Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/100/2025.20928

YOLO-ACD: Robust Lightweight Detector for Ancient Character Detection

Yucheng Song[1], Chuang Zhang[1,a,*], Ming Wu[1]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876, China
a. zhangchuang@bupt.edu.cn
**corresponding author*

Abstract: Ancient character detection serves an important role in paleographic studies and archaeological work. Unlike modern characters or text, ancient character renders common text detection algorithms inapplicable due to its complex arrangement patterns and highly variable backgrounds. Current methods commonly adopt general object detection models, which often struggle with the unique challenges of ancient characters. In this paper, we propose a lightweight ancient character detection method based on YOLO detector called YOLO-ACD. Starting from an efficient YOLO structure, to address the defects and adhesion due to coarse scanning or rubbing of the characters, we propose a novel high-frequency channel attention (HFCA) module to guide model for precise saliency extraction of strokes. To eliminate the interference of noisy backgrounds on the character area, we integrate a context-aware spatial attention (CASA) module. Furthermore, in order to overcome the limitations posed by limited training samples, we propose the incorporation of spectral clustering results as extra confidence scores, with the objective of enhancing the recall rate of the predictions. Extensive experiments on two ancient character datasets, including Chinese oracle bone script and Egyptian hieroglyphs, demonstrate that our method outperforms both mainstream and similar approaches. Compared to baseline model, the mAP is improved by 9.8-12.5%. The code of the proposed method will be available at GitHub.

Keywords: Ancient character detection, object detection, high-frequency channel attention, context-aware spatial attention, spectral clustering.

1. Introduction

As artificial intelligence (AI) drives revolutionary transformations across industries, it also opens up new possibilities for paleographic studies and archaeological research [1]–[3]. In particular, the integration of AI into paleography has transformed ancient character detection, enabling the preservation of cultural heritage and the interpretation of historical texts with unparalleled efficiency and accuracy. As an upstream task for ancient character recognition or interpretation, ancient character detection (ACD) plays a crucial role in identifying and localizing characters within complex backgrounds. Traditional methods rely on expert manual analysis, a labor-intensive process that is not only time-consuming but also prone to human error and subjectivity. In contrast, data-driven techniques enable fast detection of characters from eroded or fragmented artifacts like stone tablets or bamboo slips. However, these automated object detection algorithms struggle with ancient

DOI: 10.54254/2755-2721/100/2025.20928

characters due to their incomplete forms, fragmented strokes, and the scarcity of training samples, making feature learning challenging.

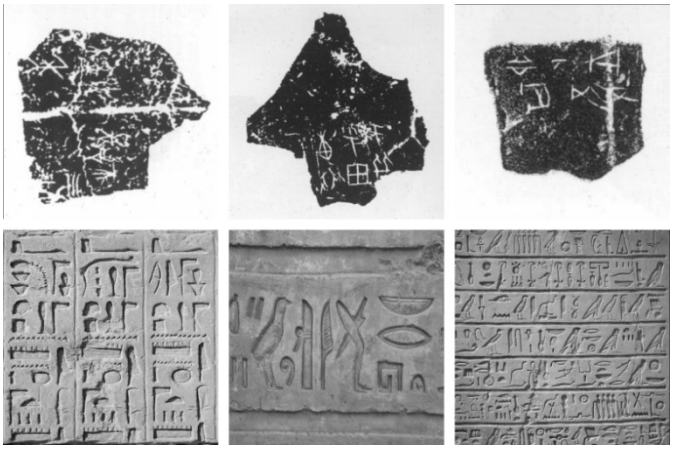

Figure 1: Sample images of two ancient characters. Top: Chinese oracle bone scripts. Bottom: Egyptian hieroglyphs.It is observable that the character arrangement is random and the forms are highly variable.

Since deep learning-based object detection models have demonstrated remarkable performance in various object detection tasks, many researchers have attempted to apply these models to ancient character detection [4]–[7]. Due to the passage of time, ancient characters such as Chinese oracle bone script and Egyptian hieroglyphs often exhibit incomplete forms, missing strokes, and significant erosion, as illustrated in Figure 1. For some extreme cases, even experts find it difficult to quickly and accurately locate the characters. Considering that archaeological work is predominantly conducted in outdoor environments, the development of an efficient object detector becomes essential to accommodate computational resource constrained devices and ensure real-time detection results. These obstacles not only highlight the significance of developing robust object detection algorithms specifically designed for ancient characters but also present significant challenges in the design of such algorithms.

