Geometrics Assisted Rubbing Generation and Semantics Enhanced Detection for Small and Dense OBI Character

Xiuan Wan[1], Yuchun Fang[1]*, Jiahua Wu[1], Shouyong Pan[2]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai (China)
[2] School of Cultural Heritage and Information Management, Shanghai University, Shanghai (China)

* Corresponding author: ycfang@shu.edu.cn

Received 15 February 2023 | Accepted 21 February 2025 | Early Access 2 October 2025

Abstract

Character detection is essential for subsequent Oracle Bone Inscription (OBI) research. However, the lack of labeled data and the complexity of small and dense OBI characters are the main difficulties in OBI detection research. In this paper, we propose a framework for rubbing generation that can automatically build up large-scale rubbing samples with verisimilar scenarios to noisy wild OBI through geometric and morphological construction combined with style transferring. Moreover, we propose a semantic-enhanced detection model aiming at small and dense OBI through the fusion of multi-resolution feature maps with the enriched feature in the YOLOv5s backbone. We introduce the higher resolution and the Soft-NMS into the proposed OBI detection model to solve the overlapping of small and dense OBI characters. The augmented dataset improves the performance of benchmark object detection models in the real OBI detection task when sufficient data is lacking. Furthermore, the proposed OBI detection model can provide easy and preferable access to OBI detection even with a small number of labeled data and obtain preferable results. Experiments ascertain the effectiveness of the proposed OBI generation framework and the proposed OBI detection model.

Keywords

Data Augmentation, GAN, NMS, Object Detection, Oracle Bone Inscription.

DOI: 10.9781/ijimai.2025.10.001

I. Introduction

As a form of cultural heritage, characters have attracted attention from researchers in recognition [1], retrieval [2], and even art of character painting [3]. Oracle Bone Inscription (OBI), often curved on bones or tortoise shells, is an ancient character in the Shang dynasty (about 1300 BC), representing the record or divination of events. Research such as deciphering [4] is much more complicated, for studying OBI requires researchers to master professional knowledge in many fields such as history, archaeology, and writing, and such philological studies are more complicated. Traditional manual deciphering is complex, inefficient, and time-consuming. On the other hand, since researchers carried out the conventional research work directly on the carrier of ancient writing, the research progress mainly depended on a very authoritative minority of experts. Moreover, OBIs are mainly stored in the form of rubbings. Hence, the detection of OBI on rubbings is one of the preconditions for the subsequent recognition [5], [6] and semantic analysis [7], which are vital in computer-aided OBI research. Effective OBI detection systems can provide a practical reference for the researchers of OBI. Also, the OBI detection system is significant in simplifying and popularizing the research of OBI. Therefore, using OBI detection models to further aid in the study of ancient characters can provide practical help for the research of OBI and has a high research value.

In recent years, with the rapid development of deep learning, convolutional neural networks (CNN) such as R-CNN [8] outperform traditional methods. Other fields like Orthopantomogram image classification [9] and surveillance video tracking [10] are booming by CNN. Furthermore, object detection models [11]–[16] have also been applied to OBI detection research. Therefore, applying object detection models to OBI [17] and carrying out OBI detection research is the new trend to help recognize and decipher OBI.

However, the need for more annotated training data makes training deep learning models difficult. Furthermore, because of its nature, OBI labeling is very costly done by professionals. Therefore, researchers usually keep their datasets private since they need to check, scan, and align different professional materials with careful and detailed manual annotation. Moreover, in contrast to regular modern text, OBI is a distinctive character that is difficult to detect. For example, as shown in Fig. 1, OBI was usually carved densely of small size and

Please cite this article as:
X. Wan, Y. Fang, J. Wu, S. Pan. Geometrics Assisted Rubbing Generation and Semantics Enhanced Detection for Small and Dense OBI Character, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 5, pp. 78-91, 2025, http://dx.doi.org/10.9781/ijimai.2025.10.001

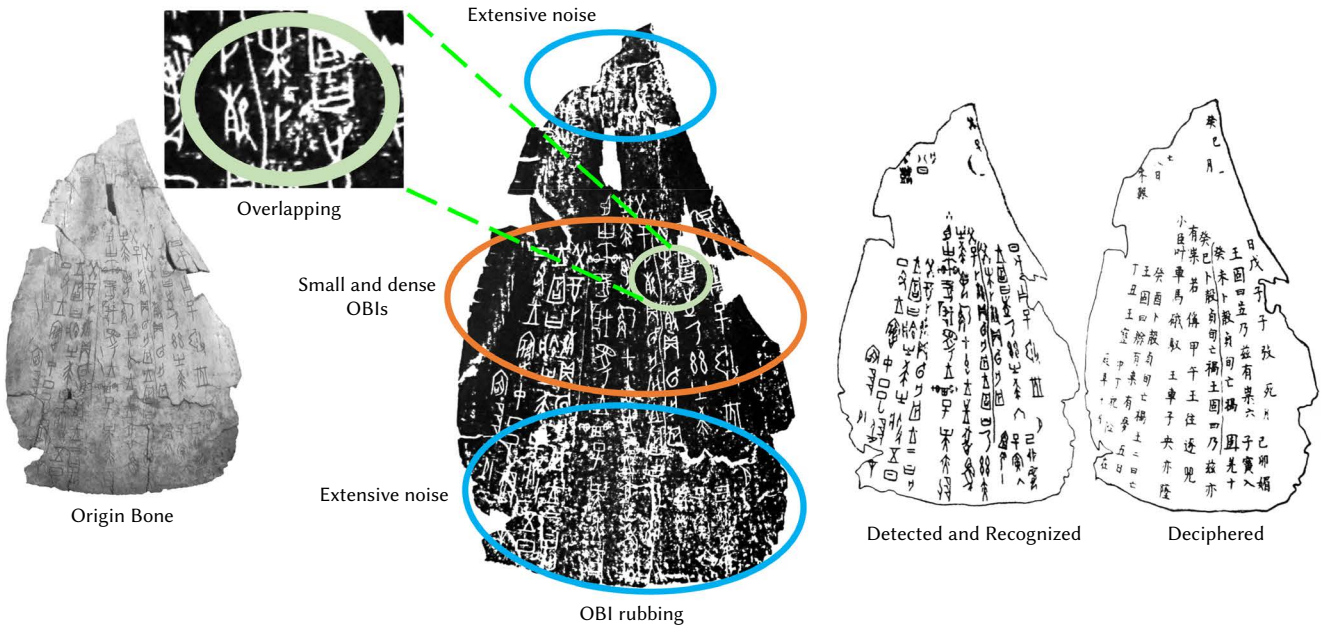


Fig. 1. Illustration of small, dense, and noisy OBI rubbing images. Rubbing images of the original bone that contains OBIs is the most common digital material for OBI research. However, the OBIs were usually carved densely in small sizes, sometimes with extensive noises and overlapping, as indicated in the red, blue, and green ellipses.

