

RESEARCH

Open Access



Detecting oracle bone inscriptions via pseudo-category labels

Xinran Fu¹, Rixin Zhou¹, Xi Yang^{1,3,4*} and Chuntao Li^{2,4*}

Abstract

Oracle bone inscriptions (OBIs) are the earliest Chinese characters and reserve abundant historical information. OBIs are detected by locating their positions in digital images. This has been a foundational task in modern archeological studies. Due to the development of deep neural networks (DNNs) in computer vision, detecting OBIs can be implemented by a more concise method instead of designing complicated hand-crafted features. However, existing models cannot perform well when noise areas are similar to some inscriptions. In this work, we present a simple but effective pseudo-label-based architecture for OBIs detection. Different from previous approaches, our method performs OBIs detection with the employment of information from multilabel annotations rather than single location information. We append a plug-and-play module that predicts the pseudo-label of an inscription after the backbone network for learning the particular structure prior to each inscription and brings this information to the backbone network by means of feature fusion. We make remarkable improvements on different backbone networks when using the proposed method on an OBIs detection dataset. The quantitative and qualitative results show that the proposed model can detect OBIs well and is an effective tool for assisting in the discovery and recognition of ancient writing.

Keywords Oracle bone inscription, Deep learning, pseudo-label, Character detection

Introduction

Oracle bone inscriptions (OBIs) are the precursors of modern Chinese characters, recording a vast amount of information about ancestors who lived 3600 years ago. They represent the most significant historical relics for understanding the development of ancient society during a specific dynasty and have become a foundational field of research in modern archeology. The detection of OBIs is an important part of ancient writing

research. Detecting OBIs is to locate the position of OBIs in digital images. Given an image containing numerous inscriptions, detection algorithms provide an explicit representation of each inscription's location, typically in the form of bounding box coordinates. With the development of computer vision technology, more research for recognizing OBIs has flourished over the past decade, with a major spectrum of works using traditional visual feature-extraction modules such as the Hough transform and handmade Gabor kernel [1–4], and some other works treat an inscription as an undirected graph composed of intersections and lines [5] and achieve this goal by topological and graph-based methods [6–9]. With the development of deep neural networks (DNNs) in the computer vision domain, especially in object detection tasks and optical character recognition, the detection of OBIs can be implemented by a more concise method instead of designing complicated hand-crafted features. Recent advances in OBIs detection follow the paradigm of

*Correspondence:

Xi Yang

yangxi21@jlu.edu.cn

Chuntao Li

850954371@qq.com

¹ School of Artificial Intelligence, Jilin University, Changchun 130000, Jilin, China

² School of Archaeology, Jilin University, Changchun 130000, Jilin, China

³ Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MoE, Changchun 130000, Jilin, China

⁴ Key Laboratory of Ancient Chinese Script, Culture Relics and Artificial Intelligence, Changchun 130000, Jilin, China



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

object detection, namely treating inscriptions carved in the bone as an object in a photo and utilizing YOLO-like convolutional neural networks to address the task [10, 11]. These regressive models are sensitive to simple changes in the complicated heuristic processing such as anchor generation [12] and non-maximum suppression (NMS) [13] and suffer from the unbalanced of samples between different categories [14]. Another method to detect OBIs follows the schema of text detection, which regards all inscriptions as the same semantic area and generates a mask for segmenting inscriptions from one image [5, 15]. A mix-up augmentation strategy that leverages information from both majority and minority classes was proposed to alleviate the imbalanced data distribution problem [16]. However, the current research on detection tasks is still relatively limited and the performance of existing models will decrease when encountering a mass of noise areas resembling certain inscriptions, which is a normal case in digital OBI images.

In this paper, we contend that two unique properties of existing OBIs dataset need to be considered. First, as the origin of Chinese characters, OBIs have the same square-shaped font but lack legible and semantic relations which formed over a long-term evolution between each other. Another property is that the number of recorded OBIs is relatively smaller than that of modern Chinese characters, so utilizing the prior information of individual inscriptions is feasible and worth exploring. Mainly taking these two properties into account, we propose a simple framework that utilizes pseudo-category labels and a supervised learning approach to capture the distinctive features of each inscription. The proposed models have fewer parameters compared to the previous one and also dispense with complex manual design processes but achieve competitive results. The incorporation of pseudo-category labels empowers the segmentation-based models with the ability to accurately identify the inscription area thus enhancing the model's precision by 1.9% and the recall by 1.8%. Consequently, this enables the model to effectively fulfill the task of detecting OBIs.

Related works

Treating OBIs detection as a text detection task is rational. Many works that rely on deep networks have been proposed to address text detection problems. A prevalent approach in recent research involves incorporating additional learnable information into the network. This supplementary information, representing various patterns of latent features, can be readily acquired and proves advantageous for the primary detection task. We review a part of the work in the following section.

Regression-based methods

There are a series of models that directly regress the boundary region of the text instances. The textBoxes series [17–19] applied a fully convolutional network to detect multi-oriented text via anchor-based quadrilateral regression. EAST [20] is an anchor-free method, that applied different predicting heads for multi-oriented contour regression. There is a score map for determining the area of a text instance and a regression of geometry parameters for determining orientation. To locate curved texts better, recent works focused both on network structure and the representation of arbitrary-shaped text regions. LOMO [21] introduced three modules to refine the text localization of a direct regression result iteratively. The corner and category information were learned by their proposed refinement module. TextRay [22] formulates the text contours in the polar coordinate system rather than widely used Cartesian coordinates and learns the geometric parameters of the text boundary. TESTR [23] follows DETR [24]'s bipartite matching design and trains the whole model using the labels of control point coordinate, query classification, and character recognition. ABCNet series [25, 26] and FCENet [27] calculate another parameterized representation of the text contour, specifically, the former used cubic Bezier curves to parameterize curved texts and predicted their control points with proposed BezierAlign module. The latter proposed the Fourier contour embedding (FCE) method, which is a new representation of text contour based on Fourier transformation theory. In addition to predicting the pixelwise classification map, this model regresses Fourier signature vectors for generating curves with arbitrary shapes.

OBIs detection with regression-based methods followed a routine and simple paradigm. Frequently used convolutional backbones with classification and regression head were utilized to directly predict bounding box priors and post-processing algorithms subsequently restrained the number of outputs to generate coordinates for each image. Previous works have tested many popular networks and obtained mediocre quantitative results [1, 11]. However, it is worth noting that the representation is much simpler for OBIs than for scene texts and bounding box coordinates are accurate enough for an inscription's location.

