Heritage Science

RESEARCH

Open Access

Human figure detection in Han portrait stone images via enhanced YOLO-v5

Junjie Zhang[1], Yuchen Zhang[1], Jindong Liu[2*], Yuxuan Lan[1] and Tianxiang Zhang[2]

Abstract

The unearthed Han Dynasty portrait stones are an important part of China's ancient artistic heritage, and detecting human images in these stones is a critical prerequisite for studying their artistic value. However, high-precision target detection techniques often result in a large number of parameters, making them unsuitable for portable devices. In this work, we propose a new human image target detection model based on an enhanced YOLO-v5. We discovered that the complex backgrounds, dense group targets, and significant scale variations of targets within large scenes in portrait stones present significant challenges for human target image detection. Therefore, we first incorporated the SPD-Conv convolution and Coordinate Attention self-attention mechanism modules into the YOLO-v5 architecture, aiming to enhance the model's recognition precision for small target images within Han portrait stones and strengthen its resistance to background disturbances. Moreover, we introduce DIoU NMS and Alpha-IoU Loss to improve the detector's performance in dense target scenarios, reducing the omission of densely packed objects. Finally, the experimental results from our collected dataset of Han Dynasty stone figure images demonstrate that our method achieves fast convergence and high recognition accuracy. This approach can be better applied to the target detection tasks of special character images in complex backgrounds.

Keywords Alpha-IoU loss, Coordinate attention self-attention modules, Enhanced YOLO-v5, Han portrait stone images, Model detection

Introduction

Han portrait stones are a pivotal element of ancient Chinese heritage, providing critical insights into the art and culture of the Han Dynasty [1]. Unearthed continuously since the 20th century, these artifacts showcase a wide array of subjects, including celestial beings, spirits, historical figures, musical and dance celebrations, carriage travels, and hunting scenes [2]. The task of accurately identifying the individuals depicted has been a consistent challenge for scholars. Traditional techniques like rubbings and reproductions, central to historical studies, often inadequately capture the detailed nuances present against complex backdrops. For example, emblematic figures represented on these stones, such as Fuxi-Nüwa and the Dancer, are not isolated. Instead, they are intricately connected with surrounding motifs like the sun, moon, yin-yang symbols, geometric shapes, and traditional attire, which significantly increases the difficulty of accurately recognizing these figures.

With the continuous progress of modern technology, the use of digital imaging technology to monitor unearthed cultural relics has become increasingly feasible. These technologies not only provide more detailed and precise graphical representations and data but also ensure that the artifacts remain undisturbed and undamaged during the monitoring process. Currently, object detection methods based on image data, as indicated in the reference [3, 4], have proven the feasibility of using computer vision techniques to detect human figures on

*Correspondence:
Jindong Liu
amosliu@stumail.nwu.edu.cn
[1] School of Arts, Northwest University, Xi'an 710127, China
[2] School of Information Science and Technology, Northwest University, Xi'an 710127, China

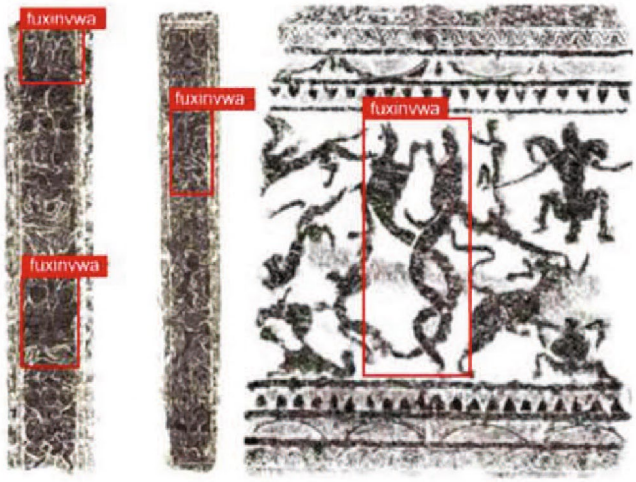
portrait stones in complex environments. Therefore, in-depth research into image detection technology is essential.

Furthermore, the integration of machine learning algorithms with digital imaging technology can enhance the accuracy and efficiency of identifying specific features and patterns in cultural artifacts. Such integration could lead to revolutionary discoveries in the way we interpret historical and cultural contexts, enabling a deeper and more comprehensive understanding of ancient civilizations. The development of technology in the field of artifact analysis emphasizes the necessity of continuous research and development in this area, providing new insights and preservation techniques for our global cultural heritage.

In recent years, significant advancements have been made in the field of object detection. Algorithms such as R-CNN [5], SSD [6], and YOLO [7] have become widely popular due to their adaptability and accuracy across a variety of images. However, as illustrated in Fig. 1, the human figures on unearthed portrait stones, with their unique artistic styles, intricate backgrounds, and diverse compositions, present a series of challenges. These challenges not only increase the complexity of object detection but also affect its accuracy and consistency. Specifically, iconic figures like Fuxi-Nüwa demonstrate a diversity in shape that significantly differs from the standard object shapes detected by conventional models. This disparity makes the task of existing object detection models more complex, hindering the effective extraction of similar features. Moreover, the distinction between object features and background becomes less clear.

In addition to figurative images, Han Dynasty stone engravings also contain rich textual symbols, presenting another significant challenge in research. The text on Han Dynasty stones not only reflects the language and writing habits of the time but is often closely integrated with figurative images and decorative patterns, forming unique cultural expressions. Therefore, the detection and recognition of these stone inscriptions are crucial for understanding the social life and culture of the Han Dynasty. However, traditional object detection models often overlook the specificity of these texts, leading to poor performance when dealing with these specific elements.

To overcome the aforementioned challenges, researchers need to develop new algorithms or improve existing ones to better accommodate the uniqueness of Han Dynasty portrait stones. This could include deep learning models specifically trained for Han Dynasty artistic styles and textual characteristics, or combining



Left figure	**Right figure**
Era: Late Eastern Han Dynasty (147-220 AD) **Cultural Relics Information:** Unearthed in Nanzhang, Yutun Town, Jining City, Shandong Province in 1970, currently housed in the Jining Museum in Shandong Province.	Era: Late Eastern Han Dynasty (147-220 AD) **Cultural Relics Information:** Unearthed in Dongjiazhuang, Anqiu City, Shandong Province in 1995, currently housed in the Anqiu City Museum.

Fig. 1 Human figures in portrait stone images

existing algorithms with semantic analysis techniques to improve the recognition rate of figures and inscriptions against complex backgrounds. Additionally, researching how to effectively separate textual and pictorial information in artifact images and conduct independent but collaborative analysis of them will be key to enhancing detection accuracy and consistency [8, 9]. Once again, an interdisciplinary research approach is particularly important here, combining archaeology, art history, computer science, and image processing technologies to deepen the understanding of the cultural and historical value of Han Dynasty stone engravings and promote the development of cultural heritage protection and research.

While YOLO-v5 performs excellently in terms of accuracy and detection speed, it encounters difficulties when addressing the unique challenges of our specific task. This is largely due to the dense clustering of targets and significant scale variations observed in the extensive scenes characteristic of portrait stones. To overcome the inherent limitations of YOLO-v5, this paper introduces an enhanced YOLO-v5 model, meticulously tailored for detecting human figures in portrait stones. Our primary goal is to improve the accuracy of object detection within such complex environments, thereby narrowing the gap between the conventional image detection capabilities of YOLO-v5 and the specialized requirements of our work.

In this paper, we have the following contributions:

(1) In our method, the introduction of the SPD-Conv convolutional layer has enhanced the model's ability to recognize small targets from a distance, while the addition of the Coordinate Attention module has optimized the model's focus and recognition of specific human images, effectively reducing the impact of complex backgrounds on detection accuracy.
(2) In our method, the implementation of DIoU NMS technology has improved the Non-Maximum Suppression (NMS) process, significantly enhancing the detector's ability to recognize targets in dense environments. Moreover, the introduction of Alpha-IoU LOSS as an optimization strategy for the loss function effectively increases the model's efficiency during the learning process, particularly demonstrating higher precision and stability in detecting human figures.
(3) A series of comparative experiments were conducted on our custom Han portrait stone human image target detection dataset to validate the practicality and advancement of our method in the field of image detection. The results of these experiments not only demonstrate the significant advantage of our method in improving detection accuracy but also exhibit our method's outstanding performance in reducing the number of model parameters.