irregular distribution, along with extensive noise caused by drilling, burning, fragmentation, and other damages for sacrificial purposes and continued to be eroded in the soil. Hence, the automatic detection of OBI characters is a challenging task for obtaining the deciphered results to serve multi-disciplinary research.

To alleviate the scarcity of data for OBI, we propose a framework to generate rubbing images from single OBI images by proposed geometric algorithms and style transfer. Standard data augmentation methods such as MixUp [18] and Mosaic [16] recombine the annotated training data by overlapping and splicing. However, these methods still need sufficient training data for detection because they only combine existing objects derived from limited training data. Nevertheless, OBIs are closely interrelated with the shape, noise, and background of the bone that contains them, for which augmentation is not reasonable by clipping and stitching.

In this paper, we propose geometric and morphological construction combined with style transferring for generating rubbing images. The rubbing image reflects the bone surface that was rubbed for OBI replica. We assume OBI rubbings as polygons called rubbing bases containing single OBIs and several types of noise. Based on this assumption, we propose to generate OBI rubbings from single OBIs by a framework of controllable placement, geometric operations, and style transfer.

Although the surface usually envelopes most OBIs, different rubbing varies in shape and detail. Hence, firstly, we propose a deliberate design to arrange the selected OBIs into divided grids and calculate the convex hull of these arranged OBIs as the background geometry, which can help generate an appropriate background for placed OBI and ensure the envelopment of OBI by generated background.

Secondly, we simplify meshing from crack simulation based on Finite Element Simulation to generate realistic cracks and holes. Compared with the general physical computing method, our proposed method can reduce complexity and computation demand by improving the occurrence mechanism of cracks and maintaining the precision of stimulation for image augmentation. In particular, we construct multi-level border triangular mesh of the background geometry by geometric methods and conduct morphological operations like erosion on the constructed mesh to get a more realistic rubbing base.

Thirdly, we propose a rubbing adversarial network to acquire a more realistic style since the natural style of noise and color is essential and defined as a style-transferring procedure. The proposed model takes the rubbing base and the segmentation of that rubbing base which divides the image into different areas as inputs, outputting the transferred rubbing image with noise and realistic style.

In addition, we propose a detection model targeted on the characteristic of OBI with superior performance. We analyze several major statistics of OBI rubbing images, which confirms that OBI is of small size and dense distribution. In this regard, we propose an OBI detection model with effective methods to provide more semantic information for tiny OBI and deal with overlapping softly. As semantic information is essential for detecting small objects, we introduce the higher resolution and larger-scale feature map into the feature fusion structure and additional detection head to enrich semantic information for OBI. Besides, dealing with overlapping is also a critical concern for performance. For this purpose, We also compare different bounding box losses and propose using Soft-NMS to relieve the problem of dense distribution and overlapping. Overall, the proposed method can achieve OBI detection with scarce training data and achieve more precise results. The main contributions of this paper are summarized below:

1. We propose an OBI rubbing image generation framework that generates OBI rubbing images by geometric and morphological construction combined with style transferring, capable of providing sufficient training data for OBI detection.
2. We propose a semantic-enhanced detection model aiming at small and dense OBI to provide more semantic information through the fusion of multi-resolution feature maps with the enriched feature in the YOLOv5s backbone, with higher resolution and the Soft-NMS handling overlapping, thus reaching better performance than competitive models.
3. Our method can provide OBI rubbing images for augmentation of OBI detection and allow researchers to perform OBI detection only using limited and scarce training data. Experiments show that with the augmentation of the proposed framework, OBI detection models gain considerable performance improvement when training data is exceptionally scarce.

II. Related Work

A. Data Augmentation

Data augmentation is a set of techniques that improves the quantity, quality, and variety of data, aiming to alleviate the problem of data scarcity, poor data quality, or data imbalance.

In image processing, the most commonly used data augmentation methods are simple and effective such as rotation, clipping, and flipping. Random noise [19] is also effective for improving robustness. For object detection, these data augmentation methods are also effective. The commonly used augmentation methods for object detection are MixUp [18], Cutout [20], and Mosaic [16]. MixUp [18] mixed two images in proportion to generate a new image. Cutout [20] clipped a region of the image with zero padding. Mosaic [16] splices four pictures together into a new image.

Besides, as generative adversarial network(GAN) [21] is capable of generating data, researchers utilized it for data augmentation of data quantity, and data imbalance such as augmentation of OBI recognition [22]. In other detection fields, for alleviating data scarcity, GAN is utilized by researchers. For example, Li et al. [23] proposed to generate a shadow image with a shadow mask guiding the position of the shadow to generate target images via GAN. However, similar ideas are hard to apply to OBI rubbing images.

Generation of OBI rubbing images requires appropriately controlling the placement of each single OBI on a proper rubbing base and properly adding noise and realistic style to the generated image. In this work, we propose a rubbing base generation framework to automatically generate realistic rubbing images with labeled OBI information using single OBI images.

B. Object Detection

According to the number of detection stages, object detection models can be divided into two-stage models and one-stage models.