To address the aforementioned issues, we proposed a robust lightweight object detector for ACD task, termed YOLO-ACD. A novel module named High-Frequency Channel Attention (HFCA) has been proposed, utilizing high-frequency channels as guidance to enhance the feature representation capabilities of the backbone network. This design aims to address the challenges posed by the incomplete and adhesion of ancient characters and the difficulty in extracting their saliency features. To effectively differentiate interested regions from noisy backgrounds, the Context-Aware Spatial Attention (CASA) module was introduced to enhance the model's attention toward character regions. Moreover, to mitigate the issues posed by limited data on deep neural networks, we incorporated a spectral clustering-based confidence optimization algorithm, termed SC-Score, into the post-processing phase.

Our main contributions are as follows:

- We propose a novel lightweight object detector, YOLO-ACD, specifically designed for ancient character detection.
- We propose the HFCA module to guide the model in extracting precise saliency features of ancient characters, introduce the CASA module to suppress noisy backgrounds and focus the model's attention on character regions, and further develop the SC-Score algorithm to enhance prediction recall rates.

DOI: 10.54254/2755-2721/100/2025.20928

- Extensive experiments on two ancient character datasets, including Chinese oracle bone script and Egyptian hieroglyphs, demonstrate that our method outperforms mainstream and similar approaches.

2. Related work

2.1. Object detection

As a fundamental task in computer vision, object detection has been extensively studied in recent years. Two-stage detectors, such as Faster R-CNN [4] and Cascade R-CNN [5], have achieved remarkable performance in various object detection benchmarks. One-stage detectors, such as YOLO [6] and SSD [7], have gained popularity due to their simplicity and efficiency, latest versions of YOLO [8]–[10] have achieved state-of-the-art performance in object detection tasks. While these detectors demonstrate impressive performance on natural images, they often struggle with ancient characters due to their unique characteristics, such as incomplete forms, fragmented strokes, and highly variable backgrounds. These challenges necessitate the development of specialized object detectors for the ACD task.

2.2. Character detection

Character detection serves as an upstream task for various text-related applications, such as optical character recognition (OCR) and ancient character recognition. To address the common challenges in this task, numerous research efforts have been proposed. Baek et al. [13] proposed a character region awareness method based on VGG-16 [14] for detecting characters in natural images. Wang et al. [15] proposed an oracle bone script detection method based on YOLOv4 [16] and a novel data augmentation strategy. Li et al. [17] proposed a lightweight character detection method based on YOLOv7-Tiny [11] for detecting Chinese oracle bone scripts. Zhao et al. [18] used the standard mAP evaluation metric for testing mainstream object detection models using the oracle bone script dataset, the results indicate that this task remains highly challenging. While numerous character detection methods have been proposed, few object detection algorithms have been specifically designed to address the unique characteristics of ancient character detection, which underscores the significance of our research.

3. Methodology

3.1. YOLO-ACD

Based on a basic lightweight YOLO structure (YOLOv8s [19]), we propose a novel YOLO-ACD model to address the few-shot ancient character detection problem. The overall architecture is shown in Figure 2. The YOLO-ACD adopts a C2f-Darknet backbone (three-stage) for multi-scale feature extraction. To adapt this basic feature extraction pipeline to practical character detection task, we propose a set of High-frequency Channel Attention (HFCA) modules at the output of the backbone to