Segmentation-based methods

Methods based on segmentation algorithms implicitly encode text regions with per-pixel masks and then transform the mask components into bounding boxes by some post-processing algorithms. DBNet [28] replaced the binarization operation with an approximate differentiable function and predicted masks for

both the text region and text border via the proposed adaptive scale fusion. CRAFT [29] modeled the relationships between different individual characters by affinity boxes and is supervised by both region map and affinity map therefore it can focus only on intra and inter-character information. The CRAFT algorithm predicted character-level bounding boxes for input images which is different from other works. Some recent works [30, 31] addressed this problem by cooperating with an attention mechanism [32], and using a convolutional network to extract hidden features that need to optimize as signals of foreground text, then a subsequent transformer architecture served as a post-processing module to get final parameterization of each text region.

The segmentation-based method for detecting OBIs adopted the ideas from CRAFT because the inscriptions in an image are independent and lack linguistic meaning. Liu et al. [5] followed the label generation mode as CRAFT and combined the hourglass network [33] with the progressive post-processing algorithm as their main design. The proposed segmentation-based detector avoided the demand for anchors and could learn character spatial regions. Lin et al. [34] also designed an oracle radical extraction and recognition framework based on the segmentation-based method.

Dataset

We use an open-access OBIs dataset provided by the Key Laboratory of the Ministry of Education for Oracle Information Processing, Anyang Normal University [11]. This

dataset comprises 9,154 images meticulously annotated with bounding box details, primarily emphasizing the detection task. Following the structure of certain optical character recognition datasets introduced in recent years [35–37], which maintain a ratio of 4:1:5 between the training set, validation set, and testing set respectively, we partitioned all images into three subsets. Specifically, 3688 images were allocated for training, 900 images for validation, and 4566 images for evaluating the performance of models. The partition mode leads to a relatively strict ratio between training samples and testing samples and requires a model to have greater capabilities. Specifically, we just divided the dataset by the order of file names without considering other properties of images. This dataset is used to train the proposed models and accomplish the task of detecting OBIs. In addition, we chose another dataset [2] which is made of different separate OBIs to obtain pseudo-category information. This dataset includes images of a single inscription which are clipped manually from rubbing images and these images of a single inscription were divided into 125 categories. We use this dataset to train a 34-layer ResNet model and obtain the pseudo-labels. We also test this ResNet, which can achieve 98% classification accuracy, guaranteeing label availability.

Method

Model architecture

First, we present the complete pipeline of our proposed method, which is illustrated in Fig. 1. The models we used in this paper are general and composed of a

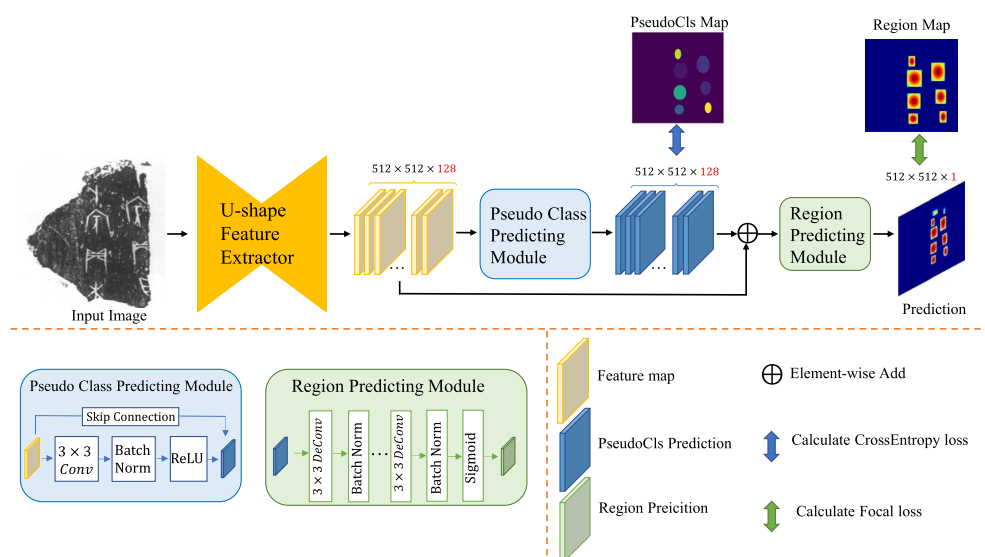


Fig. 1 Illustration of our proposed method comprises a U-shape Feature Extractor, a Pseudo Category Module, and a Region Predicting Module. The output of the Pseudo Label Module will be merged with the initial feature map, while the Region Predicting Module generates a pixel-level mask for inscriptions

downsampling module $\mathbf{D}(\cdot)$ and an upsampling module $\mathbf{U}(\cdot)$, which serve as feature extractors in our work. Given an image $I \in \mathbb{R}^{H \times W \times 3}$, a model encodes it to a high-dimension feature map $\mathbf{F} = \mathbf{U}(\mathbf{D}(I)) \in \mathbb{R}^{H \times W \times d}$. After obtaining the feature map, we append two prediction modules for our purpose, that is, we regard these models as latent state extractors rather than end-to-end networks, and the outputs of these backbone networks are high dimensional codes of each image. First, a pseudo-label predicted branch $f_p(\cdot)$ which is composed of standard 3×3 convolutional layers, batch-norm layers, nonlinearity activation layers, and residual connections propagates forward it to a prediction mask $\mathbf{M}_{pc} = f_p(\mathbf{F}) \in \mathbb{R}^{H \times W \times d}$ that each vector of a location represents the probability of an individual character. This mask has two effects. We use it to supervise the model for understanding the particular structure of inscriptions and we add it to the feature map obtained by backbones in an element-wise manner as a fusion to introduce the useful structure information to the original feature, namely $\mathbf{F}' = \mathbf{F} + \mathbf{M}_{pc}$. Then a region predicted branch $f_r(\cdot)$ propagates forward this fusion feature \mathbf{F}' to the final outputs $\mathbf{M}_r = f_r(\mathbf{F}') \in \mathbb{R}^{H \times W}$

which is the final prediction of the model and has the same shape as the input image. In this predicted mask each magnitude of a pixel represents the probability of the inscription area. The region predicting branch has a relatively simple design with sequential transposed convolution layers, batch-norm layers, and a Sigmoid activation function for numerical stability. Finally, the region

map and pseudo-label map generated beforehand will be used to optimize these two outputs.

Specifically, we choose three encoder-decoder structure neural networks as our backbones: UNet, U^2 Net and Hourglass-Net [33, 38, 39]. This kind of neural network is composed of a module that downsamples an input image to a low-resolution feature map as a latent representation of the relevant image and a module whose function is to upsample the feature map to the output mask which has the same shape as the input image and is used for supervision.