Related works

With the widespread application of deep learning technologies in computer vision, traditional object detection algorithms like R-CNN [5], Fast R-CNN [10], and Faster R-CNN [11] are increasingly being supplanted by more efficient methods. Modern research bifurcates into two primary algorithm types: two-stage detection algorithms, which recognize objects through region proposals followed by classification, and single-stage detection algorithms, such as YOLO [7], SSD [6], and CornerNet [12], known for their direct prediction of object locations and classifications. Additionally, recent progress includes the deployment of enhanced deep learning models in niche areas like portrait stone figure detection, utilizing techniques like multi-scale sliding windows and feature sharing to boost accuracy and efficiency [13].

When choosing small object detection technologies, a balance between efficiency and accuracy is essential. Single-stage detection algorithms, celebrated for their swift performance, directly ascertain the location and category of objects, a critical feature for real-time applications like video surveillance and mobile robotics. However, this speed can compromise accuracy. Conversely, two-stage detection algorithms, which first generate numerous candidate objects and then employ advanced techniques for evaluation, offer higher precision, vital for error-sensitive areas such as autonomous driving and security surveillance. This necessitates a strategic choice between the rapid yet less precise single-stage algorithms and the slower but more accurate two-stage methods based on specific application requirements.

Introduced by Ross Girshick in 2014, the R-CNN framework combined candidate regions with CNNs for object detection [5]. Although a significant advancement, its inefficiency stemmed from processing each region individually, hampering its real-time application. Addressing this, the SPPNet, proposed by He K and colleagues in 2015, incorporated the Spatial Pyramid Pooling layer, enhancing computational efficiency [14]. Girshick further evolved this concept into Fast R-CNN [11], which, despite its speed improvements, still relied on separate proposal generation. The introduction of Faster R-CNN by S. Ren's team resolved this by integrating a Region Proposal Network (RPN) [12], facilitating near real-time proposal generation. This innovation, coupled with the Feature Pyramid Network (FPN) introduced by T.Y. Lin in 2017 [15], significantly advanced multi-scale object detection.

Recent technological strides have seen the advent of algorithms like YOLOv4 and EfficientDet, which enhance

both speed and accuracy, especially in constrained environments. YOLOv4 integrates new backbone architectures and spatial pyramid pooling to accelerate and refine detection [16], while EfficientDet leverages compound scaling and an optimized structure for peak performance [17]. These innovations mark significant milestones in the evolution of object detection, striving for a harmonious balance between speed, accuracy, and resource utilization.

Despite their expedited design, one-stage detection algorithms, renowned for their real-time application efficiency, have outshone traditional multi-stage counterparts in both speed and accuracy, underscored by developments like YOLO [7] and SSD [6]. However, challenges remain, particularly in complex detection scenarios. In response, 2017s introduction of RetinaNet by T.-Y. Lin introduced "focal loss" [18], addressing the class imbalance. The novel approaches of CornerNet [12] and CenterNet [19] further diversified detection methodologies. Additionally, 2020s DETR [20] and subsequent Deformable DETR introduced transformative shifts with the use of self-attention mechanisms. The YOLO series continued to evolve with YOLO-v4 [21], YOLO-v6 [22], and YOLO-v7 [23], enhancing detection precision, speed, and adaptability. Nonetheless, the YOLO approach, skipping region proposals, faces accuracy challenges with small or densely clustered objects, underlining the ongoing quest for balanced object detection solutions. As shown in Table 1, a comprehensive comparison of object detection algorithms has been conducted.

In summary, although existing object detection algorithms have achieved significant results, they still face several challenges: (1) These algorithms often struggle to find a balance between detection speed and accuracy; (2) Detection becomes particularly challenging when the appearance of the object closely resembles the background. To effectively address these issues, this paper proposes an enhanced YOLO-v5 model, specifically designed for the detection of figurative images in stone carvings. This algorithm improves the performance of detecting stone-carved figures in complex environments from multiple perspectives.

Object detection method for portrait stone images based on enhanced YOLO-v5 model

Conducting object detection of human images in excavated Han dynasty stones poses numerous challenges, primarily due to the background's complexity, the diversity of target forms, and the scenes' disarray. Initially, these Han dynasty stones were part of funerary art, typically manifesting in various forms, such as coffins, walls, beams, steles, arches, and lintels, as illustrated in Fig. 2 and Fig. 3. This not only enriches the cultural heritage but also increases the complexity of detection. Moreover, the characters depicted on these Han dynasty stones are closely integrated with their backgrounds, forming intricate patterns, making accurate object detection from the complex backgrounds particularly challenging. For instance, distinguishing mythological figures such as Fuxi-Nüwa in the Han dynasty is particularly challenging due to their varied depictions and close integration with the environment. The portrayals of Fuxi-Nüwa vary from human heads with snake bodies to figures holding compasses, as well as symbols of authority like the sun and the moon, adding complexity to object detection. Therefore, addressing these challenges requires a comprehensive approach that considers the background's complexity, the diversity of the targets, and the overall structure of the scenes to enhance the accuracy and consistency of detection.

Object detection modle based on YOLO-v5

YOLO is a widely-recognized series of object detection models. YOLO-v5, in comparison to its predecessors,

Table 1 Comparison of object detection algorithms

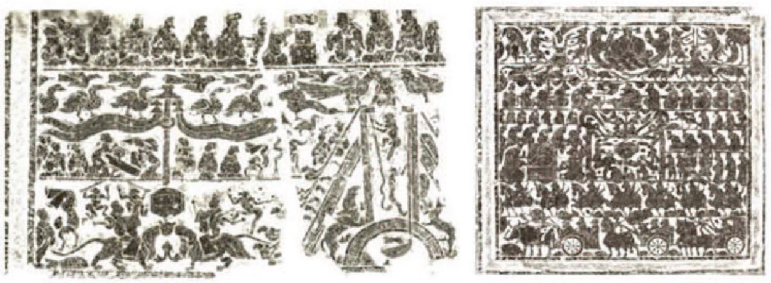
Refs.	Advantages	Disadvantages
[5]	Improves detection accuracy	Low efficiency, not for real-time
[10]	Faster than R-CNN	Requires separate region proposal
[11]	Near real-time proposals, high accuracy	Slower than single-stage methods
[7]	High-speed, suitable for real-time	Lower accuracy for small objects
[6]	Balances speed and accuracy	Less accurate in complex scenes
[12]	Innovative, improves accuracy	Complex, slow inference
[14]	Efficient, adaptable to sizes	Requires effective region proposal
[15]	Excellent in multi-scale detection	Complex, high resource needs
[16]	Increases speed and accuracy	Room for improvement in complex scenes
[17]	Efficient, balances speed and accuracy	May require tuning for specific scenarios

Upper left figure	**Upper right figure**
Era: In the third year of the Yuanhe reign of Emperor Zhang of the Eastern Han Dynasty (86 AD) **Cultural Relics Information:** Pingyi Town, Pingyi County, Shandong Province, currently located at the Pingyi County Cultural Relics Management Office.	**Era:** During the reign of Emperor Zhang of the Eastern Han Dynasty (76-88 AD) **Cultural Relics Information:** The South Xiaotang Mountain in Xiaolipu Village, Xiaoli Town, Changqing County, Shandong Province, currently located at the Cultural Relics Management Office of Changqing County, Shandong Province.
Below figure **Era:** Late Western Han to Early Eastern Han (9-88 AD) **Cultural Relics Information:** Unearthed in Hanjiapu Village, Qufu City, Shandong Province in 1937, currently housed in the Cultural Relics Management Committee of Qufu City, Shandong Province.	