DOI: 10.54254/2755-2721/100/2025.20928

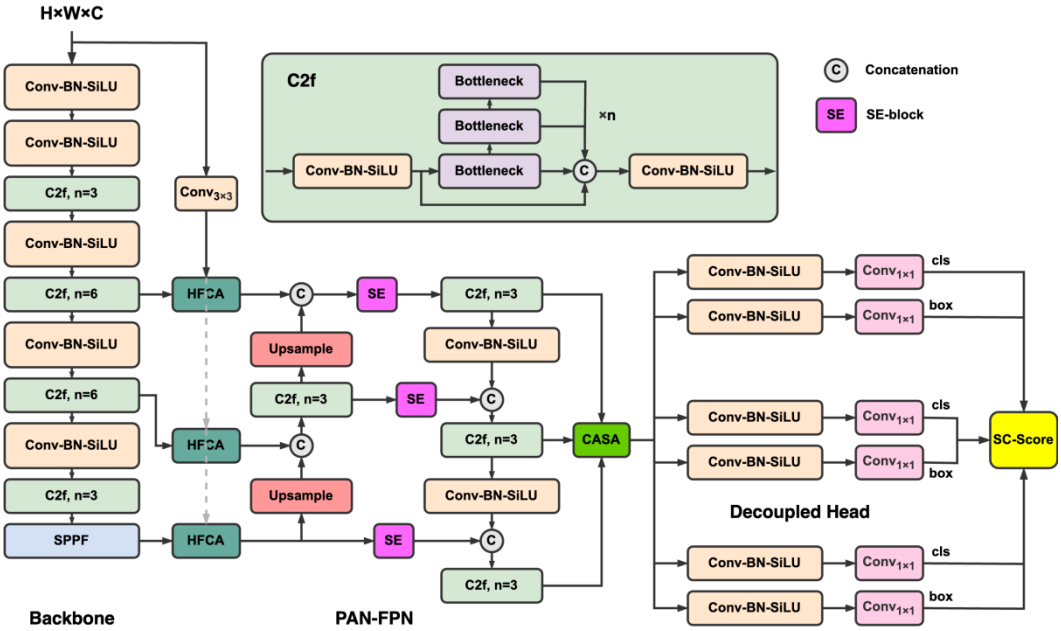

Figure 2: The overall architecture of the proposed YOLO-ACD. The basic YOLO structure is integrated with novel High-frequency Channel Attention (HFCA) modules and decoupled head with Diffusive-ODE structure to enhance the detection performance.

enhance the character-related saliency features. We employ a three-level PAN-FPN neck for feature fusion, integrating SE-blocks [20] into its lateral connections to enhance salient features. A Context-Aware Spatial Attention (CASA) module is connected at the output of the neck for emphasizing regions that likely belong to characters, while suppressing irrelevant backgrounds. Lightweight network structures often struggle to learn robust features effectively due to limited training samples, leading to suboptimal prediction performance. To address this limitation, we introduced a post-processing algorithm based on spectral clustering, termed SC-Score, to enhance the model's recall. We will focus on explaining HFCA, CASA, and SC-Score in the following sections.

3.2. High-frequency Channel Attention

The structure of the High-frequency Channel Attention (HFCA) module is shown in Figure 3. Given that most ancient character rubbings are single-channel grayscale images, a 3×3 convolutional layer is employed to extract their dual-channel high-frequency features. These features are then concatenated with the original image and dimensionally adjusted to form the final high-frequency representation I_{hf} . This process can be expressed as:

$$I_{hf} = concat\left(Conv_{3\times3}(I), \zeta(I)\right) \tag{1}$$

where I represents original image. ζ refers to downsampling using an average pooling layer. $Conv_{3\times3}$ denotes a 3×3 convolutional layer. concat denotes the concatenation operation. To guide the features extracted by the backbone using high-frequency information, we employed a non-local [21] cross attention mechanism. To match the dimensions of the feature maps at different stages of the backbone, we adjusted I_{hf} through 1×1 convolution φ and resize operation:

$$I_{hf}^{i} = bilinear\left(\phi\left(I_{hf}\right)\right) \tag{2}$$

where bilinear refers to the bilinear interpolation operation.

DOI: 10.54254/2755-2721/100/2025.20928

The non-local cross-attention mechanism is calculated as:

$$X^{'} = LN\left(X_i, \mu\left(d_i g(X_i)\right)\right), where\ d_i = \frac{\theta(X_i)\phi\left(I_{hf}^{i}\right)^{\top}}{N} \tag{3}$$

where X_i is the feature at the i-th stage of the backbone. g, θ, φ, and μ are 1×1 convolutional layers. N is the normalization factor with the shape of $H_i \times W_i$. d_i is the similarity matrix between X_i and I_{hf}^{i}. Different from the original non-local self-attention mechanism, the HFCA module utilizes high frequency information to guide the original feature in a cross-attention manner. The HFCA also utilizes layer normalization LN rather than summation to enhance the stability of the training process. The output of the HFCA module is calculated as:

$$X = X^{'} + CA\left(X^{'}\right) \tag{4}$$

where CA denotes the channel attention mechanism [22].