Two-stage object detection models are also called sparse detection models. R-CNN [8] proposed to use a heuristic algorithm to select some regions and extract the candidate regions' corresponding features using a convolutional neural network and used a support vector machine for classification. SPP-Net [24] used spatial pyramid pooling so that the fully connected layer can adapt feature map input of different sizes. Fast R-CNN [25] proposed to only performes feature extraction once and use a fully connected layer to re-correct and achieve better performance. Faster R-CNN [11] proposed generating numerous anchors on the feature map and used Region Proposal Network (RPN) to get candidate regions. Further improvements based on Faster R-CNN like Cascade R-CNN [26] proposed combination of multi-stage detectors to achieve better performance.

One-stage object detection models are based on the idea of simultaneous region extraction and classification. YOLOv1 [13] and YOLOv2 [14] proposed dividing the image into several grids and predicted the object in each grid. SSD [12] adopted the anchor mechanism in Faster R-CNN to perform multi-scale prediction on feature maps. YOLOv3 [15] also adopted the anchor mechanism but used the prior anchors obtained by clustering and the Feature Pyramid Network (FPN) [27] to predict on multi-scale feature maps. YOLOv4 [16] further adopted the CSP backbone [28] and introduced spatial pyramid pooling and a modified Path Aggregation Network (PAN) [29] structure. The data augmentation method with better localization loss achieved high accuracy and speed simultaneously. YOLOv5 further simplifies the network structure as much as possible under engineering experiments. It achieved fast speed and a tiny model size under a design similar to YOLOv4. In recent years, Transformer [30] has been proven to perform well in image tasks [31]. Furthermore, DETR [32] proposed to use Transformer structure to directly input the feature map output by the backbone into the encoder-decoder structure to generate the result to the feedforward neural network for prediction, which is a precedent for the Transformer-structured object detection model.

C. OBI Detection

Due to the lack of data, previous OBI detection researchers usually constructed their private datasets and directly applied object detection models. Meng et al. [33] proposed to use SSD for OBI detection on a small Oracle detection dataset. Fujikawa et al. [34] proposed to use YOLO-tiny for OBI detection before classification. Xu [35] proposed to use YOLOv2 to detect and identify OBI radicals on an Oracle bone character dataset. Liu et al. [36] proposed to use Mask R-CNN for OBI detection and achieved a preferable performance. Xinth et al. [37] produced the first public OBI detection dataset tested by the mainstream object detection models. They proposed the YOLO-SPPG and YOLO-ASPP detection models based on improving the YOLOv3 network neck by combining different pooling layers to achieve better performance.

III. Methods

The overview of the proposed methods is shown in Fig. 2. The rubbing base generation constructs a primary rubbing base e using single OBI images by geometric and morphological operations. Rubbing Base Construction (RBC) and Rubbing Transformation (RT) algorithms are designed to prepare a rubbing base in the rubbing base generation module. RBC builds the rubbing base by appropriately controlling each OBI's placement and constructing a multi-level border mesh on a convex hull to place these OBIs. RT adds cracks, holes, and irregular shapes to the constructed rubbing base from RBC. The rubbing adversarial network generator takes rubbing base e and segmentation s to transfer e to a realistic and noisy style as rubbing images for training data of the detection model. Accordingly, real rubbing images are used to train the rubbing adversarial network.

Meanwhile, an OBI detection model of improvements aiming at enriching semantic information and dealing with overlapping is proposed. These improvements aim at characteristics of OBI for performance enhancement. The proposed model provides a preferable performance on OBI detection for researchers.

A. Rubbing Base Generation Framework

As shown in Fig. 3, geometric and morphological operations are proposed in the rubbing base generation module to construct the background rubbing base.

1. Algorithm RBC

In the RBC algorithm, the first step is placing OBIs for further procedure and obtaining label locations. N single OBI images $I=\{I_1,...,I_N\}$ are randomly selected as candidates. Then I is arranged on a canvas equally divided into $\lceil\sqrt{N}\rceil+2$ rows and $\lceil\sqrt{N}\rceil+2$ columns. Only the central $\lceil\sqrt{N}\rceil\times\lceil\sqrt{N}\rceil$ grids $G=\{G_1,\ldots,G_{\lceil\sqrt{N}\rceil\times\lceil\sqrt{N}\rceil}\}$ are used for avoiding bad cases of placement in the border. Each selected single OBI image I_i is placed in a randomly selected grid G_i. If the placement is successful (no overlapping), the grid G_i is no longer used, and the box position (x_1, y_1, x_2, y_2) of I_i is recorded as the label $Label_i$. After placing all OBI images, all these labels information $Label=\{Label_1, ..., Label_N\}$ are acquired to form the segmentation mask required to be fed into the rubbing adversarial network.

Real rubbing images are more than just some OBIs, whose background geometry is also significant. Hence the convex hull is adopted in the second step, where a convex hull is the smallest polygon that envelope a set of points. The ordered convex hull vertexes $V=[V_1, V_2,..., V_M]$, $V_i=(x, y)$ are calculated from Label as the simple