Then we introduce them briefly and the sketches of their design are shown in Fig. 2. UNet used 3×3 convolutional layers, ReLU layers, and max pooling layers successively to extract the features of an image and used sequential interpolation layers to obtain the segmentation mask from this downsampled feature map. UNet was first proposed to address the problem of biomedical image segmentation, this method has shown great ability in pixel-level classification tasks and yielded good results. U^2 Net replaced simple convolutional layers with an innovatively proposed Residual U-shape (RSU) block in both the encoder and decoder layers. The RSU block is a U-Net-like module and is stacked sequentially to build the whole cascaded model. U^2 Net was used in salient object detection which involves discovering the main subject in a given image. Hourglass-Net settled the problem of human pose estimation and was named for its use of using conv-deconv module whose shape is similar to an hourglass. The hourglass module can be stacked recursively to constitute a network and is set up as follows: convolutional layers and max pooling layers

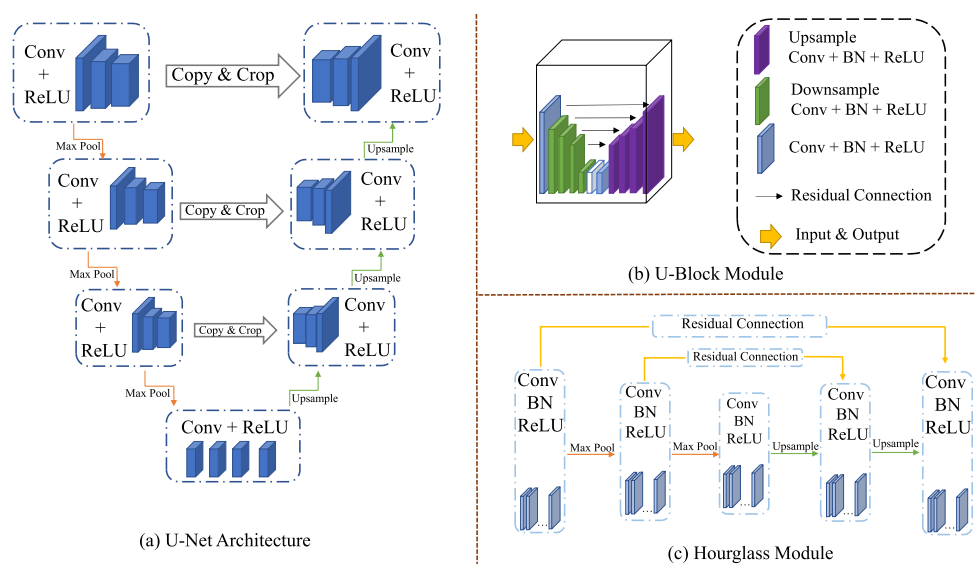


Fig. 2 The illustration of the major components of each backbone. **a** displays the entire architecture of U-Net, **b** displays the Residual U-Block that is proposed in U^2 Net. **c** displays the hourglass module which is a subassembly of Hourglass-Net

are used to process features down to a very low resolution. Then the following subassemblies implement upsampling and combining operations of features across scales.

Label preparation

Given an image of the oracle bone, annotations of the inscription region are described by a set of coordinates of two vertices and need to be converted into mask-type supervision. In this study, we create a region map to depict the location of each inscription and a pseudo-category map to incorporate individual information regarding different inscriptions into the model. Both the region and pseudo-category maps are 2D arrays. Figure 3 shows the process of generating two maps for an image.

In the process of generating a region map, we employ an approximate probability density function of a two-dimensional Gaussian distribution for each bounding box. This allows us to calculate the probability value of pixels within the bounding box, which is then utilized to generate a mask with the same shape as the image. The value of each pixel in the mask is calculated by the following formula:

$$G(x, y) = \frac{A}{2\pi\Sigma^2} e^{-\frac{x^2+y^2}{2\Sigma^2}}, \quad (1)$$

where $x \in [x_1, x_2], y \in [y_1, y_2]$ and x_1, x_2 are the upper-left and lower-right abscissa of a bounding box, and y_1, y_2 are the upper-left and lower-right ordinates of a bounding box. All coordinates should be normalized to $[-1, 1]$ so that the mean of the Gaussian distribution is in the center of the region mask. A is a coefficient that expands the value for each pixel. This operation simulates a situation

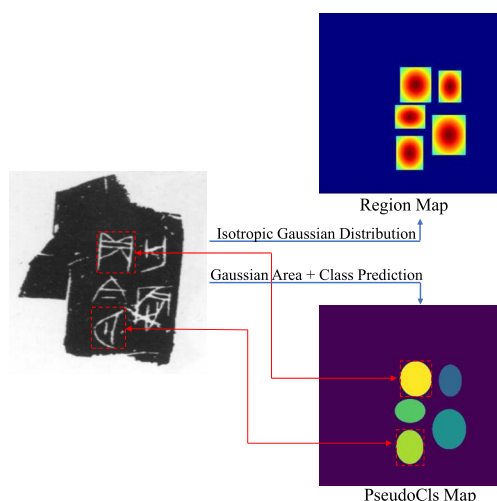


Fig. 3 Process of label generation for region map and pseudo-label map. In Region Map, the color of red means the center of an inscription, and the color of cold tune means the boundary. In PseudoCls Map, each color represents a category

in which the center of the area has a high value and the boundary has a relatively low value, which means that the center of an inscription is more important to focus on and that the peripheral area has less importance. At the top of Fig. 3, the color of gradually darker red means the center of an inscription (high value pixels) and the gradual cold-tuning color indicates the boundary of an area (low value pixels)

In addition to generating a region map, we use a category map to supervise an inscription's individuality information. First, we use a lite-scale ResNet [40] $f(\cdot)$ which was trained previously in the classification dataset mentioned before [2] to obtain the category of each inscription. The process is as follows:

$$c = f(I), c \in \mathbb{N}, I \in \mathbb{R}^{H \times W \times 3}, \quad (2)$$

where I is an image of an inscription that was cropped from a rubbing and c is its predicted category. Then, for each rubbing image, we crop the inscriptions by using their coordinates and feed them into ResNet, which is trained on the classification dataset mentioned before to obtain their category labels. After obtaining these category labels, we replace the probability value of each pixel whose value is greater than a threshold with the predicted category to obtain the category map. At the bottom of Fig. 3, different areas of the inscription have different colors since they are different inscriptions. This map is called a pseudo-category map because all categories are obtained by the output of a pre-trained neural network rather than labeling by archeological workers. The category map provides more specific details of an area and introduces supervisory information on the shape and construction of different inscriptions.