3.3. Context-Aware Spatial Attention

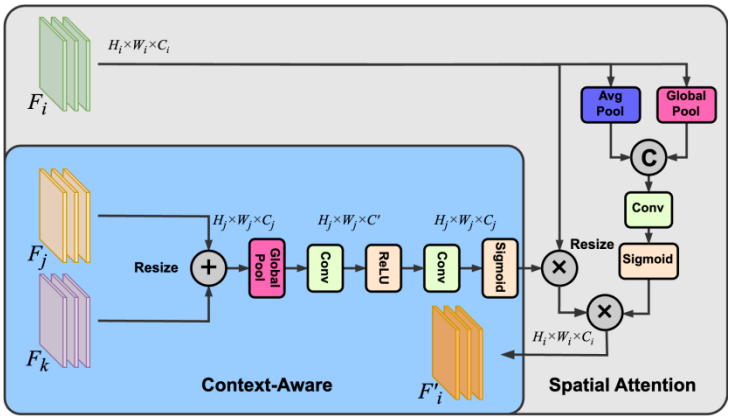

Figure 4: The structure of the Context-Aware Spatial Attention (CASA) module.

The concept of Context-Aware Spatial Attention (CASA) has been explored in various research studies [23], [24]. In our implementation, the CASA module is designed to refine neck features by incorporating multi-scale contextual information and applying spatial attention to highlight regions of interest. As shown in Figure 4, the CASA module consists of two parts: context-aware and spatial attention. For each feature level F_i from the PAN-FPN with shape $H_i \times W_i \times C_i$, the context-aware aggregates multi-scale features F_j and F_k from other levels. These features are fused and processed through global pooling and convolutional layers to generate the context-aware feature F_i :

$$F_i^c = F_i \cdot ConvLayers\left(\rho\left(F_j + F_k\right)\right) \tag{5}$$

where ρ denotes the global pooling operation. $ConvLayers$ refers to a series of convolutional layers with activation functions. The spatial attention is applied to the F_i :

$$F_i^s = sigmoid\left(Conv\left(\zeta(F_i), \rho(F_i)\right)\right) \tag{6}$$

where sigmoid denotes the sigmoid activation function. ζ denotes the average pooling operation. The final output of the CASA module is calculated as:

$$F_i^{'} = F_i^c \cdot F_i^s \tag{7}$$

DOI: 10.54254/2755-2721/100/2025.20928

3.4. SC-Score

The Spectral Clustering Score (SC-Score) is a post-processing algorithm that works alongside Non-Maximum Suppression (NMS) to improve the recall rate of the model's predictions. The core idea is to leverage a spectral clustering [25] to refine confidence scores of predicted boxes based on their similarity to a centroid derived from training set boxes in feature space. This ensures predictions align with the learned distribution of ancient character features, improving recall and filtering noisy predictions. We first extract the feature embeddings $f_{gt} = [b1, b2, \ldots, bn]$ of training set boxes using VGG-16 [14]. To calculate a clustering centroid c:

$$c = \frac{1}{n}\sum_{i=1}^{n} b_i \tag{8}$$

where bi is the feature embedding for the ith ground-truth box.During inference stage, we obtain the feature embeddings $f_{pred} = [p1, p2, \ldots, pm]$ of predicted boxes. Following the clustering algorithm described in [25], each predicted box is assigned to $k = 2$ cluster (relevant and irrelevant) based on its similarity to the precomputed centroid c derived from the training set.

To adjust the confidence scores of predicted boxes, if p_i is assigned to the relevant cluster, its confidence score is updated as:

$$sc_i = sc_i + \lambda \cdot exp\left(-\frac{\|p_i - c\|^2}{2\sigma^2}\right) \tag{9}$$

where α is a hyperparameter that controls the scaling factor. σ is a hyperparameter that controls the similarity threshold.