Fig. 2 Variations in the Han portrait stone image dataset

boasts enhanced detection speed and accuracy. Its architecture is primarily composed of three components: the Backbone, Neck, and Head. Within YOLO-v5, the backbone network utilizes CSPMarknet53, an advanced iteration of Darknet53. Remarkably, CSP-Darknet53 elevates the network's performance without increasing its depth or parameter count. The core structure of CSP (Cross-Stage Partial Network) in CSPDarknet53 segregates and facilitates interaction between information from different stages, thereby amplifying the network's representational capacity. Specifically, CSPDarknet53 bifurcates the input at each stage into two branches: one dedicated to convolutional operations and the other to down-sampling. The outputs from these branches are subsequently concatenated and further convolved. This design strategy aims to mitigate computational complexity while circumventing issues related to gradient vanishing and explosion. Furthermore, the backbone of YOLO-v5 integrates the SPP (Spatial Pyramid Pooling) module, which pools feature maps across various scales, ensuring the preservation of critical information during feature extraction.

In the YOLO-v5 model, the PANet (Path Aggregation Network) [24] acts as the neck network, whose main function is to merge feature maps of different scales to facilitate object detection tasks. The structure of PANet consists of three main components: an upsampling module, an adaptive feature pooling module, and a downsampling module, as shown in Fig. 4. The downsampling module is responsible for extracting feature maps, while the upsampling module is used to restore the sizes of the feature maps. The adaptive feature pooling module integrates feature maps of different scales, aiming to improve the model's accuracy in segmentation tasks. In this way, PANet enhances the model's ability to detect objects of various sizes, thereby improving the overall detection performance.



Left figure	Right figure
Era: Late Eastern Han Dynasty (147-189 AD) **Cultural Relics Information:** Unearthed in the early years in Chengnan, Jining City, Shandong Province, currently housed in the Cultural Relics Management Office of Yutai County, Shandong Province.	**Era:** Late Eastern Han Dynasty (147-189 AD) **Cultural Relics Information:** Unearthed in Xihukou Village, Sangcun Town, Tengzhou City, Shandong Province in 1958, currently housed in the Tengzhou City Museum.

Fig. 3 Complex Backgrounds in the Han portrait stone image dataset

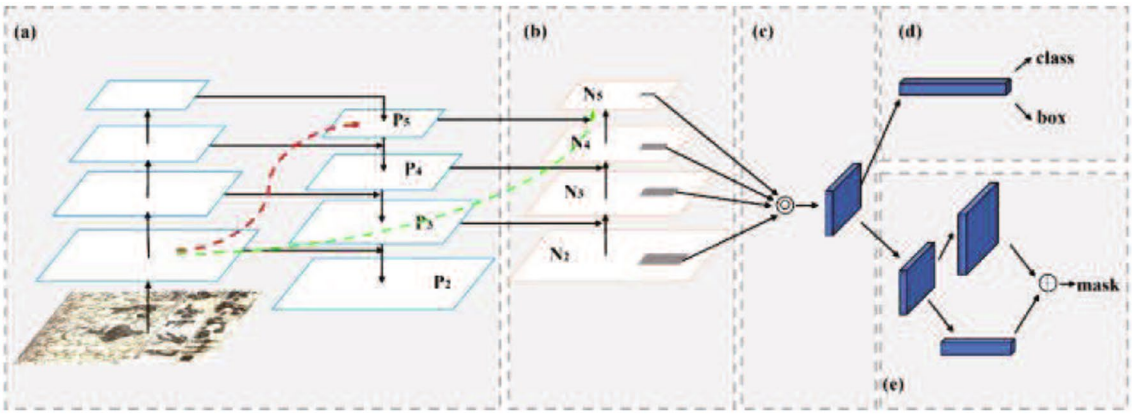


Fig. 4 PANet Framework Diagram **a** FPN Backbone Network; **b** Bottom-Up Path Enhancement; **c** Adaptive Feature Pooling; **d** Bounding Box Branch; **e** Fully Connected Fusion

Enhanced YOLO-v5 model for object detection in Han dynasty stone images

Despite the significant adaptability and accuracy that modern object detection algorithms such as RCNN, YOLO, and SSD have demonstrated in various environments, they still exhibit limitations when facing specific challenges, particularly in detecting human images in excavated Han dynasty relief stones. These challenges include the diversity of the Han relief stones, the complexity of the backgrounds, and the intricate combination of elements, all of which can impact the performance of the algorithms. In particular, the traditional YOLO-v5 model faces challenges such as chaotic backgrounds, dense clusters of objects, and significant scale variations of objects in large scenes when detecting figures against these complex backgrounds, which can all degrade detection performance.

In light of this, we propose an enhanced YOLO-v5 model specifically designed to overcome the unique challenges encountered in object detection of human images against the backdrop of Han dynasty relief stones. This model is built upon the framework of YOLO-v5, incorporating the following enhanced modules to address the detection difficulties in complex environments, thereby improving detection accuracy and consistency in these specific scenarios.

(1) Introduction of SPD-Conv Convolutional Layers: We have replaced the conventional convolutional layers in YOLO-v5 with SPD-Conv layers, enhancing the detection capability for distant small objects.
(2) Coordinate Attention Self-Attention Modle: The model can focus more on small-object human figures, reducing interference from complex backgrounds.
(3) DIoU NMS: The DIoU NMS technique has been utilized, improving Non-Maximum Suppression (NMS) to optimize the post-processing of object detection results, thereby enhancing the detector's ability to recognize densely packed targets.
(4) Alpha-IoU LOSS: Proposed Alpha-IoU LOSS optimizes the model's loss function, thereby improving the model's learning effectiveness and enhancing its ability to detect human images.

The structure of the enhanced YOLO-v5 network model we designed is shown in Fig. 5, where the red frame indicates the introduction of the SPD module, and the blue frame indicates the addition of the Coordinate Attention self-attention module. Next, we will provide a detailed analysis of these modles:

SPD-Conv

Convolutional Neural Networks (CNNs) have achieved significant progress in computer vision tasks such as object detection and image classification. However, their performance significantly decreases when dealing with lower-resolution images or smaller object sizes. This issue arises because small and large objects may coexist within the same image, with larger objects often dominating the feature learning process, making it difficult to detect smaller objects. This problem originates from common designs in traditional CNN architectures, such as stridden convolutions and pooling layers, which can reduce computational costs but lead to the loss of fine-grained information and decreased efficiency in feature representation. To address this issue, we introduce a new type of CNN module named SPD-Conv, designed to replace traditional stridden convolutions and pooling layers. SPD-Conv combines space-to-depth (SPD) layers with non-strided convolution (Conv) layers, enabling more