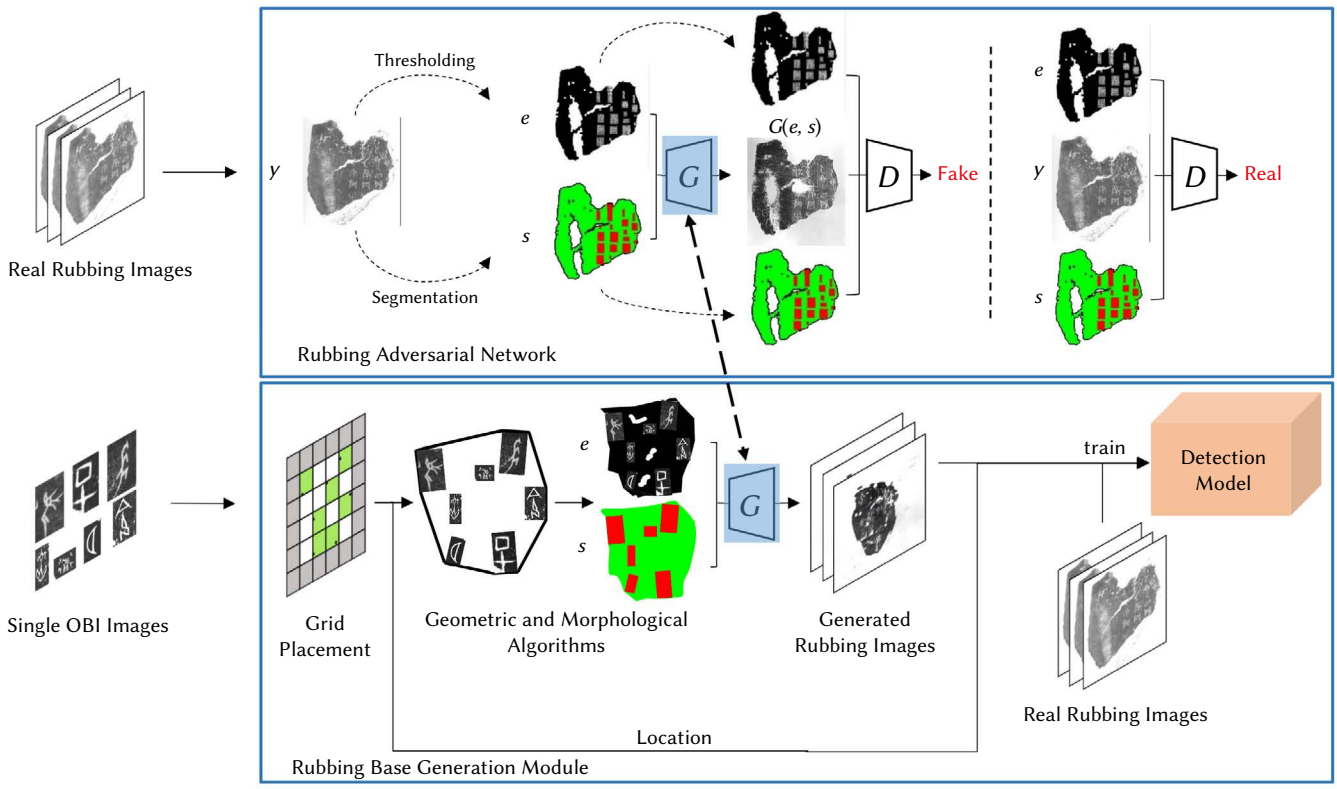


Fig. 2. The pipeline of the rubbing base generation framework.

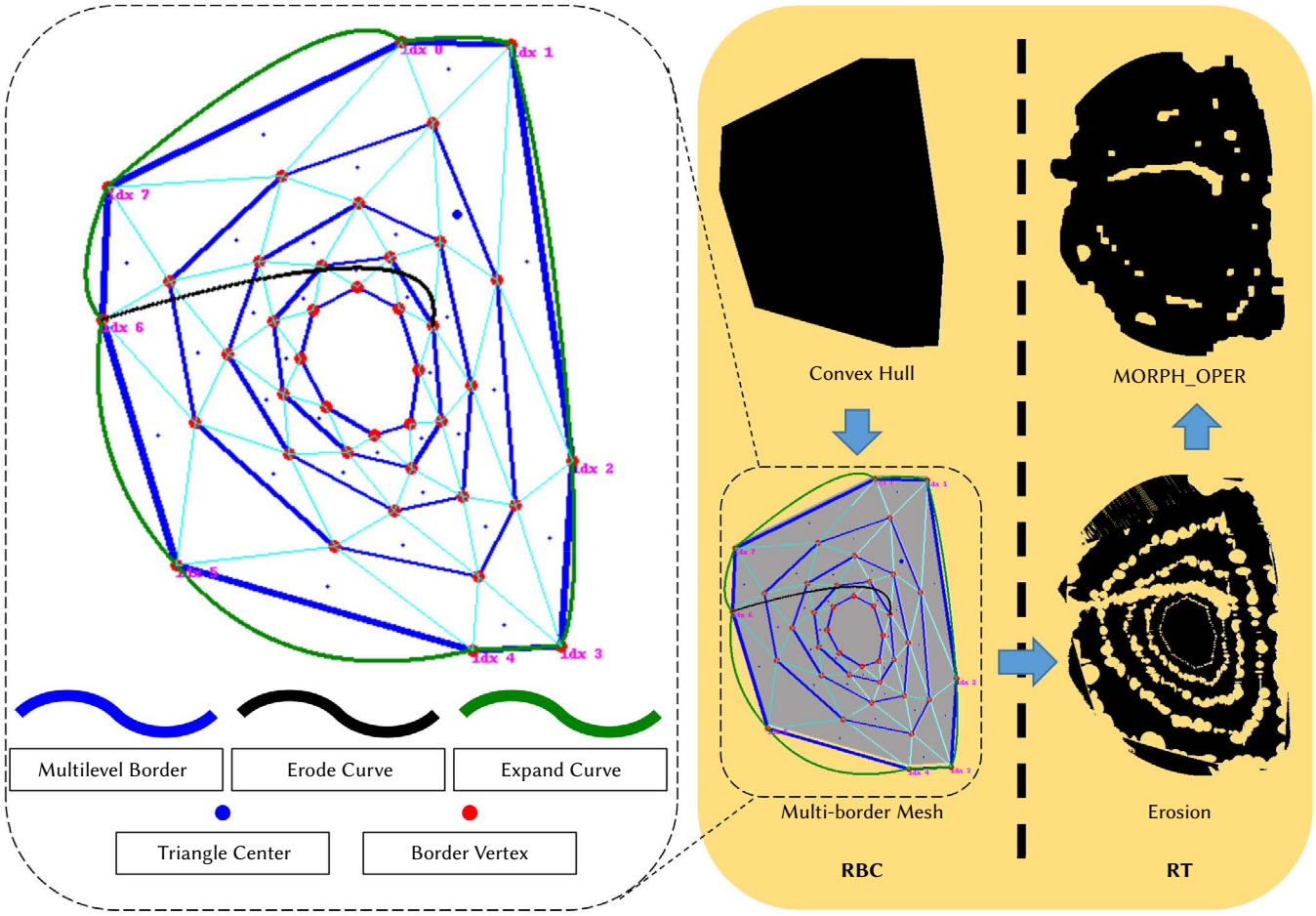


Fig. 3. The geometric and morphological operation to generate rubbing base.

and regular rubbing base that envelopes all placed OBIs. Moreover, the center point $P_c = (x_c, y_c)$ of the convex hull vertexes V is calculated.