4. Experiments

4.1. Datasets and evaluation metrics

We evaluate the proposed YOLO-ACD on two ancient character public datasets: Chinese oracle bone script [26] and Egyptian hieroglyphs [27]. The Chinese oracle bone script dataset contains 3,066 images with 18,273 annotated characters, where 2,799 images are used for training and 267 images for testing. The Egyptian hieroglyphs dataset contains 105 images with 16,587 annotated characters, where 87 images are used for training and 18 images for testing. However the Egyptian hieroglyphs dataset is extended from 35 images by simple data augmentation, therefore the training samples are limited.

We report the Average Precision (AP) as the primary evaluation metric. We also report the recall rate to provide a comprehensive evaluation of the SC-Score. Besides, we report the Floating Point Operations (FLOPs) and the number of parameters (Params) to evaluate the model's efficiency.

4.2. Implementation details

All the experiments are conducted on a computer with an NVIDIA RTX 3090 GPU. The YOLO-ADC is implemented using PyTorch 1.13.1 and MMYOLO [19]. Other object detection models (except for YOLOv10 [28]) that are included in this section are implemented using MMYOLO and MMDetection [29] under the same environment.

For the training process, SGD optimizer is employed with a batch size of 12 and the initial learning rate set to 0.01. We train the YOLO-ACD for 100 epochs, other models are trained for different epochs according to their convergence speed. All experimental results are obtained at the same input resolution of 640×640. Mosaic augmentation [16] is applied for all experiments.

DOI: 10.54254/2755-2721/100/2025.20928

4.3. Comparisons

Table 1: Comparisons with state-of-the-art object detection models on Oracle bone/Egyptian datasets.

Models	AP(%)	AP50(%)	FLOPS	Params
YOLOv5s	0.386/0.189	0.771/0.319	8.1G	12.3M
YOLOXs	0.411/0.193	0.796/0.327	13.4G	8.9M
YOLOv8s	0.420/0.192	0.759/0.320	14.9G	11.4M
YOLOv8m	0.461/0.211	0.811/0.339	40.1G	26.1M
YOLOv10s	0.437/0.199	0.761/0.323	21.6G	7.2M
Faster R-CNN	0.403/0.184	0.788/0.340	63.2G	41.4M
Ours	0.464/0.218	0.807/0.348	21.5G	12.1M

We compare the proposed YOLO-ACD with several state-of-the-art object detection models, including YOLOv5, YOLOX [30], YOLOv8 [12], YOLOv10 [28], and Faster R-CNN [6] with ResNet-50 [31] backbone. The results are listed in Table 1. It can be observed that the proposed YOLO-ACD outperforms all other models in terms of AP, and has a competitive AP50 with YOLOv8m, while maintaining a relatively low FLOPs and Params. Compared to the latest YOLOv10s which has similar computational cost, YOLO-ACD outperforms by 6.2% and 9.5% in terms of AP on the two datasets, respectively, with only a small increase in FLOPs and parameters. Compare to two-stage Faster R-CNN, YOLO-ACD outperforms by 15.1% and 2.4% in terms of AP on the two datasets, respectively, with only around 30% of the FLOPs and parameters. As a result, the proposed YOLO-ACD demonstrates a superior trade-off between performance and efficiency compared to other mainstream object detection models.

4.4. Ablation Study

To evaluate the effectiveness of our proposed method, we use YOLOv8s as baseline and conduct ablation experiments on the proposed HFCA, CASA, and SC-Score, the

Table 2 : Ablation experiments of proposed innovations on Oracle bone/Egyptian datasets. Note the baseline model is YOLOv8s, which does not include SE-blocks in the neck.