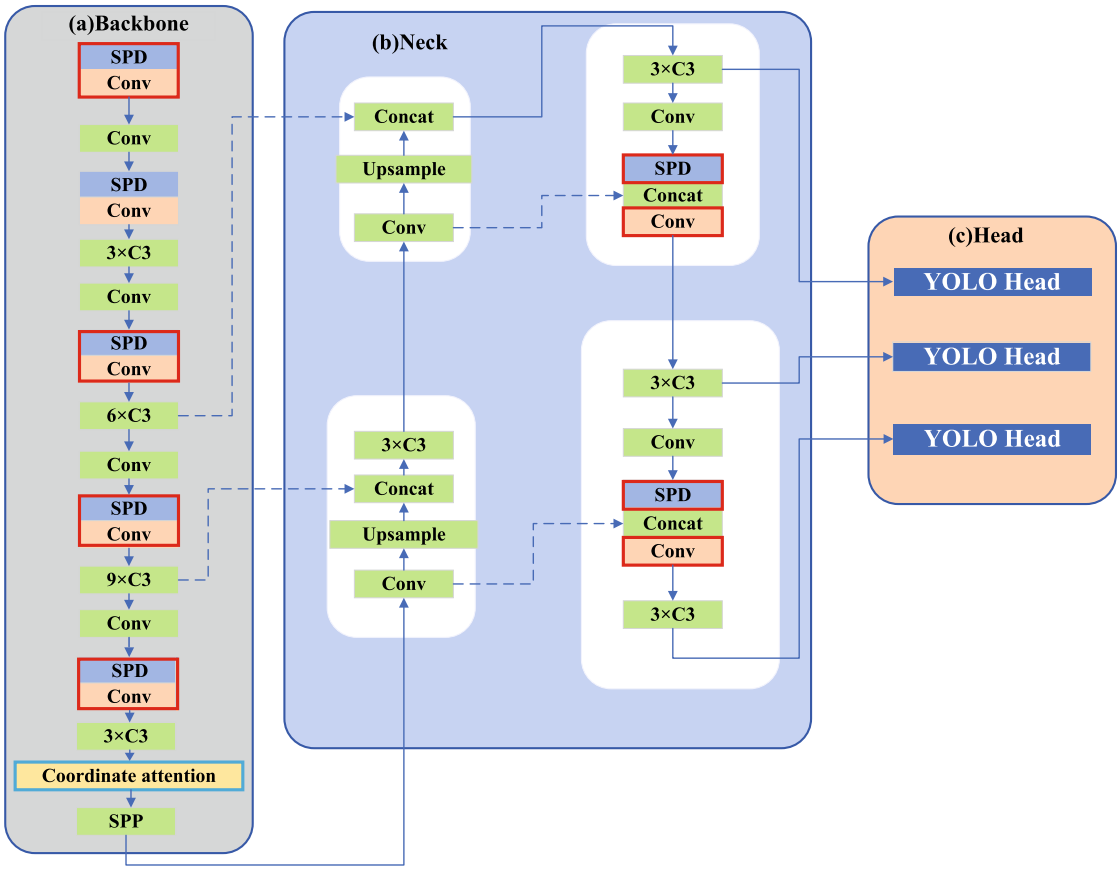


Fig. 5 Structure of the enhanced YOLO-v5 network model

effective preservation of fine-grained information and improved efficiency in feature representation, thereby enhancing the accuracy and efficiency of object detection. Specifically, SPD-Conv demonstrates its superiority when applied to the Fu Xi and Nu Wa image dataset, which contains a large number of small object targets.

(1) SPD (space-to-depth) layer

As shown in Fig. 6 (a)(b)(c), given scale = 2, four sub-feature maps are obtained: $f_{0,0}$, $f_{1,0}$, $f_{0,1}$ and $f_{1,1}$, each with a shape of $(\frac{S}{2}, \frac{S}{2}, C_1)$. Additionally, the original X is down-sampled by a factor of 2.

These sub-feature maps are then concatenated along the channel dimension to obtain a feature map X', whose spatial dimensions are reduced by a factor of *scale*, while the channel dimension is expanded by a factor of $scale^2$. In other words, SPD transforms the feature map $X(S, S, C_1)$ into the feature map $X'(\frac{S}{scale}, \frac{S}{scale}, scale^2 C_1)$. Figure 6 (d) provides an illustration of this process when $scale = 2$ is applied.

(2) Conv (non-strided convolution) layer

After the SPD feature transformation layer, an additional Conv($scale = 1$) layer is added, with $C_2 < scale^2 C_1$, further transforming the feature map $X'(\frac{S}{scale}, \frac{S}{scale}, scale^2 C_1)$ into $X''(\frac{S}{scale}, \frac{S}{scale}, C_2)$. The reason for using the Conv layer is to preserve all discriminative feature information as much as possible. Otherwise, if a 3×3 filter with $scale = 3$ is used, the feature map would be "downsized", but each pixel would be sampled only once; if $scale = 2$, asymmetric sampling would occur, where the sampling times for even and odd rows/columns differ. In general, $scale > 1$ would lead to loss of information.

As shown in Fig. 6, this paper has modified the YOLOv5 framework by replacing its $scale = 2$ convolution layers with SPD-Conv layers. In the backbone network of YOLOv5, a total of five locations require adjustments, especially in the neck region, which includes two $scale = 2$ convolution layers. Furthermore, after each stridden convolution layer in the neck of YOLOv5, a concatenation layer is implemented. In this paper, we have retained this arrangement of the concatenation layer between SPD-Conv and regular Conv layers.

Through experiments, we found that YOLOv5-SDP demonstrates good performance in detecting small object human images, being able to identify two types of instances, namely Fuxi-Nüwa images. However, in complex background images, particularly those with overlapping and merged human figures, some instances of Fuxi-Nüwa images still cannot be effectively detected.

Coordinate attention self-attention module

The Han dynasty stones contain diverse and complex combinations of human images, which often closely resemble their surrounding environments. This complexity not only enriches the narrative but also increases

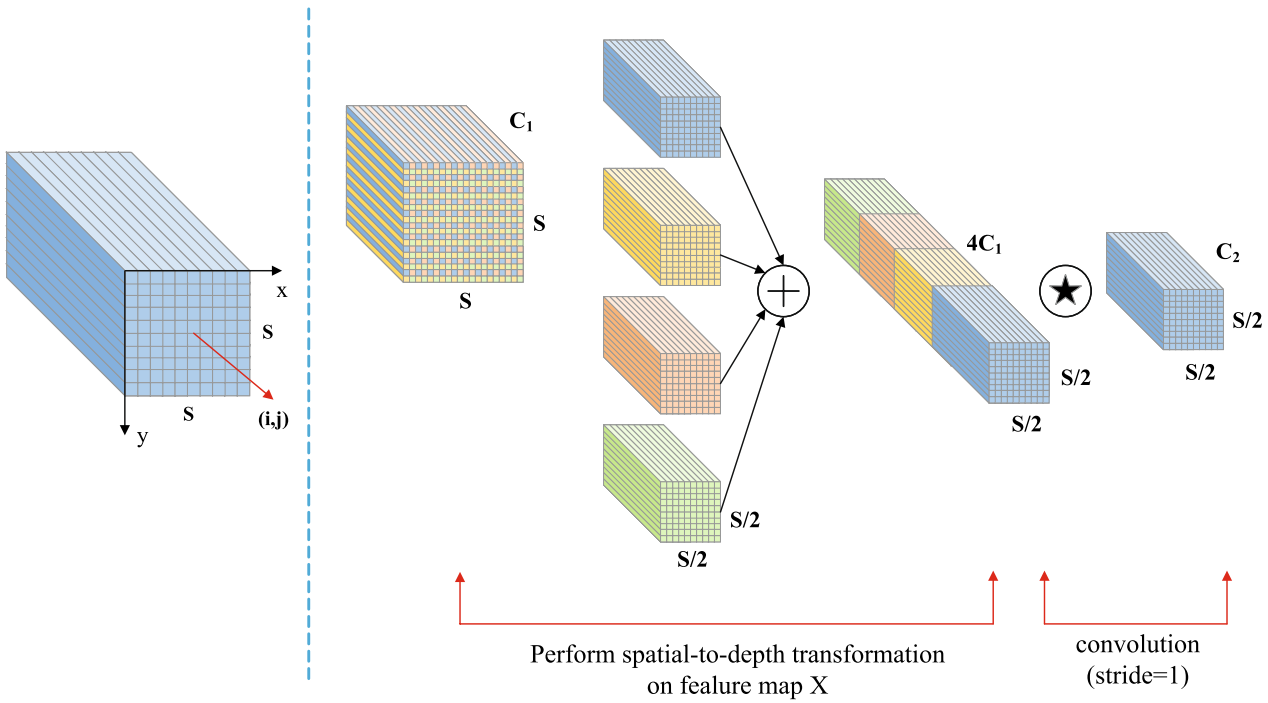


Fig. 6 Instance of SPD-Conv at $scale = 2$

the difficulty of detecting human targets. Therefore, it is important to focus on the similarity and uniqueness of the images with their surroundings to reduce the chances of missing or falsely detecting targets.

To enhance the performance and feature expression capabilities of the model, we have incorporated a Self-Attention Module (SAM) into the YOLOv5 framework. SAM is capable of learning the relationships between different positions within the input feature maps and applying weighted processing to the features, thereby strengthening the model's ability to recognize targets and suppress background interference, further improving classification accuracy and detection precision.

Furthermore, we have specifically introduced the Coordinate Attention mechanism, an effective feature weighting method that captures the correlation between different positions in the feature map by calculating the coordinates of feature points. This approach allows for a more precise focus on the importance of objects at different positions within the image, thus enhancing the model's performance in object detection tasks. Compared to traditional attention mechanisms, Coordinate Attention not only emphasizes the relationships between channels but also highlights the spatial relationships on the feature map, integrating position information into the attention calculation. This enables the model to perceive the positional information on the feature map more effectively. By enhancing the network's expressive capability, Coordinate Attention helps the model to understand the input image more deeply and locate targets more accurately. The structure of Coordinate Attention is shown Fig. 7.