To obtain an irregular shape and stimulate cracks or holes, we construct a multi-level borders mesh $B = [B_1, B_2, B_3, B_4, B_5], B_i = \{(x, y)\}$, where B_i are sets of border vertexes. In particular, the first border B_1 is the convex hull vertexes V. Each border B_i is extended from the former border B_{i-1} for the next four borders. For each adjacent vertex pair (x_1, y_1), (x_2, y_2) in B_{i-1}, their distance d and midpoint $P_m = (x_m, y_m)$ are calculated. A new vertex $P_{new} = P_m + 0.25d \frac{\overrightarrow{P_m P_c}}{|\overrightarrow{P_m P_c}|}$ is added into B_i, where $\overrightarrow{P_m P_c}$ is the vector from P_m pointing to P_c. Meanwhile, (P_m, P_c, P_{new}) represents a triangle area, and the center points $T = [T_1, ..., T_M]$, $T_i = (x, y)$ of these triangle areas are recorded for the latter erosion operation.

2. Algorithm RT

In the RT algorithm, some post-processing operations are conducted on the multi-level border mesh constructed by RBC. Expansions and erosions based on the multi-level border mesh and morphological operations are conducted to make the rubbing base more realistic and irregular. For such geometric purposes, the bezier curve is adopted. A bezier curve is a smooth curve controlled by several points, and three points can define a quadratic bezier curve, as shown in Equation (1).

$$C = (1-t)^2 P_1 + 2t(1-t)P_2 + t^2 P_3, t \in [0,1] \quad (1)$$

where P_1, P_2, P_3 are controlling points, and t is the step variable ranging from 0 to 1.

This way, the rugged and jagged convex hull border B_1 is extended by bezier curves to obtain a more smooth and more realistic shape. For each border line from the adjacent vertex pair (x_1, y_1), (x_2, y_2) in B_i, we calculate their distance d and a point P_d on this borderline, where P_d cannot be too close to the endpoints of the line. Hence $P_d = (ux_1 + (1-u)\, x_2, uy_1 + (1-u)\, y_2)$, where u is randomly sampled from [0.2, 0.8]. Then the outer point $P_{expand} = P_d + wd \frac{\overrightarrow{P_c P_d}}{|\overrightarrow{P_c P_d}|}$ is obtained as the point P_2 of a quadratic bezier C curve controlled by $P_1 = (x_1, y_1)$and $P_3 = (x_2, y_2)$, where w is randomly sampled from [0, ⅓] because the curve does not need to be too curving. The area enclosed by the line and this quadratic bezier curve C is filled with black as expansion.

The border of real rubbing images is usually rough and irregular. Hence, white ellipses of small random radius as erosion are created on each border line from the adjacent vertex pair (x_1, y_1), (x_2, y_2) in B_i. Furthermore, the same erosion is conducted to stimulate cracks, and the holes on the quadratic bezier C curve defined by randomly selected $P_1 \in B_1$, $P_2 \in T$, and $P_3 \in B_4$ because curves usually start at the outer border.

Finally, several morphological operations are conducted to eliminate the flaw of discrete sampling in bezier curve expansion and erosion. They are ERODE, OPEN, CLOSE, OPEN, and ERODE sequentially.

3. Rubbing Adversarial Network

After RBC and RT, a rubbing base e is obtained, which contains several single OBIs with the control of their placement. However, e still needs to include a realistic style of noise and color. To further transfer e into a more realistic style, a generative adversarial network is proposed to realize such a style-transferring goal. First, we denote the segmentation of the rubbing area, OBI, and background of a rubbing image as s. Then the trained generator G is used for style transferring the rubbing base processed by RT, and the discriminator D is discarded in the stage.

The training procedure of rubbing adversarial net is diagrammed in Fig. 2. Considering an actual rubbing image y, the threshold result of y is taken as e, and its corresponding segmentation is s. The generator G is trained to output a fake rubbing image from input e and s, while the discriminator D is trained to distinguish fake from real with corresponding e and s. The adversarial loss of the rubbing adversarial net is expressed as Equation (2).

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}\ [\log D(e, s, y)] + \mathbb{E}\ [\log(1 - D(e, s, G(e, s))] \quad (2)$$

where G is trained to generate high-quality fake rubbings to minimize this objective to confuse D and D is trained to distinguish fake rubbings from real rubbings to maximize it correctly.

In order to keep the general information of the rubbing base e, taking L1 Loss as a pixel-level restriction is beneficial [38], [39] for less blurring. However, applying L1 loss to OBI rubbing images may cause instability or even divergence in training because the noises on rubbing images are unstable and random. Hence, the image is divided into different areas for loss calculation, where complete, bone, and char are areas according to the segmentation areas indicated by segmentation s and the rubbing base e. A division sample is shown in Fig. 4. The rubbing loss is divided into three parts according to the segmentation, which can guide the generator G to generate better-quality rubbing images and avoid divergence in training. The rubbing loss is computed in Equation (3).

$$\mathcal{L}_{rubbing}(G) = \alpha \mathcal{L}_{full}(G) + \beta \mathcal{L}_{bone}(G) + \gamma \mathcal{L}_{char}(G) \quad (3)$$

where α, β, and γ are balance factors for the contents of different area.