Baseline	HFCA	CASA	SC-Score	AP(%)	AP50(%)	AR(%)	FLOPS	Params
√	×	×	×	0.420/0.192	0.759/0.320	0.560/0.223	14.891G	11.358M
√	√	×	×	0.435/0.202	0.771/0.325	0.576/0.233	21.501G	2.050M
√	×	√	×	0.427/0.194	0.762/0.321	0.566/0.226	14.894G	11.361M
√	×	×	√	0.417/0.195	0.746/0.323	0.563/0.229	14.891G	11.358M
√	√	√	×	0.461/0.209	0.811/0.338	0.586/0.240	21.503G	12.053M
√	√	√	√	0.461/0.216	0.810/0.344	0.583/0.248	21.503G	12.053M

results are shown in Table 2. It can be observed that the incorporation of HFCA and CASA modules significantly improves the performance of the baseline model from 0.420/0.192 to 0.461/0.209 in terms of AP, with only 6.612 GFLOPs and 0.695M Params increase. The results indicate that the HFCA and CASA modules effectively improve the model's performance by enhancing the feature extraction capabilities using attention mechanisms. As for the SC-Score, which brings a significant improvement in AR (Average Recall) on the Egyptian hieroglyphs dataset, demonstrating its effectiveness in enhancing the recall rate of the model's predictions under limited data scenario.

DOI: 10.54254/2755-2721/100/2025.20928

4.5. Qualitative Results

We provide qualitative results of the proposed YOLO-ACD from two aspects: attention maps and detection results. The attention maps from the middle neck layer in the proposed YOLO-ACD are shown in Figure 5. For each set of image in Figure 5, the first row shows the original image, the second row shows the attention map from the baseline model (YOLOv8s), and the third row shows the attention map from the YOLO-ACD. It can be observed that the proposed HFCA and CASA modules effectively guide the model to focus on the character regions, enhancing the feature representation capabilities of the backbone network. The detection results are shown in Figure 6, where the proposed YOLO-ACD demonstrates more robust performance than the baseline model in detecting ancient characters.

5. Conclusion

In this paper, we proposed a novel lightweight object detector, YOLO-ACD, specifically designed for ancient character detection. The proposed YOLO-ACD integrates the HFCA and CASA modules to enhance the feature representation capabilities of the feature extraction process and guide the model to focus on character regions. The SC-Score algorithm is introduced to enhance the recall rate of the model's predictions under limited data scenario. The proposed method achieves the state-of-the-art performance and a superior trade off between performance and efficiency compared to other object detection models, demonstrating its effectiveness and application potential in real-world ancient character detection.

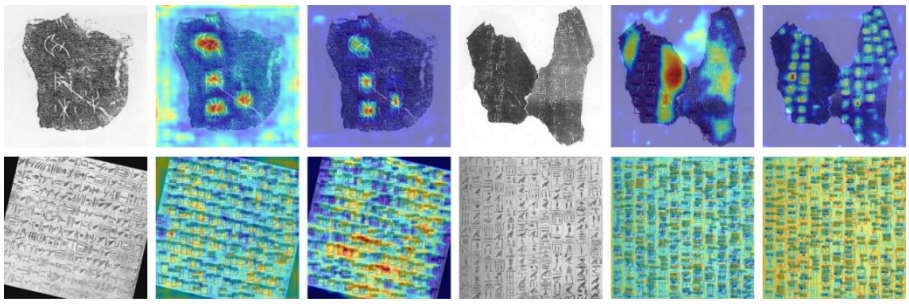

Figure 5: Attention maps from the neck layers (middle level) in the proposed YOLO-ACD.

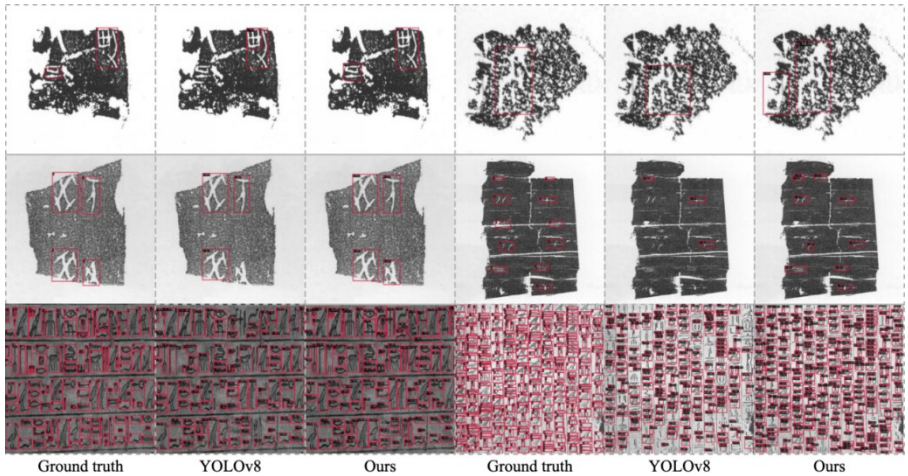

Figure 6: Detection results of the proposed YOLO-ACD.