Loss function

The total loss function can be expressed as a weighted sum of the value for the region map and the value for the pseudo-category map:

$$\mathcal{L} = \lambda_1 \mathcal{L}_r + \lambda_2 \mathcal{L}_{pc}, \quad (3)$$

where L_r is the loss for the region map and L_{pc} is the loss for the pseudo-label map. λ_1 and λ_2 are set to different values because the region branch is the main task and the pseudo-branch should have a relatively weak effect on all the parameters. This task is essentially a binary classification for each pixel, hence, we apply focal loss, a balanced binary cross entropy loss [41] for the region map to overcome the imbalance of the number of positives and negatives:

$$\mathcal{L}_r = -\alpha(1 - p_t)^\gamma \log(p_t) \quad (4)$$

where p_t is an indicator function value that equals the model's estimated probability for the positives and equals one minus probability for negatives, γ and α are proposed focusing hyperparameters. As for the pseudo-category map, we choose cross-entropy loss to measure the prediction since the objective is to classify pixels into different categories. The formula is as follows:

$$\mathcal{L}_{pc} = \sum_{c=1}^M y_c \log(p_c), \quad (5)$$

where M is the number of classes, p_c is the probability for each class that is calculated by the model and y_c is the target of this prediction.

Inference

In the inference period, we only use the final output of the region map to generate inscription bounding boxes since the information of the pseudo-label has been integrated into the whole model. The box formation process consists of two steps: The region map is first binarized with a constant threshold to generate a binary map whose pixels only have two values of 1 or 0. Then connected areas are obtained from the binary map by a contour searching algorithm [42]. These predicted connected areas are formed by the coordinates of vertices and encircle the inscriptions that need to be detected.

Experiments

Implementation details

Our implementation is based on Python and PyTorch. In the training process, all the network parameters are initialized randomly and all the networks are trained from scratch instead of pretraining. For UNet and U²Net, we use their original models and we stack three hourglass modules as Hourglass-Net. To keep the scale of the parameters close, we change the dimensions of the middle layers. We also modify the last output dimensions of these models for the sake of elementwise fusion. We use one NVIDIA A40 GPU to train each model with a batch size of 8. During training, the input image will be augmented by color jitter and directly resized to resolution of 512×512 and the settings are the same for prediction and evaluation. Adam optimizer [43] is used to train all networks with a learning rate of $3e-4$ in the beginning and the learning rate is modified by an exponential warm-up function in each iteration. For the loss function, the hyperparameters α and γ in the focal loss are set to 0.25 and 2.0, the weights of λ_1 and λ_2 in the total loss are set to 1 and 0.1, respectively.

Evaluation metrics

We first introduce the concept of the intersection over union (IoU). IoU describes the extent of overlap of arbitrary shapes that can be a measure of the similarity. For two plane figures, the IoU can be calculated by:

$$IoU = \frac{A_{inter}}{A_{union}} \in [0, 1] \quad (6)$$

where A_{inter} is the area where two figures intersect and A_{union} is the sum of the areas of two figures. It is obvious that the greater the region of overlap is, the greater the magnitude of the IoU and the more similar the two figures are.

To evaluate the matching quality between the predicted bounding boxes and the ground truth, we use three comprehensive evaluation indicators, the *Precision*(P), *Recall*(R) and *F₁-score*(F_1). The calculation formulas are as follows:

$$P = \frac{TP}{TP + FP}, \quad (7)$$

$$R = \frac{TP}{TP + FN}, \quad (8)$$

$$F_1 = \frac{2 \times P \times R}{P + R}, \quad (9)$$

where TP , FP and FN are abbreviations for three terms: true positive, false positive, and false negative. The demarcation of samples depends on the IoU. Specifically, the bounding box is assigned to a true positive sample if the IoU value between a bounding box and a ground truth is larger than a threshold that is commonly set to 0.5 for impartiality, in the opposite case we assign a bounding box to a false positive sample and if a ground truth is not matched by any bounding box we assign it to a false negative sample.

Results and discussion

Quantitative results

To verify the effectiveness of our proposed model for modeling inscription instances, we compare the performance of the detection results on the newly divided test set. Figure 4 shows the sizes of the different models and the results of our proposed methods alongside their corresponding baselines in the training set, validation set, and test set. Our proposed method has made obvious progress. As shown in Fig. 4a, the scale of each model is limited to 10 MB which means a relatively fair comparison and is for the purpose of being lightweight. As