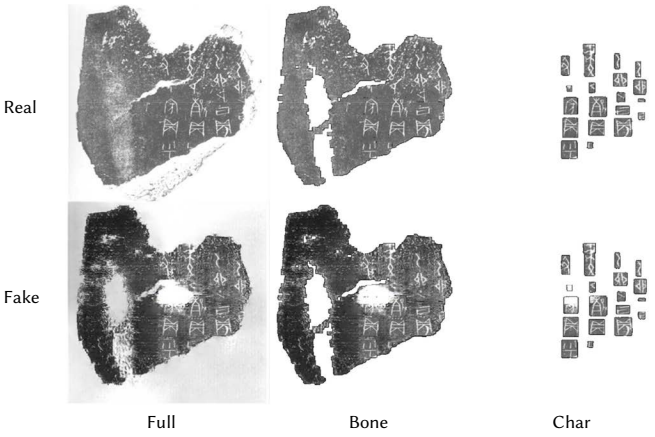


Fig. 4. The division of rubbing images. The complete area is the whole rubbing image. The bone area is the OBI rubbing with single OBIs, and the character area only contains the single OBI images. They are derived from segmentation gained by threshold processing of real rubbing images and the generated rubbing base.

In order to keep the whole image information, Smooth L1 loss is adopted instead of L1 loss to make G insensitive to outlier pixel noises for the complete area. The total image loss calculated between the whole area of the fake rubbing image and the real rubbing image is computed in Equation (4).

$$\mathcal{L}_{full}(G) = \mathcal{L}_{SmoothL1}(G(e, s)_{full}, y_{full}) \quad (4)$$

where the fake rubbing image is denoted as $G\ (e, s)_{full}$, the real rubbing image is denoted as y_{full}.

To make G more insensitive to unstable noises, L1 loss after a 2 x 2 MaxPool2d layer is adopted for the bone area. The bone area loss calculated between the bone area of the fake rubbing image and the real rubbing image is computed in Equation (5).

$$\mathcal{L}_{bone}(G) = \mathcal{L}_{L1}(MaxP(G(e, s)_{bone}), MaxP(y_{bone})) \quad (5)$$

where the fake rubbing bone area is denoted as $G\ (e, s)_{bone}$ and the real rubbing bone area is denoted as y_{bone}, and the MaxP(·) is the MaxPool operation.

Since the OBIs are the detection targets, the L1 loss is adopted to keep the character area consistent with input e. The character area loss calculated between the character area of the fake rubbing image and the real rubbing image is computed in Equation (6).

$$\mathcal{L}_{char}(G) = \mathcal{L}_{L1}(G(e,s)_{char}, y_{char}) \tag{6}$$

where the fake rubbing character area is denoted as $G(e,s)_{char}$ and the real rubbing character area is denoted as y_{char}.

The final objective is denoted as Equation (7).

$$\arg\min_{G}\max_{D} \mathcal{L}_{GAN}(G,D) + \lambda\mathcal{L}_{rubbing}(G) \tag{7}$$

where λ is a balance factor for the balance of adversarial loss and the rubbing loss.

The structure of the rubbing adversarial network is shown in Fig. 5. The U-Net [40] structure is adopted as the generator G, and the network is derived from [39], which contains down-sampling blocks and up-sampling blocks with skip-connection. A simple 5-layer CNN is used as the discriminator D, which is composed of sequential Conv-BN-LeakyReLU units. The input channels of five convolution layers are 9, 64, 128, 256, and 512, with all kernel sizes set to 4.

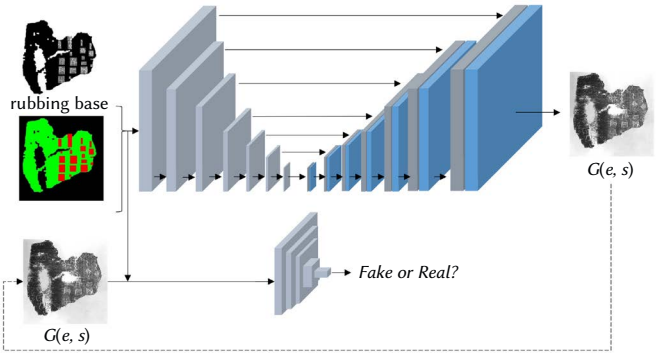


Fig. 5. The structure of the rubbing adversarial network.

B. Detection Model

Since OBI is a complicated character, missing components may change the meaning of an OBI, and the model performance for detection is essential. Therefore, a detection model is proposed, which is enhanced from YOLOv5 [41] backbone for its flexibility in deployment and efficient training with the effective Mosaic [16] data augmentation. Furthermore, four significant improvements are proposed in the detection model for more rigorous detection ability. The structure of the proposed detection model is shown in Fig. 6.

Resolution. For object detection models based on feature maps obtained by convolutional networks, a larger input resolution can lead small targets easier to detect because tiny objects have a smaller range of semantics in the feature map after multi-layer convolution down sampling. As a result, tiny objects have more semantic information in the feature map with a larger resolution, making them easier to detect. For this reason, the input resolution is enlarged to 960 while the original input resolution is 640.

Feature Fusion. In the original YOLOv5s model, only three scales of the feature map of 1/8, 1/16, and 1/32 are fed to the FPN and PAN structures and prediction heads for feature fusion. Since multi-scale feature fusion is essential for the detection performance of small objects [42]. Hence, by adding a 1/4 scaled feature map to participate in detection in FPN and PAN feature fusion structure, it should be able to add lower-level detail location semantic features to the model for small objects. For this reason, two C3 layers similar to CSP Bottleneck [28], and two convolution layers are inserted to fuse the 1/4 feature map in FPN and PAN structure for an extra detection head P2.