The implementation process of Coordinate Attention starts by extracting a set of feature maps from the input image. Then, the coordinate values of each point within these feature maps are normalized to produce relative coordinates that are independent of the actual size of the image. This step ensures that the location information of the features is standardized. Subsequently, these normalized coordinates are fed into a neural network to assess the similarity between feature points. In this way, a weighting mechanism based on similarity is applied to features at different positions, thus enhancing the expressiveness of the features and improving the overall performance of the model.

In summary, by integrating the coordinate information of objects into the self-attention mechanism, Coordinate

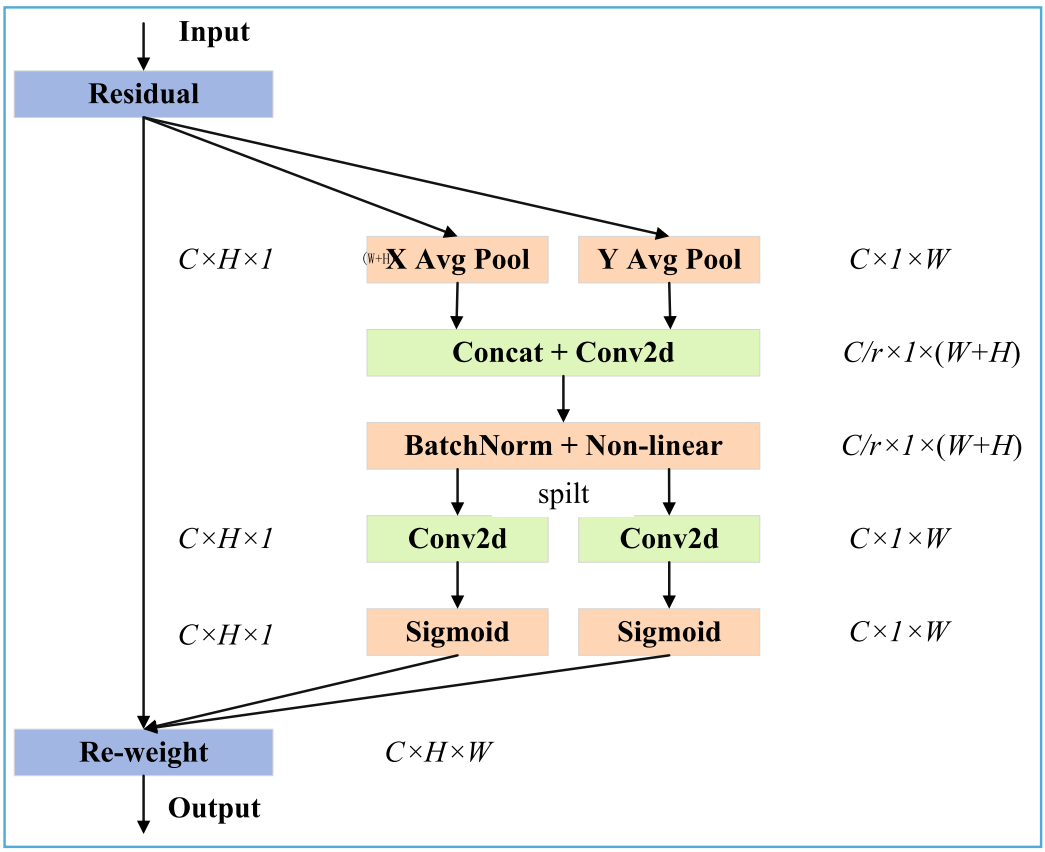


Fig. 7 Coordinate attention self-attention module

Attention provides a new perspective for the model. This mechanism allows the model to capture the spatial positions of objects on the feature maps more accurately, thereby achieving higher accuracy and efficiency in complex tasks such as object detection. Finally, we have successfully integrated the Coordinate Attention module into the backbone network of YOLOv5, especially before the SPP (Spatial Pyramid Pooling) layer, as shown in Fig. 4. This arrangement enables the module to process input data in the early stages of the model, helping the model to more deeply analyze the spatial information of objects, thereby enhancing detection performance.

DIoU NMS

In YOLOv5, the default Non-Maximum Suppression (NMS) algorithm primarily filters candidate boxes by calculating the Intersection over Union (IoU). NMS employs an iterative approach, starting with the highest-scoring box, and progressively removes boxes that overlap significantly with it until a set of mutually independent high-quality detection results is retained. However, NMS has certain limitations: firstly, if the overlap of the detection boxes exceeds a preset threshold, NMS sets their scores to zero, which might lead to missed detections of real objects. Secondly, the performance of NMS largely depends on the setting of the overlap threshold and score threshold; selecting these parameters often requires manual adjustment, making it challenging to adapt to different scenes and datasets. Moreover, the traditional IoU metric only considers the location and size of detection boxes and does not account for the distance between boxes, which may not be ideal for handling small or densely packed targets.

To address these issues, Distance IoU (DIoU) NMS has been introduced. This method uses DIoU as the metric, which can more accurately represent the distance relationship between detection boxes, thereby improving the detection performance for small and densely packed objects. DIoU adds a penalty term to the IoU loss, directly minimizing the normalized distance between two detection boxes, thus optimizing the detection results, as shown in Eq. (1).

$$\begin{aligned} R_{DIoU} &= \frac{\rho^2(b,b^{gt})}{c^2} \\ [10pt]S_i &= \begin{cases} S_i, IoU - R_{DIoU}(M,B_i) < \varepsilon \\ [10pt]0, IoU - R_{DIoU}(M,B_i) \geq \varepsilon \end{cases} \end{aligned} \tag{1}$$

In Equation (1), R_{DIoU} denotes the penalty term introduced. b and b^{gt} respectively represent the centers of the predicted and the ground truth boxes. The function $\rho(\cdot)$ signifies the Euclidean distance between these two centers, while c is the diagonal length of the smallest enclosing box that covers both the predicted and the ground truth boxes. Lastly, ε refers to the threshold for the Intersection over Union (IoU).

The DIoU algorithm not only considers the overlap between boxes but also the distance between their center points, providing a more accurate representation of the distance relationship between the boxes. This refined approach improves the sorting and filtering process of the boxes. Especially when there is a significant distance between two boxes, DIoU reduces their IoU value, effectively minimizing unnecessary box overlaps. As shown in Fig. 8, compared to traditional methods, DIoU NMS is more effective in retaining the correct detection boxes.

The operational process of the DIoU NMS algorithm is similar to that of the traditional IoU-based NMS and mainly includes the following steps:

(1) Sort all predicted boxes in descending order based on their confidence scores.
(2) Select the predicted box with the highest confidence score and add it to the output list.
(3) For the remaining predicted boxes, compare them with the boxes already in the output list. If their DIoU value exceeds a set threshold, set their score to 0; otherwise, add them to the output list.
(4) Repeat steps (2) and (3) until all predicted boxes have been processed.

The detection results of DIoU NMS and the original NMS at different thresholds are compared in Fig. 8. It is evident that DIoU NMS surpasses the original NMS across all thresholds. DIoU NMS records its lowest performance at a threshold of $\varepsilon = 0.43$, scoring 83.24, whereas the original NMS reaches its peak performance at a threshold of $\varepsilon = 0.48$, with a score of 83.14. The comparison highlights that DIoU NMS still exceeds the peak performance of the original NMS by 0.10 percentage points, even in its least favorable scenario.

Owing to its consideration of the distance and shape information between objects, the DIoU NMS algorithm significantly improves upon the traditional IoU NMS, especially in scenarios involving overlapping or similar objects. Consequently, DIoU NMS bolsters the detector's ability to identify densely clustered objects while diminishing the likelihood of missing such targets.