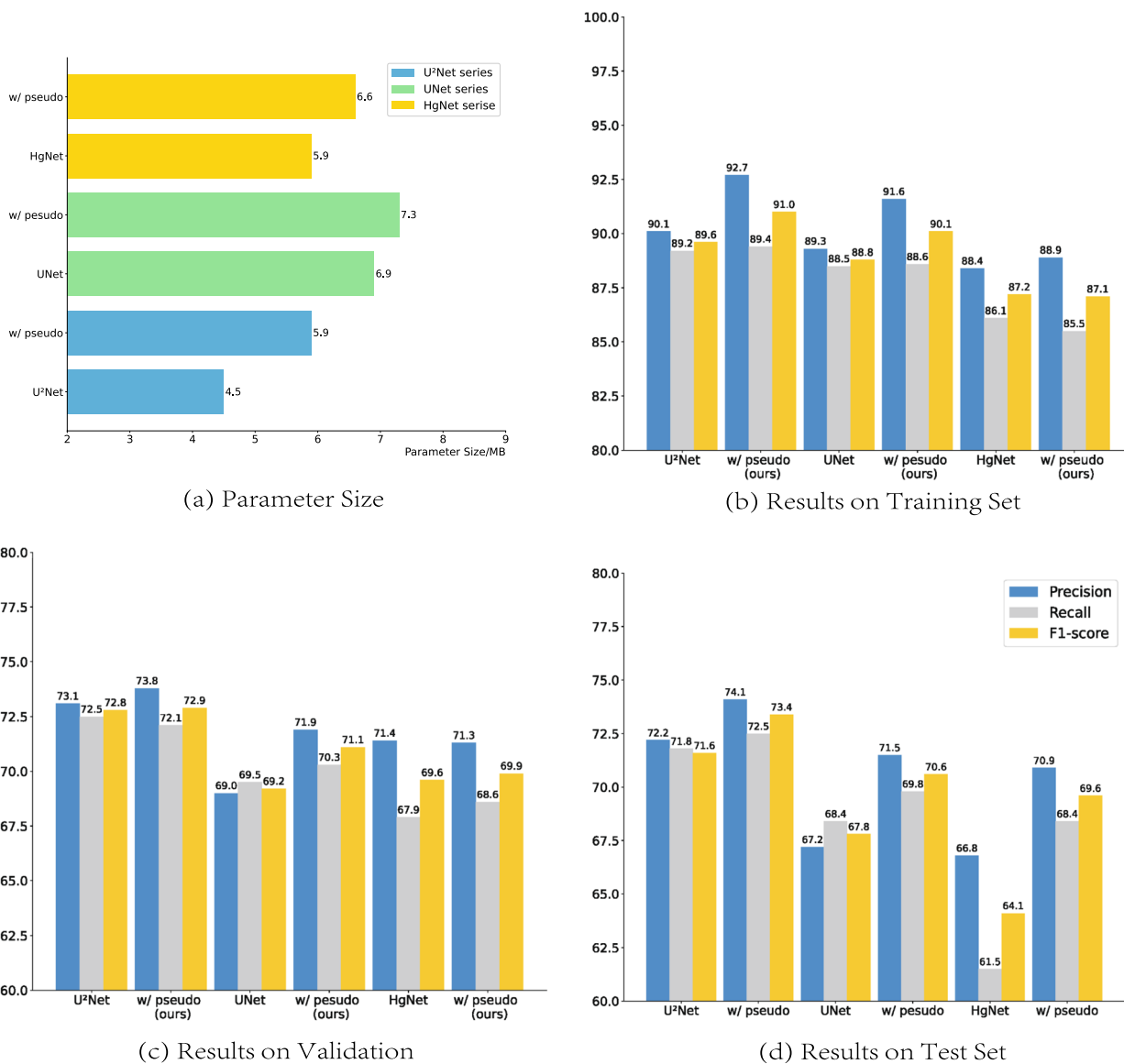


Fig. 4 **a** Shows the parameter size of different models, each color represents a series of backbone. **b–d** show the comparison of quantitative results between different models in the training set, validation set, and test set respectively, each color represents a metric

shown in Fig. 4d UNet equipped with this classification branch achieves a 4% improvement in precision and a 1.4% improvement in recall. Assisted by the classification branch, Hourglass-Net increases the magnitude of precision by 4.1% and increases the magnitude of recall by 7.1%. Compared with the U²Net baseline, the additional branch counterpart yields a better result in which the precision is improved by 1.9% and the recall is improved by 1.8% with only a few extra parameters. Note that the U²Net, which is a competitive backbone network for detecting OBIs, has the smallest number of parameters but obtains the best results. Based on the experimental

results, the newly introduced subtask that predicts the probable category of inscriptions endows unique signals of each inscription to models and renders superior results.

We also compare the results of our models and some regression models on the training, validation, and test sets, and show them in Table 1. According to the results, segmentation models equipped with our module achieve a higher index than RetinaNet which is trained from scratch. Compared with YOLO, which is pre-trained on the Large ImageNet dataset, our proposed methods have higher precision than YOLO. We also compared

Table 1 Comparative result between regression models and segmentation models with our proposed module

	Precision			Recall			F1-Score		
	Train	Val	Test	Train	Val	Test	Train	Val	Test
RetinaNet [41]	60.5	57.3	56.1	62.3	56.4	55.4	61.4	56.9	55.4
YOLO [44]	47.6	44.1	46.3	47.2	45.6	46.9	47.4	44.8	46.6
YOLO(pre-train)	85.4	73.2	73.7	82.4	72.8	74.1	83.8	72.7	73.9
Hourglass(ours)	88.9	71.3	70.9	85.5	68.6	68.4	87.1	69.9	69.6
U ² Net(ours)	92.7	73.8	74.1	89.4	72.1	72.5	91.0	72.9	73.4
UNet(ours)	91.6	71.9	71.5	88.6	70.3	69.8	90.1	71.1	70.6

Bold represents the best test results

Table 2 Parameters of models

	RetinaNet	YOLO	Hourglass(ours)	U2Net(ours)	UNet(ours)
Parameter	121.5 MB	250.3 MB	6.6 MB	6.9 MB	5.9 MB

the parameter sizes between regression models and our methods. Table 2 shows that our methods based on segmentation are more lightweight than regression models but achieve comparable results. These results show the

ability of our proposed methods and it is worth mentioning that our methods have much simpler implementation because of the removal of some heuristic algorithms such as anchor generation, anchor matching, and the NMS process which are indispensable in regression models.

Qualitative results

To explore what the proposed module has learned and its effect, we illustrate the predicting results in some low-resolution images with adjacent inscriptions and their corresponding outputs of the pseudo-label predicting

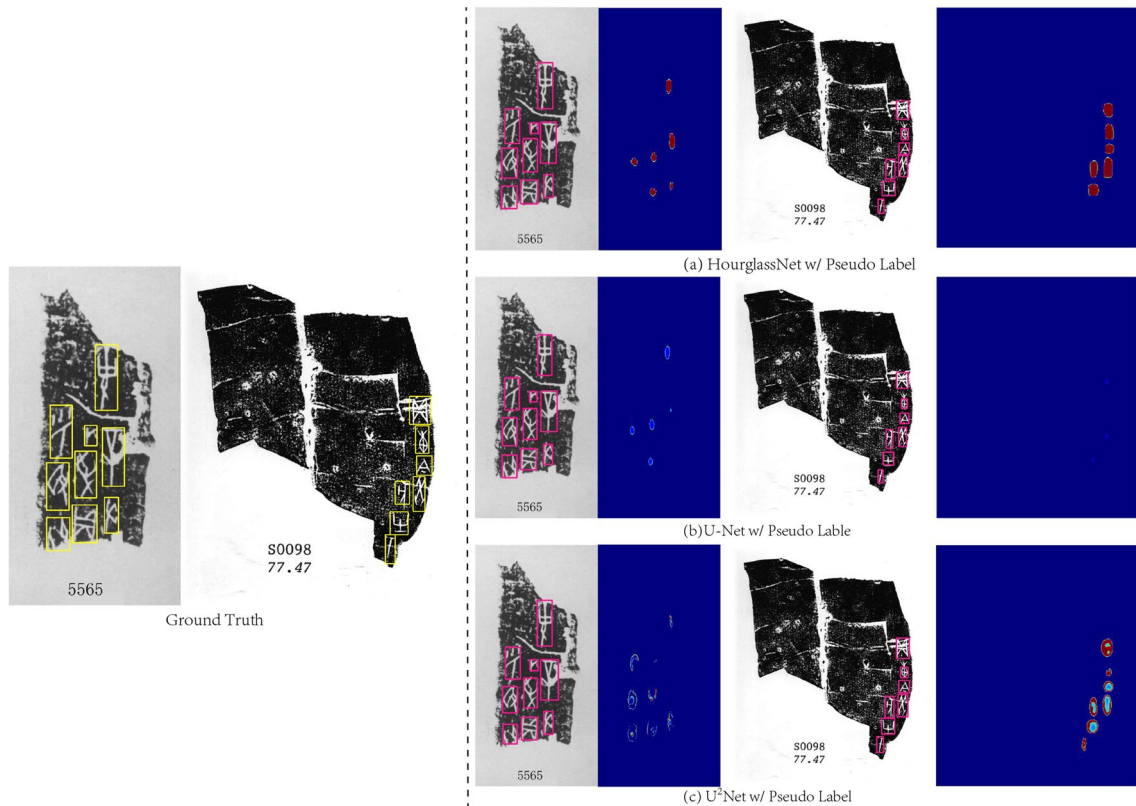


Fig. 5 Illustration of results and their relative pseudo-label mask. It can be seen that the proposed pseudo-label predicting module focuses on the region of inscriptions and generates different signals from the region of backgrounds

module in Fig. 5. As shown in the results of each first column, models can deal with low-resolution situations well. As shown in the heat maps of each second column, the areas of inscription have different values from that of strokes or background which means the proposed module can focus on the region of inscriptions and generate different signals from the region of backgrounds. In the forward propagating process, this high-dimension feature map serves as a kind of prior information of inscription and guides the next region predicting module to generate accurate results through the fusion of element-wise adding operation.