Box Loss. In the earlier object detection model, the loss used to evaluate the degree of deviation between the predicted box and ground truth is Smooth L1 or L2 loss. However, these losses are sensitive to scale change. Further, the IOU coefficient is used to predict the regression of the box [43], but when the two box does not overlap at all, no matter how far they deviate, their IOU coefficients are 0. GIOU [44] proposes to add the difference between the minimum circumscribed rectangle

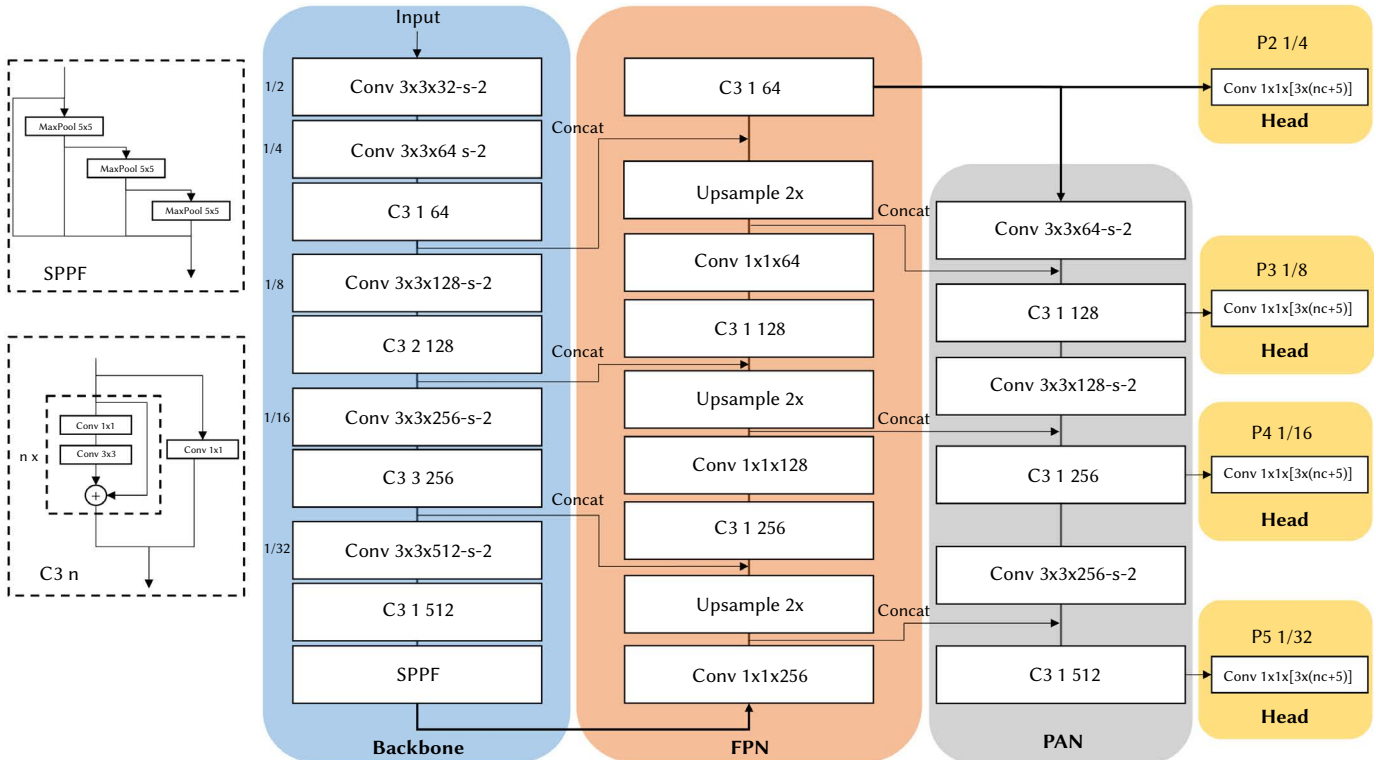


Fig. 6. The structure of the proposed OBI detection model.

and the union of the two rectangles to measure the degree of deviation of the two rectangles. DIOU [45] proposes to use the square of the ratio between the diagonal distance of the smallest circumscribed rectangle and the distance between the centers of the two rectangular boxes to measure the degree of offset between the two boxes. CIOU [45] proposes introducing the respective aspect ratios of the predicted box and the ground-truth box for further measurement. These box losses are evaluated in the experiment for selecting box loss adopted in the proposed model.

NMS. NMS (Non-Maximum Suppression) is an algorithm used for redundant screening of the prediction by the object detection model because common competitive detection models predict overlapping results. However, NMS has certain defects. For example, when the distribution of targets is dense or the overlapping area is large, satisfactory performance cannot be obtained by setting threshold, where overlapping is handled by discarding prediction with lower confidence score. Soft-NMS [46] is adopted in the proposed model instead of NMS to solve this defect. Soft-NMS makes a simple improvement: instead of directly culling other boxes, the score decay is performed according to their overlapping extent.

IV. Experiments

A. Dataset

The proposed method consists of two main parts, the OBI detection model and the rubbing generation framework. First, they are trained on the same dataset independently, as detailed below. Then, the two parts of the proposed method are evaluated on the OBI-Detection dataset.

The OBI-Detection dataset is derived from the dataset collected by Xing et al. [37]. It is an OBI rubbing image dataset for OBI detection. The dataset is collected from 9134 scanned OBI rubbings with 51864 OBI annotations. We split the dataset into 9:1 for training and testing. There are 8,221 training samples and 913 testing samples. All OBI detection models are trained on the OBI-Detection dataset. Besides, the proposed rubbing adversarial network is also trained on the same dataset, using real rubbing images and segmentation as the inputs of the generator. Single OBI images are selected from the training set as the data source for the rubbing generation framework to generate rubbing images.

B. OBI Analysis

OBI analysis is conducted to reveal that the OBI has a small size and dense distribution characteristics. There are 51,684 OBI annotations on 9134 rubbing images in the OBI-Detection dataset. We analyze the number (Num) of OBIs per image, the box scale per Ground Truth (GT), and the polygon scale per image. The polygon scale refers to the quotient of the sum of OBI areas over the convex hull areas of OBIs in an image. As shown in Fig. 7, the density and overlapping of the OBI distribution can be quantitatively analyzed by calculating the polygon scale to analyze the distribution of OBIs.

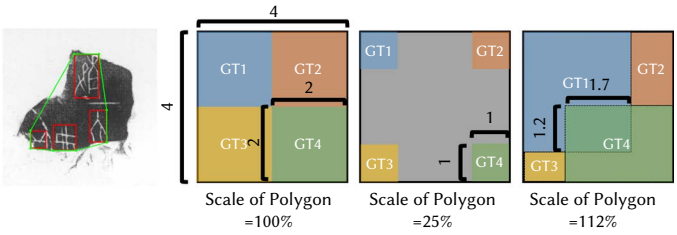


Fig. 7. The polygon scale statistic examples.