Alpha-IoU lOSS

The positional loss function utilized in YOLOv5 is CIoU (Complete Intersection over Union), which is a distance metric tailored for bounding box alignment in object detection tasks. CIoU refines the original IoU to more precisely measure the distance between two bounding boxes. Nonetheless, CIoU faces challenges in effectively

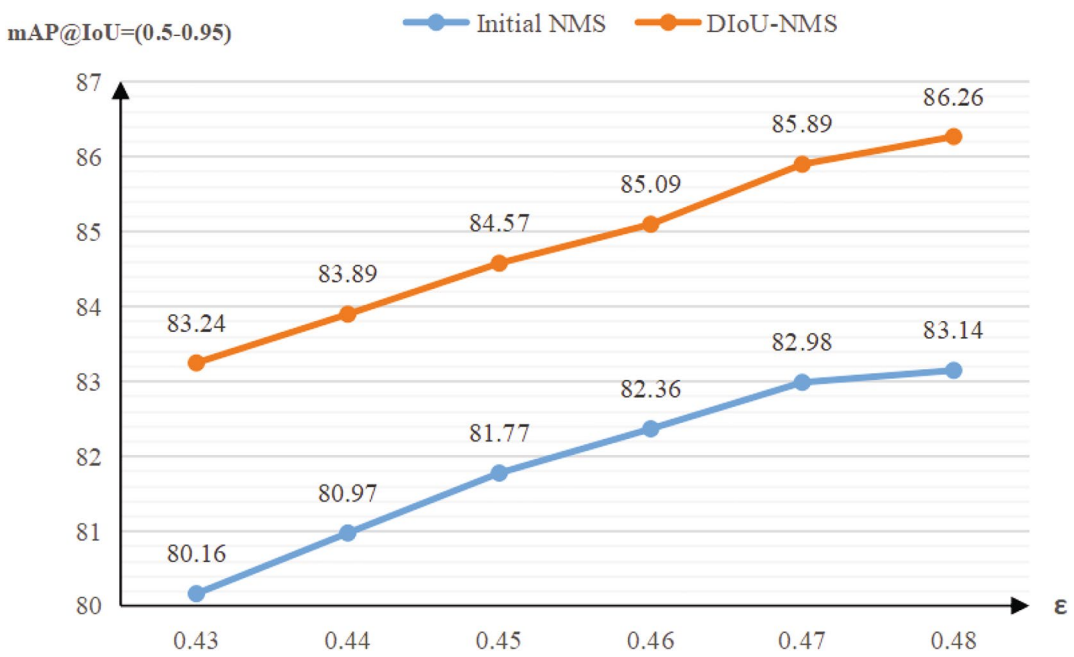


Fig. 8 Comparing the Results of DIoU-NMS and Initial NMS at Different Thresholds

managing bounding boxes with extreme aspect ratios and may encounter issues of gradient explosion during the training phase of object detection, resulting in unstable training.

To overcome these issues, Alpha-IoU introduces an additional learnable hyperparameter, alpha, which adjusts the distance metric between bounding boxes. This modification aims to alleviate some of the shortcomings associated with CIoU. Throughout the training process, the neural network autonomously adjusts the alpha value towards an optimal level to enhance the overlap between the predicted and actual bounding boxes, as illustrated in Eq. (2). By doing so, the method boosts the model's robustness and capacity for generalization. Consequently, Alpha-IoU contributes to an improvement in both the precision and stability of object detection.

$$\begin{aligned} \mathcal{L}_{\mathrm{CIoU}} &= 1 - IoU + \frac{\rho^2(b,b^{gt})}{c^2} + \beta v \\ &\downarrow \\ \mathcal{L}_{\alpha-\mathrm{CIoU}} &= 1 - IoU^{\alpha} + \frac{\rho^{2\alpha}(b,b^{gt})}{c^{2\alpha}} + (\beta v)^{\alpha} \end{aligned} \tag{2}$$

Experimental

Dataset and evaluation metrics

Datasets To verify the effectiveness of our model, we constructed a target detection dataset of Han Dynasty portrait stone images and conducted ablation and comparative experiments using the proposed method. The sources of the dataset are detailed in Table 2, and the data collection was carried out according to the following process:

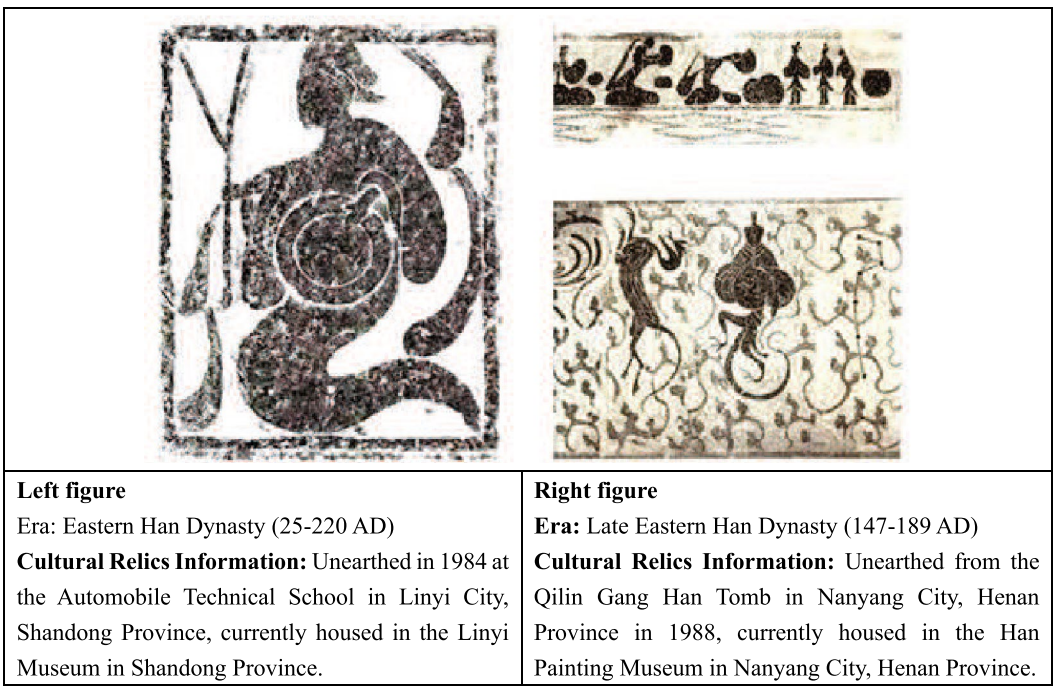
(1) Equipment: The Han Dynasty portrait stone images were captured using a Nikon D600 camera with a resolution of over 20 million pixels.
(2) Image clarity requirements: The camera equipment was fixed to prevent shaking and other issues during the shooting process.
(3) Lighting requirements: To accommodate the diversity of outdoor environments, the captured image data includes conditions such as strong illumination, shadows, and normal lighting.

The constructed dataset consists of three parts: the training set, the validation set, and the test set, and the data were manually annotated. Specifically, the training set contains 123 images, the validation set contains 80 images, and the test set contains 82 images. The Han Dynasty portrait stone image dataset includes two targets, namely Fuxi-Nüwa, and music and dance scenes, with some of the data shown in Fig. 9.

Evaluation metrics Evaluation metrics are crucial for assessing the advantages and limitations of object detection models because they provide a means to quantify the performance of the model on specific tasks. Accurate evaluation metrics can not only help us compare the performance of different models but