To provide an intuitive comparison, we illustrate the partial results of the proposed method and their corresponding baselines in figures. Figure 6, 7 and 8 display the prediction results of different networks for large resolution images. In the zoom-in image of Fig. 6, our proposed method can generate the result of an inscription whose strokes are very light. Figure 7 shows that the U² Net equipped with the proposed method has a better detection result than its initial edition and detects the pony-size inscription and ignores the area of cracks. Figure 8 shows the results for the HG-Net series. The proposed method leads to more precise boundary detection. The pseudo-class branch based on novel supervision proposed in this work can not only improve the utilization of the inscription's features but also avoid disturbing noise

areas and further improve detection accuracy. The models equipped with the proposed classification branch are more adept at partitioning the boundary in dense situations than basic models therefore detecting more inscriptions with small areas and slender strokes. Moreover, the newly proposed models also reduce the crack area fault detection.

In Fig. 9, we display the predictions and labels with bounding boxes with different colors on images that have a mass of noisy areas. The yellow boxes are the ground truth bounding boxes and the pink boxes are the predicted bounding boxes. There are two effects that merge the features of the pseudo-label. The models have a better capability to distinguish the area of each inscription exactly. Additionally, the models can generate clear boundaries when the inscriptions are dense. Therefore it can be seen that the predictions do not appear in the crack areas and the pink predicted boxes overlap the yellow ground-truth boxes well.

Conclusion

The identification of inscriptions in rubbings has been a problem because of the noisy area of scratches and cracks. This paper applies computer vision technology to detect traditional Chinese OBIs. We introduce a novel subtask that predicts the pseudo-category of individual OBIs. Under this subtask the pseudo-class

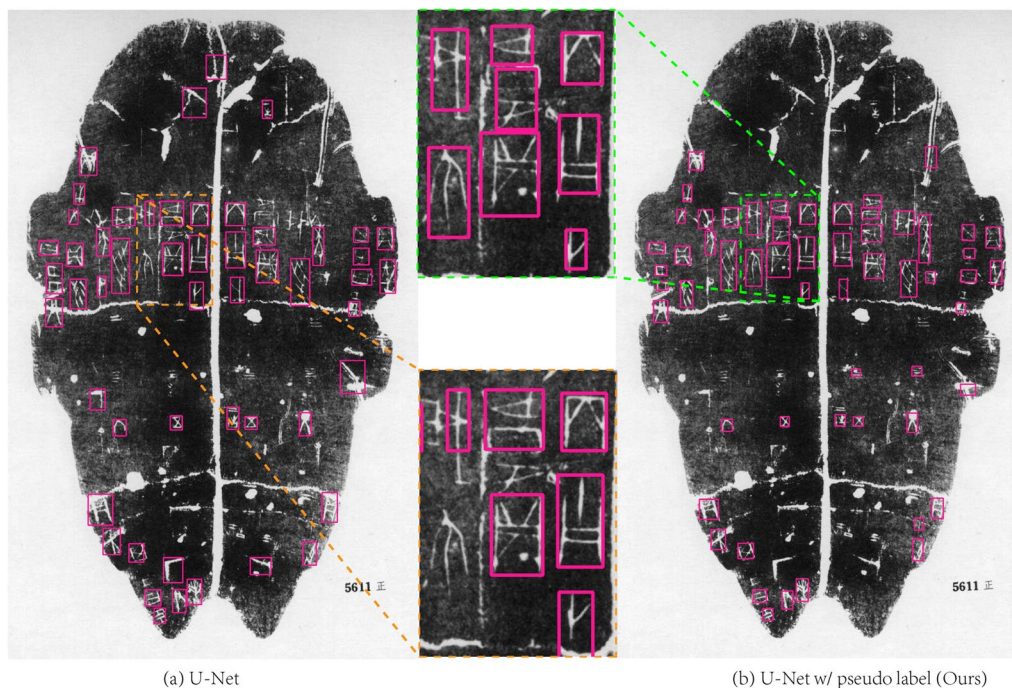


Fig. 6 Qualitative comparison of U-Net series. Zooming in the figures offers a better view of the capability in a situation where inscriptions are very dense. Compared with the original U-Net, the proposed method can detect dense inscriptions more accurately

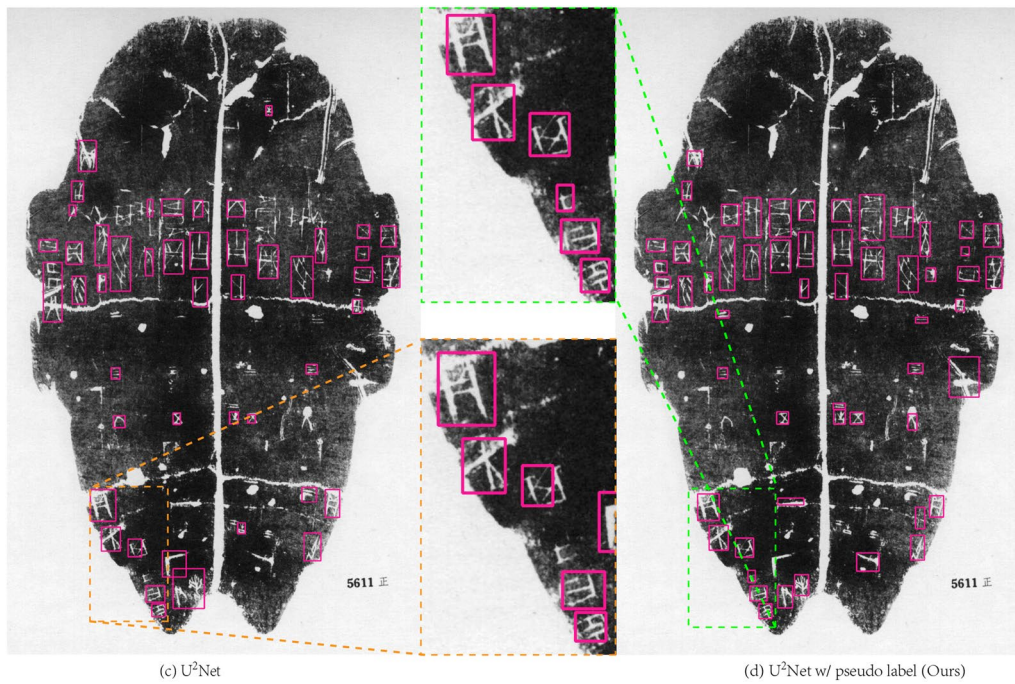


Fig. 7 Qualitative comparison of U²Net series. The zoomed-up figures offer a better view of the capability in a situation where inscriptions are in small size. Compared with the original U²Net, the proposed method can detect inscriptions with a very small size better