DOI: 10.54254/2755-2721/100/2025.20928

References

[1] D. Chapinal-Heras and C. D´ıaz-Sanchez, "A review of ai applications in human sciences research," Digital Applications in Archaeology and Cultural Heritage, p. e00288, 2023.

[2] S. M. Griffin, "Epigraphy and paleography: bringing records from the distant past to the present," International Journal on Digital Libraries, vol. 24, no. 2, pp. 77–85, 2023.

[3] M. Tenzer, G. Pistilli, A. Bransden, and A. Shenfield, "Debating ai in archaeology: applications, implications, and ethical considerations," Internet Archaeology, no. 67, 2024.

[4] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on dronecaptured scenarios," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 2778–2788.

[5] M. Yan, S. Wang, and Z. Lu, "Improving end-to-end object detection by enhanced attention," in 2024 International Joint Conference on Neural Networks (IJCNN). IEEE, 2024, pp. 1–8.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 6, pp. 1137–1149, 2016.

[7] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6154–6162.

[8] J. Redmon, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition,2016.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands,October 11–14, 2016, Proceedings, Part I 14. Springer, 2016, pp. 21–37.

[10] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie et al., "Yolov6: A single-stage object detection framework for industrial applications," arXiv preprint arXiv:2209.02976, 2022.

[11] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 7464–7475.

[12] G. Jocher, J. Qiu, and A. Chaurasia, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[13] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9365–9374.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[15] N. Wang, Q. Sun, Q. Jiao, and J. Ma, "Oracle bone inscriptions detection in rubbings based on deep learning," in 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), vol. 9. IEEE, 2020, pp. 1671–1674.

[16] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[17] Y. Li, H. Chen, W. Zhang, and W. Sun, "Lightweight oracle bone character detection algorithm based on improved yolov7-tiny," in 2024 IEEE International Conference on Mechatronics and Automation (ICMA). IEEE, 2024, pp. 485–490.

[18] P. Zhao and Y. Liu, "Oracle bone inscriptions detection based on standard evaluation metric," in 2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML). IEEE, 2022, pp. 49–55.

[19] M. Contributors, "MMYOLO: OpenMMLab YOLO series toolbox and benchmark," https://github.com/open-mmlab/mmyolo, 2022.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[21] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.

[22] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

[23] H. Tang, X. Liu, K. Han, X. Xie, X. Chen, H. Qian, Y. Liu, S. Sun, and N. Bai, "Spatial context-aware self-attention model for multi-organ segmentation," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 939–949.

[24] J. Zhang, J. Ren, Q. Zhang, J. Liu, and X. Jiang, "Spatial context-aware object-attentional network for multi-label image classification," IEEE Transactions on Image Processing, vol. 32, pp. 3000–3012, 2023.

[25] U. Von Luxburg, "A tutorial on spectral clustering," Statistics and computing, vol. 17, pp. 395–416, 2007.

DOI: 10.54254/2755-2721/100/2025.20928

[26] guangdong, "Trr dataset," https://universe.roboflow.com/guangdong/trrnq2vd , apr 2024, visited on 2024-12-22. [Online]. Available:

[27] https://universe.roboflow.com/guangdong/trr-nq2vd image segmentation task, "segmentation dataset," https://universe.roboflow.com/image-segmentation-task/segmentation cpobd , apr 2023, visited on 2024-12-22. [Online]. Available: https://universe.roboflow.com/image-segmentation-task/segmentation cpobd

[28] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," arXiv preprint arXiv:2405.14458, 2024.

[29] MMDetection Contributors, "OpenMMLab Detection Toolbox and Benchmark," Aug. 2018. [Online]. Available: https://github.com/openmmlab/mmdetection

[30] Z. Ge, "Yolox: Exceeding yolo series in 2021," arXiv preprint arXiv:2107.08430, 2021.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.