Table 2 Dataset Sources

Figs.	Description	Excavation date	Excavation site
Fig.1 Left	Serpent-tailed humans, fish-bodied humans, mythical creatures images	1970	South of Yutun Town, Jining City, Shandong Province
Fig.1 Right	Double-headed tiger, Fuxi-Nüwa images	Dec 1959-Mar 1960	Dongjiazhuang, Anqiu City, Shandong Province
Fig.2 Upper Left	Imperial Saint Que East Que	1932	"Babu Top" north of Pingyi County City, Shandong Province
Fig.2 Upper Right	Stone shrine images	1954	Xiaotangshan, Changqing County, Shandong Province
Fig.2 Lower	Stone sarcophagus partition images	1937	Hanjiapu Village, Qufu City, Shandong Province
Fig.3 Left	Late Eastern Han Dynasty stone images	1973	South of Jining City, Shandong Province
Fig.3 Right	Queen Mother of the West, preacher figures, drum construction images	1958	Xihukou Village, Sangcun Town, Tengzhou City
Fig.9 Left	Fuxi holding the sun image	1984	Automotive Technical School, Linyi City, Shandong Province
Fig.9 Upper Right	Music, dance, and a hundred plays images	Late 1980s to early 1990s	Qilin Gang, Wolong District, Nanyang City, Henan Province
Fig.9 Lower Right	Celestial phenomena images	Late 1980s to early 1990s	Qilin Gang, Wolong District, Nanyang City, Henan Province
Fig.10 Left	Serpent-tailed humans, fish-bodied humans, mythical creatures images	1970	South of Yutun Town, Jining City, Shandong Province
Fig.10 Middle	Lishi Màozhuāng Tomb No. 4 lintel images	1919	Màozhuāng Tomb No. 4, Lishi, Shanxi Province
Fig.10 Right	Xihe holding the sun image	1990	Gaoli Village, Guoli Township, Zoucheng City, Shandong Province
Fig.11 Left	Acrobatics, kitchen scenes images	1968	Near the Normal School, Zoucheng City, Shandong Province
Fig.11 Middle	Wu Family Shrine left and right chamber east wall stone images	Qing Dynasty, Qianlong 51st year	North of Wuzhai Mountain Village, Jiaxiang County, Shandong Province
Fig.11 Right	Figures, drum construction, mythical creatures images	1970	South of Yutun Town, Jining City, Shandong Province



Left figure	**Right figure**
Era: Eastern Han Dynasty (25-220 AD) **Cultural Relics Information:** Unearthed in 1984 at the Automobile Technical School in Linyi City, Shandong Province, currently housed in the Linyi Museum in Shandong Province.	**Era:** Late Eastern Han Dynasty (147-189 AD) **Cultural Relics Information:** Unearthed from the Qilin Gang Han Tomb in Nanyang City, Henan Province in 1988, currently housed in the Han Painting Museum in Nanyang City, Henan Province.

Fig. 9 Partial Data Display of the Han Portrait Stone Image Dataset

also guide the improvement and optimization of future models. Below are the main evaluation metrics we have chosen:

(1) AP (Average Precision): AP is one of the most common evaluation metrics in the field of object detection, representing the average precision value. It is obtained by calculating the area under the precision-recall curve, and an increase in the AP value indicates an improvement in model performance.
(2) APS: In the evaluation of object detection, targets with an area less than a certain pixel value are classified as small objects. The APS value measures the model's detection capability for small objects, with higher APS values indicating better detection performance for small targets.
(3) APM: Defined as the average precision for medium-sized targets, whose area falls between two predetermined pixel values. A higher APM value signifies stronger detection capability for medium-sized targets.
(4) APL: For targets whose area exceeds a certain pixel threshold, they are classified as large objects. The APL value indicates the model's performance in detecting large objects, with higher APL values indicating superior detection capabilities for large targets.
(5) mAP: Represents the mean of the Average Precision across all categories and is a crucial metric for evaluating the overall performance of an object detection algorithm. An increase in mAP reflects an improvement in the model's performance across all categories. By integrating the results of these metrics, we can observe whether the model meets the needs of detecting human images in the complex environment of portrait stones.

Experimental details

In this experiment, we utilized a dataset specifically designed for target detection of Han Dynasty portrait stone figures. The experimental environment was set up as follows: the operating system was Ubuntu 16.04, with computational resources including four NVIDIA TITAN Xp graphics cards and an Intel(R) Xeon(R) CPU E5–2678 v3 @ 2.50GHz processor. The deep learning framework used was PyTorch, with a version number of 1.8.0, and the Python version was 3.7. For model training, we utilized the official yolov5.pt as the pre-trained model. During the training process, the initial learning rate was set to 0.001, batch size was set to 16, and the number of epochs was set to 1000.

Ablation experiment

In this section, we present a comprehensive set of ablation experiments conducted on the Han Dynasty portrait stone human image object detection dataset to validate the effectiveness of each module. To ensure the convincingness of the ablation results for each module, we specifically selected the SPD-Conv (SPD), Coordinate Attention mechanism (CA), DIoU NMS (DIoU), and Alpha-IoU LOSS (Alpha) modules, and employed the mAP@IoU=(0.5−0.95) metric for both quantitative and qualitative analysis.

According to the results shown in Table 3, after integrating the SPD-Conv module into the base YOLOv-5 framework, the detection performance for Fuxi-Nüwa on the mAP(50) and mAP(50–95) evaluation metrics increased by 0.40 and 0.11 points, respectively. For the Dancer figure detection, the improvements were 0.79 and 0.49 points, respectively. Since the Han Dynasty portrait stone image dataset contains very few small targets, the scope for performance improvement is limited. The most notable performance enhancement was observed with the addition of the Coordinate Attention self-attention module: the detection metrics for Fuxi-Nüwa increased by 2.07 and 1.20 points, respectively, on mAP(50) and mAP(50–95), and for the Dancer figures, the increases were 1.48 and 0.86 points, respectively. This improvement is primarily due to the high similarity between the figures of Fuxi-Nüwa, as well as the Dancer figures, and their backgrounds. The Coordinate Attention module enhances the model's focus on these specific targets, reducing background interference with detection performance. Meanwhile, due to the scarcity of dense objects in the dataset, the performance improvements from introducing the DIoU and Alpha modules are also relatively limited.

Table 3 Results from the ablation experiments on the Han portrait stone image dataset for detecting Fuxi-Nüwa, and Dancer scenes. mAP(50) (mAP@IOU=(0.5)); mAP(50–95) (mAP@IOU=(0.5−0.95))

Method	mAp(50)		mAp(50-95)	
	Fuxi-Nüwa	Dancer	Fuxi-Nüwa	Dancer
YOLO-v5	64.02	43.96	47.36	27.10
YOLO-v5+SPD	64.42	44.75	47.47	27.59
YOLO-v5+CA	66.09	45.44	48.56	27.96
YOLO-v5+DIoU	64.15	44.06	47.39	27.16
YOLO-v5+Alpha	64.73	44.71	47.86	27.63
YOLO-v5+SPD+CA+DIoU+Alpha	**66.75**	**45.93**	**49.04**	**28.53**

Further investigation has revealed that our proposed enhanced YOLO-v5 model for Han stone relief human figure image object detection has achieved significant performance improvements. Particularly in the detection of Fuxi-Nüwa, the improvement in mAP(50) reached 2.73, ultimately achieving a score of 66.75; while the increase in mAP(50–95) was 1.68, ultimately reaching 49.04. In the detection of Dancer figures, the mAP(50) increased by 1.97, ultimately reaching 45.93; and the mAP(50–95) increased by 1.43, ultimately reaching 28.53. The partial detection results on the test set of the Han stone relief image dataset are shown in Fig. 10 and Fig. 11.

Comparative experiment

To validate the effectiveness of the proposed model, we selected a range of classic object detection algorithms and conducted five rounds of experiments on the Han Dynasty portrait stone figure image detection dataset. The best results from these experiments were chosen for comparison. As Table 4 demonstrates, the experimental outcomes confirm the superior performance of our proposed method in detecting targets within Han Dynasty portrait stone figure images, particularly excelling in tasks involving complex background human image detection. Moreover, our method reached scores of 39.64 and 53.65 on the APS and APM evaluation metrics, respectively, outperforming other methods, especially in detecting small targets. Compared to the second-best YOLOv7 algorithm, our proposed method achieved improvements of 3.48 and 1.31 on the APS and APM metrics, respectively.