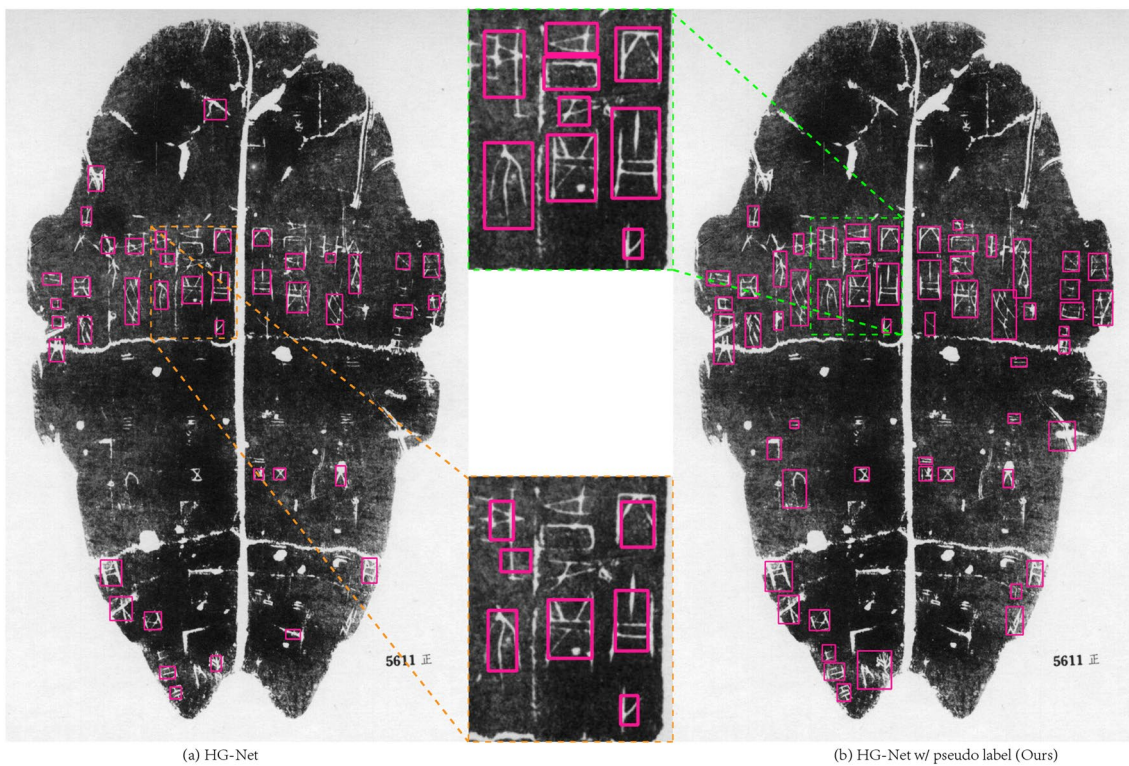


Fig. 8 Qualitative comparison Hourglass-Net series. Zooming in on the figures offers a better view of the capability in a situation where inscriptions are very dense. Compared with the original Hourglass-Net, the proposed method has a greater ability to detect dense inscriptions

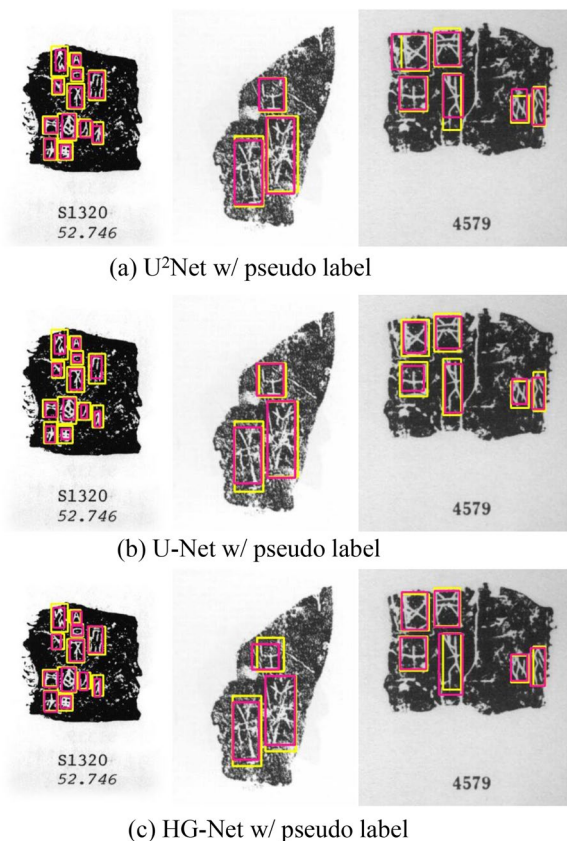


Fig. 9 Results of images that contain many noisy areas

predicting branch guides the model in understanding the form prompt of different inscriptions and the proposed method provides unique information about inscriptions to its corresponding area predictor in a fusion manner and increases the likelihood of detection with an additional parameter. Experiments on a challenging dataset demonstrated that the proposed method which acts as a plug-and-play module is suitable for different backbone networks and recognizes inscriptions precisely in different situations. Additionally, the establishment of an OBIs detection model can automatically generate corresponding visual labels for OBIs, facilitate the construction of a font database of OBIs, and play a positive role in developing archeology. It is worth exploring more distinct OBIs properties to enhance the generalizability of deep learning based models. In the future, the focus of our work will be exploring more methods that can utilize the information of individual inscriptions and how to introduce useful information to the model effectively.

Acknowledgements

Not applicable.

Author contributions

XY developed the research idea and provided valuable suggestions for this manuscript. XF conducted the experiments in this manuscript. XF and RZ wrote this manuscript. CL provided specialized knowledge of ancient writing.

Funding

This work is supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No.62206106).

Availability of data and materials

The datasets used during the current study are available from <http://jgw.aynu.edu.cn/>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

The Author confirms: that the work described has not been published before; that it is not under consideration for publication elsewhere; that its publication has been approved by all co-authors, if any; that its publication has been approved.

Competing interests

The authors declare that they have no competing interests.

