T.; Liu, Y.; Song, Z.; Ji, R. An open benchmark for oracle bone rubbing image retrieval. *npj Herit. Sci.* **2025**, *13*, 292. [Google Scholar] [CrossRef]
23. Zhou, R.; Qiu, P.; Zhang, Q.; Li, C.; Yang, X. Ladder-side mixture of experts adapters for bronze inscription recognition. *npj Herit. Sci.* **2026**, *14*, 30. [Google Scholar] [CrossRef]
24. Duque Domingo, J.; Gómez-García-Bermejo, J.; Zalama, E. Egyptian shabtis identification by means of deep neural networks and semantic integration with Europeana. *Appl. Sci.* **2020**, *10*, 6408. [Google Scholar] [CrossRef]
25. Li, B.; Yang, J.; Liang, Y.; Hu, X.; Ding, Z.; Peng, X.; Han, S.; Qin, P.; Luo, D.; Jin, T.; et al. OBIMD: A multi-modal dataset for contextual interpretation of oracle bone inscriptions. *Sci. Data* **2026**, *13*, 681. [Google Scholar] [CrossRef]
26. Li, Y.; Liu, C.; Lou, Y.; Shen, T.; Wu, Y.; Guo, J.; Li, Y.; Zhang, M. Integrating colored LiDAR and YOLO semantic segmentation for design feature extraction in Chinese ancient architecture. *npj Herit. Sci.* **2025**, *13*, 316. [Google Scholar] [CrossRef]
27. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [Google Scholar] [CrossRef]

28. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [Google Scholar] [CrossRef]
29. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [Google Scholar] [CrossRef]
30. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *Computer Vision: ECCV 2016*; Springer: Cham, Switzerland, 2016; pp. 21–37. [Google Scholar] [CrossRef]
31. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Computer Vision: ECCV 2020*; Springer: Cham, Switzerland, 2020; pp. 213–229. [Google Scholar] [CrossRef]
32. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007. [Google Scholar] [CrossRef]
33. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, *111*, 257–276. [Google Scholar] [CrossRef]
34. Tong, K.; Wu, Y. Deep learning-based detection from the perspective of small or tiny objects: A survey. *Image Vis. Comput.* **2022**, *123*, 104471. [Google Scholar] [CrossRef]
35. Fu, X.; Zhou, R.; Yang, X.; Li, C. Detecting oracle bone inscriptions via pseudo-category labels. *Herit. Sci.* **2024**, *12*, 107. [Google Scholar] [CrossRef]
36. Tao, Y.; Fu, X.; Pang, H.; Yang, X.; Li, C. Clustering-based feature representation learning for oracle bone inscriptions detection. *npj Herit. Sci.* **2025**, *13*, 296. [Google Scholar] [CrossRef]
37. Guo, M.-H.; Xu, T.-X.; Liu, J.-J.; Liu, Z.-N.; Jiang, P.-T.; Mu, T.-J.; Zhang, S.-H.; Martin, R.R.; Cheng, M.-M.; Hu, S.-M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [Google Scholar] [CrossRef]
38. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [Google Scholar] [CrossRef]
39. Gao, C.; Zhang, Q.; Tan, Z.; Zhao, G.; Gao, S.; Kim, E.; Shen, T. Applying optimized YOLOv8 for heritage conservation: Enhanced object detection in Jiangnan traditional private gardens. *Herit. Sci.* **2024**, *12*, 31. [Google Scholar] [CrossRef]
40. Oksuz, K.; Cam, B.C.; Kalkan, S.; Akbas, E. Imbalance problems in object detection: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3388–3415. [Google Scholar] [CrossRef]
41. Yang, Z.; Sinnott, R.O.; Bailey, J.; Ke, Q. A survey of automated data augmentation algorithms for deep learning-based image classification tasks. *Knowl. Inf. Syst.* **2023**, *65*, 2805–2861. [Google Scholar] [CrossRef]
42. Jocher, G.; Qiu, J. Ultralytics YOLO11 (Version 11.0.0). 2024. Available online: https://github.com/ultralytics/ultralytics (accessed on 22 August 2026). [Google Scholar]
43. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans. Cybern.* **2022**, *52*, 8574–8586. [Google Scholar] [CrossRef]
44. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 21002–21012. Available online: https://proceedings.neurips.cc/paper_files/paper/2020/hash/f0bda020d2470f2e74990a07a607ebd9-Abstract.html (accessed on 9 November 2026). [Google Scholar]
45. Padilla, R.; Passos, W.L.; Dias, T.L.B.; Netto, S.L.; da Silva, E.A.B. A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics* **2021**, *10*, 279. [Google Scholar] [CrossRef]
46. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In *Computer Vision: ECCV 2014*; Springer: Cham, Switzerland, 2014; pp. 740–755. [Google Scholar] [CrossRef]
47. Du, L.; Wang, Y. Bi-YOLO: A novel object detection network and dataset for components of China heritage buildings. *J. Build. Eng.* **2024**, *97*, 110817. [Google Scholar] [CrossRef]
48. Huggett, J. Is big digital data different? Towards a new archaeological paradigm. *J. Field Archaeol.* **2020**, *45*, S8–S17. [Google Scholar] [CrossRef]

49. Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; Loy, C.C. Domain generalization: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4396–4415. [Google Scholar] [CrossRef]
50. Jocher, G.; Qiu, J.; Liu, M.; Lyu, S.; Akyon, F.C.; Kalfaoglu, M.E. Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models. *arXiv* **2026**, arXiv:2606.03748. [Google Scholar] [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.