At the same time, considering the presence of numerous small and occluded targets within the dataset, we have enhanced the detection capabilities for small and dense targets by incorporating SPD-Conv convolution, using DIoU NMS, and implementing the Alpha-IoU Loss module, which has reduced the missed detection of dense targets. Consequently, our method has demonstrated higher accuracy and robustness in detecting small and dense targets. It has performed exceptionally well on the APL evaluation metric, surpassing other leading algorithms by 3.28 compared to YOLOv7, by 3.91 compared to YOLOX, and by 5.75 compared to Deformable DETR, thanks to the introduction of the Coordinate Attention mechanism, which further focuses the model's attention on the targets and enhances detection performance. On the mAP (50–95) metric, our proposed method

Left figure	Middle figure	Right figure
Era: Late Eastern Han Dynasty (147-220 AD) **Cultural Relics Information:** Unearthed in Nanzhang, Yutun Town, Jining City, Shandong Province in 1970, currently housed in the Jining Museum in Shandong Province.	Portrait of the Right Pillar of Tomb Gate No. 4 in Lishi Mamaozhuang **Cultural Relics Information:** In 1990, it was unearthed from the No. 4 Han Tomb in Majiazhuang, Lishi County, Shanxi Province. currently housed in the Han Painting Stone Museum in Lishi County, Shanxi Province.	**Era:** Late Eastern Han Dynasty (147-189 AD) **Cultural Relics Information:** In 1990, it was unearthed in Gaoli Village, Guoli Township, Zoucheng City, Shandong Province. currently stored in Mengmiao, Zoucheng City, Shandong Province.

Fig. 10 Partial examples of the detection results for Fuxi-Nüwa using the enhanced YOLO-v5 model

Left figure	Middle figure	Right figure
Era: Early Eastern Han Dynasty (25-88 AD) **Cultural Relics Information:** Unearthed near the Normal School in Zoucheng City, Shandong Province in 1968, currently housed in Mengmiao, Zoucheng City, Shandong Province.	**Era:** Built around the second year of Emperor Huan of the Eastern Han Dynasty **Cultural Relics Information:** In the 51st year of the Qianlong reign of the Qing Dynasty (1786 AD), it was unearthed in the north of Wuzhaishan Village, Jiaxiang County, Shandong Province. currently stored in the Wushi Temple in Jiaxiang County, Shandong Province.	**Era:** Late Eastern Han Dynasty (147-189 AD) **Cultural Relics Information:** Unearthed in 1970 by Chengnan Zhang, Yutun Town, Jining City, Shandong Province, currently housed in the Jining Museum, , Shandong Province.

Fig. 11 Partial examples of the detection results for Dancer using the enhanced YOLO-v5 model

Table 4 Comparative Experiments on the Han Portrait Stone Image Object Detection Dataset

Method	APS	APM	APL	mAP(50-95)
SSD [6]	44.68	39.73	31.91	21.92
CornerNet [12]	48.96	41.10	36.17	23.29
CenterNet [25]	55.32	43.84	38.30	24.66
RetinaNet [18]	53.19	42.47	42.55	24.66
YOLOv7 [23]	59.57	45.21	44.68	26.03
YOLOX [26]	61.70	46.58	46.81	27.40
DETR [20]	55.32	42.74	38.30	24.66
Deformable DETR [27]	57.45	45.21	40.42	26.03
Our Method	**66.75**	**47.94**	**49.04**	**28.53**

also achieved the best detection results. Therefore, our research method can more effectively adapt to the task of detecting portrait stone figure images in complex environments.

Conclusion

In this paper, we propose an enhanced YOLOv-5 model, which includes SPD-Conv convolution, the Coordinate Attention self-attention module, as well as the DIoU NMS and Alpha-IoU Loss modules. We utilize the SPD-Conv convolution and Coordinate Attention module to improve the model's detection capability for small objects and its resistance to background interference. Moreover, by integrating the DIoU NMS and Alpha-IoU Loss modules, we have effectively enhanced the model's performance in detecting dense objects, significantly reducing the rate of missed detection of dense objects. Experimental results demonstrate that our method significantly surpasses other existing methods in the task of detecting human images in Han portrait stone, achieving efficient and accurate detection performance. However, despite the clear advantages of the proposed method, there are certain limitations. Specifically, due to the large number of parameters in the Coordinate Attention module, even after incorporating strategies intended to reduce the model's parameter count, the final size of the model still exceeds that of other lightweight models. Meanwhile, in future work, we will delve into the potential improvements that the latest YOLO-v8 and YOLO-v9 models may bring in small object detection. This exploration aims to improve the model's ability to extract shallow features to address the limitations of our current method. Finally, we hope our work can inspire further application of deep learning in the study of Han portrait stones.

While ensuring accurate detection of human images in Han portrait stones, in-depth research into model lightweight is also explored.

Acknowledgements
National Social Science Foundation Youth Project "Suburban Music and the Great Unification Theory of the Han Dynasty"(18CZS011); The 71st batch of China Postdoctoral Science Foundation Exploring the Origins of "Yueji": "Cultural Dissemination and the Formation of Confucian Music Classics in the Warring States Qin and Han Dynasties"(2022M712581).

Availability of data and materials
Some data, models, and code generated or used during the study will be available under reasonable request from the corresponding author.

Declarations

Competing interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: 11 December 2023 Accepted: 2 April 2024
Published online: 10 April 2024

References

1. Chang K.-c. Early chinese civilization 1976;**23**.
2. Ebrey P. Later han stone inscriptions. Harvard J Asiatic Stud. 1980. 40:325–53.
3. Zou Z, Chen K, Shi Z, Guo Y, Ye J. Object detection in 20 years: a survey. Proc IEEE. 2023. 111:257–76.
4. Li Q, Chen Y, Zeng Y. Transformer with transfer cnn for remote-sensing-image object detection. Remote Sensing. 2022. 14:984.
5. Girshick R, Donahue J, Darrell T, Malik. Jitendra, rich feature hierarchies for accurate object detection and semantic segmentation. Proc IEEE Computer Vision Pattern Recogn. 2014. 98:580–7.
6. Wei L, Dragomir A, Dumitru E, Christian S, Scott R, Cheng-Yang F, Berg, A.C. Ssd. 2016. Single shot multibox detector. European Computer Vision(ECCV). 21–37
7. Redmon J, Divvala S, Girshick R, Farhadi ACB. Ali: You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. 779–788
8. Cerra D, Plank S, Lysandrou V, Tian J. Cultural heritage sites in danger-towards automatic damage detection from space. Remote Sensing. 2016. 8:781.
9. 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. Heritage Sci. 2024. 12:31.
10. Girshick R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision. 2015;1440–1448
11. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inform Process Syst. 2015. https://doi.org/10.1109/TPAMI.2016.2577031.
12. Law H, Deng J. Cornernet: Detecting objects as paired keypoints. Proceedings of the European conference on computer vision (ECCV). 2018;734–750
13. Botifoll M, Pinto-Huguet I, Arbiol J. Machine learning in electron microscopy for advanced nanocharacterization: current developments, available tools and future outlook. 2022. Nanoscale Horizons.
14. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Machine Intell. 2015. 37: 1904–16.
15. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. 2117–2125
16. Bochkovskiy A, Wang C-Y, Liao H-YM. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint. 2020. arXiv:2004.10934.
17. Tan M, Pang R, Le QV. Efficientdet: Scalable and efficient object detection, 2020. 10781–10790
18. Lin TY, Goyal P, Girshick R, He K, Dollár P. Mdct: Multi-kernel dilated convolution and transformer for one-stage object detection of remote sensing images. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision. 2017. 2980–2988
19. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q. Centernet: Keypoint triplets for object detection. Proceedings of the IEEE/CVF international conference on computer vision. 2019. 6569–6578
20. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. European conference on computer vision. 2020. 213–229
21. Jiang J, Fu X, Qin R, Wang X, Ma Z. High-speed lightweight ship detection algorithm based on yolo-v4 for three-channels rgb sar image. Remote Sensing. 2021;13:1909.
22. Li C, Li L, Jiang H, Weng K, Geng Y, Li L, Ke Z, Li Q, Cheng M, Nie W, Li Y, Zhang B, Liang Y, Zhou L, Xu X, Chu X, Wei X, Wei X. Yolov6: A single-stage object detection framework for industrial applications. arXiv preprint. 2022. arXiv:2209.02976.
23. Wang CY, Bochkovskiy A, Liao H-YM. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. 7464–7475.
24. Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018;8759–8768.
25. Zhou X, Wang D, Philipp K. Objects as points. arXiv preprint. 2019. arXiv:1904.07850.
26. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint. 2020. arXiv:2010.04159.
27. Ge Z, Liu S, Wang F, Li Z, Sun J. Yolox: Exceeding yolo series in 2021. arXiv preprint . 2021. arXiv:2107.08430.

Publisher's Note