Received: 14 November 2023 Accepted: 21 March 2024

Published online: 03 April 2024

References

- Meng L. Two-stage recognition for oracle bone inscriptions. In: International conference on image analysis and processing. 2017.
- Lin M. Recognition of oracle bone inscriptions by extracting line features on image processing. In: International conference on pattern recognition applications and methods. 2017.
- Gefei F, Shaotong G. Feature extraction method of oracle bone inscriptions based on mathematical morphology. *J Chin Inf Process*. 2013;27:79–86.
- Guo J, Wang C, Roman-Rangel E, Chao H, Rui Y. Building hierarchical representations for oracle character and sketch recognition. *IEEE Trans Image Process*. 2016;25:104–18.
- Liu G, Chen S, Xiong J, Jiao Q. An oracle bone inscription detector based on multi-scale Gaussian kernels. *App Math J Chin Univ Ser B*. 2021;12:224–39.
- Feng L, Xinlun Z. Recognition of Jia Gu Wen based on graph theory. *J Electron*. 1996;18:41–7.
- Meng L, Izumi T. A combined recognition system for oracle bone inscriptions. *Int J Adv Mechatron Syst*. 2017;7:235–7.
- Qingsheng L. Recognition of inscriptions on bones or tortoise shells based on graph isomorphism. *Comput Eng Appl*. 2011;47:112–4.
- Ying-min T. A graphic-based method for Chinese Oracle-bone classification. *J Beijing Inf Sci Technol Univ*. 2010;25:92–6.
- Fujikawa Y, Li H, Yue X, Aravinda CV, AmarPrabhu G, Meng L. Recognition of Oracle bone inscriptions by using two deep learning models. 2021. [arXiv:2105.00777](https://arxiv.org/abs/2105.00777).
- Xing J, Liu G, Xiong J. Oracle bone inscription detection: a survey of oracle bone inscription detection based on deep learning algorithm. In: International conference on artificial intelligence, information processing and cloud computing. 2019.
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed SE, Fu CY, et al. SSD: single shot multibox detector. In: European conference on computer vision. 2015.
- Neubeck A, Van Gool L. Efficient non-maximum suppression. In: 18th International conference on pattern recognition (ICPR'06), vol. 3; 2006. p. 850–5.

14. Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. In: IEEE conference on computer vision and pattern recognition (CVPR), vol. 2016; 2016. p. 761–9.
15. Lin X, Chen S, Zhao F, Qiu X. Radical-based extract and recognition networks for Oracle character recognition. *Int J Doc Anal Recognit (IJ DAR)*. 2022;25:1–17.
16. Li J, Wang QF, Zhang R, Huang K. Mix-up augmentation for Oracle character recognition with imbalanced data distribution. In: Document analysis and recognition—ICDAR 2021. 2021.
17. Liao M, Shi B, Bai X. TextBoxes++: a single-shot oriented scene text detector. *IEEE Trans Image Process*. 2018;27(8):3676–90.
18. Liao M, Shi B, Bai X, Wang X, Liu W. TextBoxes: a fast text detector with a single deep neural network. In: AAAI conference on artificial intelligence. 2016.
19. Ma J, Shao W, Ye H, Wang L, Wang H, Zheng Y, et al. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans Multimed*. 2018;20(11):3111–22.
20. Zhou X, Yao C, Wen H, Wang Y, Zhou S, He W, et al. East: an efficient and accurate scene text detector. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 5551–60.
21. Zhang C, Liang B, Huang Z, En M, Han J, Ding E, et al. Look more than once: an accurate detector for text of arbitrary shapes. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), vol. 2019; 2019. p. 10544–53.
22. Wang F, Chen Y, Wu F, Li X. TextRay: contour-based geometric modeling for arbitrary-shaped scene text detection. In: Proceedings of the 28th ACM international conference on multimedia; 2020. p. 111–9.
23. Zhang X, Su Y, Tripathi S, Tu Z. Text spotting transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2022 June. p. 9519–28.
24. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. 2020. [arXiv:2005.12872](https://arxiv.org/abs/2005.12872).
25. Liu Y, Chen H, Shen C, He T, Jin L, Wang L. ABCNet: real-time scene text spotting with adaptive Bezier-curve network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2020.
26. Liu Y, Shen C, Jin L, He T, Chen P, Liu C, et al. ABCNet v2: adaptive Bezier-curve network for real-time end-to-end text spotting. *IEEE Trans Pattern Anal Mach Intell*. 2022;44(11):8048–64.
27. Zhu Y, Chen J, Liang L, Kuang Z, Jin L, Zhang W. Fourier contour embedding for arbitrary-shaped text detection. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), vol 2021; 2021. p. 3122–30.
28. Liao M, Zou Z, Wan Z, Yao C, Bai X. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Trans Pattern Anal Mach Intell*. 2023;45(1):919–31.
29. Baek Y, Lee B, Han D, Yun S, Lee H. Character region awareness for text detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 9365–74.
30. Tang JR, et al. Few could be better than all: feature sampling and grouping for scene text detection. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2022; p. 4553–62.
31. Xie X et al. Toward understanding WordArt: corner-guided transformer for scene text recognition. In: European conference on computer vision. 2022.
32. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems. 2017;30.
33. Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. In: European conference on computer vision (ECCV). 2016.
34. Lin X, Chen S, Zhao F, Qiu X. Radical-based extract and recognition networks for Oracle character recognition. *Int J Doc Anal Recognit (IJ DAR)*. 2022;09(25):1–17.
35. Nayef N, Yin F, Bizid I, Choi H, Feng. ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification—RRC-MLT. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR), vol 01; 2017. p. 1454–9.
36. He M, Liu Y, Yang Z, Zhang S, Luo C, Gao F, et al. ICPR2018 contest on robust reading for multi-type web images. In: 2018 24th international conference on pattern recognition (ICPR); 2018. p. 7–12.
37. Liu Y, Jin L, Zhang S, Zhang S. Detecting curve text in the wild: new dataset and new solution. 2017. [arXiv:1712.02170](https://arxiv.org/abs/1712.02170).
38. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. 2015. [arXiv:1505.04597](https://arxiv.org/abs/1505.04597).
39. Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, Jagersand M. U2-Net: going deeper with nested U-structure for salient object detection. *Pattern Recognit*. 2020;106: 107404.
40. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), vol. 2016; 2016. p. 770–8.
41. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection.
42. Suzuki S, Be K. Topological structural analysis of digitized binary images by border following. *Comput Vis Graph Image Process*. 1985;30:32–46.
43. Kingma DP, Ba J. Adam: a method for stochastic optimization. *CoRR*; 2014. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
44. Redmon J, Divvala SK, Girshick RB, Farhadi A. You only look once: unified, real-time object detection. In: IEEE conference on computer vision and pattern recognition (CVPR), vol. 2015; 2016. p. 779–88.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.