These statistical results are shown in Table I. The mean number of OBI per image is 5.66, and the median number of OBI per image is 4, indicating OBI is not much on rubbing images. While the mean scale of OBI is only 1.487% of the image, confirming that OBI is a kind of small object. The mean polygon scale of OBIs per image is 61.24%, and the median polygon scale is 61.38% of the image. Hence, over half of the convex hull area is occupied by OBIs. Also, the max polygon scale is 148.1%, meaning some extreme overlapping cases exist. These results indicate that OBIs have a compact layout with some overlapping cases.

TABLE I. Results of OBI Analysis

Statistics Type	Num / Image	Scale (%) / GT	Scale (%) / Image
Min	1	0.00069	1.968
Max	158	30.35	148.1
Mean	5.66	1.487	61.24
Median	4	0.9909	61.38

C. Experiment Setup

Since the proposed method consists of a rubbing generation framework and an OBI detection model, the training process is divided into two parts. Pytorch 1.11.0 is adopted as the deep learning platform.

For training the proposed detection model, SGD is adopted as the optimizer with the weight decay set to 5×10^{-4}, and the parameter momentum set to 0.8. The number of epochs is set to 100. The initial learning rate is $\times 10^{-2}$, and the parameter momentum is 0.8. The final learning rate is adjusted to $\times 10^{-3}$ as the cosine annealing.

For training the rubbing adversarial network, Adam is adopted as the optimizer. The learning rate is 3×10^{-5}. Balance factor α is set to 0.3, β is set to 0.4, γ is set to 0.3, and λ is set to 100.

A total of 8,221 rubbing images are generated as the data source for selection in the augmentation of training detection models along with real training samples.

D. Performance Evaluation

Appropriate metrics are essential to evaluate the performance of classification and location in the detection model. In experiments, the **AP0.5** and **AP0.5:0.95** are metrics used for evaluation, for which higher is better.

IOU (Intersection over Union) is used by Object detection models to measure the closeness between prediction and GT, which is shown in Equation (8).

$$IOU = \frac{|Prediction \cap GT|}{|Prediction \cup GT|} \tag{8}$$

IOU ranges from 0 to 1, representing the extent of overlapping between the prediction bounding box and the GT bounding box. A high IOU means stricter prediction to GT. Under a certain IOU threshold, any prediction that has IOU with a corresponding GT lower than the IOU threshold is considered a false prediction.

Precision and *Recall* are fundamental indicators for performance, which are shown in Equation (9) and Equation (10).

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

Similarly, *Precision* and *Recall* are calculated from TP, FP, and FN. TP represents prediction with the correct category and location, FP represents the wrong prediction, and FN represents the target missed by the model. *Precision* refers to the ratio of correct predictions among all targets found by the model, reflecting the model's ability to find targets correctly. *Recall* refers to the ratio of targets found by the model among all labeled targets, reflecting the model's ability to find all

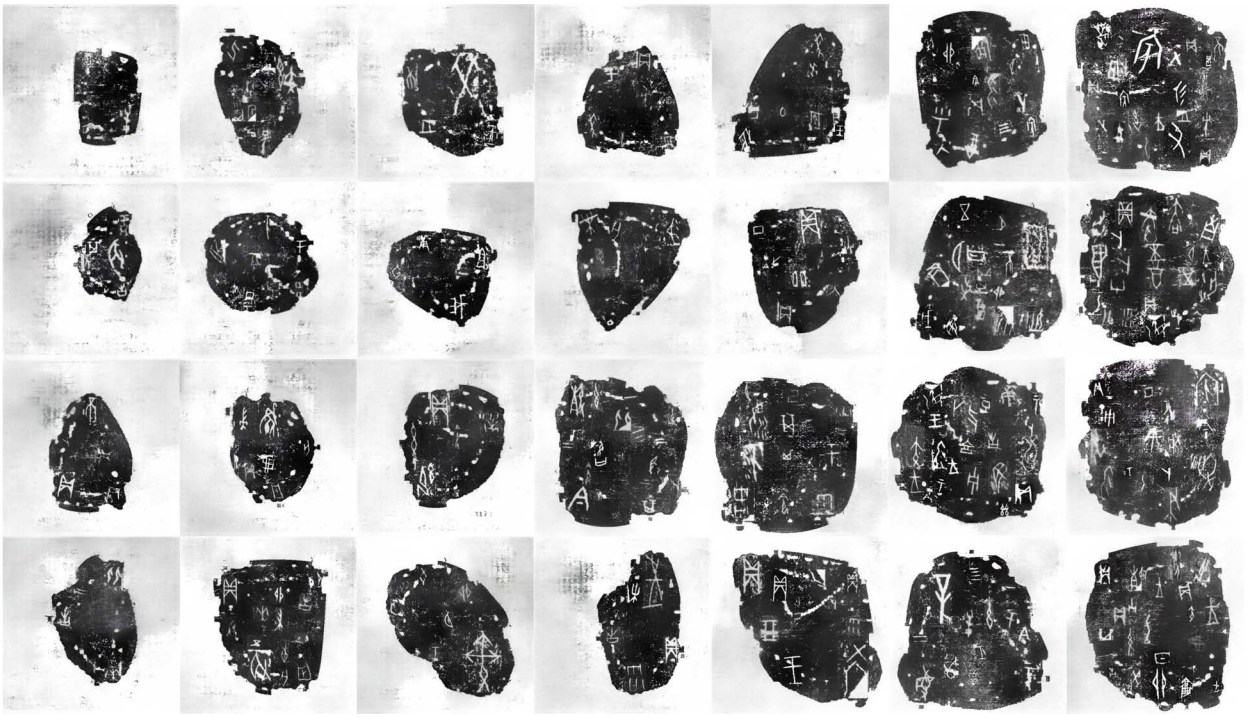


Fig. 8. Examples of rubbing images generated by the rubbing generation network.

occurring targets. By setting different confidence thresholds, Precision and Recall may vary. Generally, high *Precision* will result in low *Recall*.

AP0.5 and AP0.5:0.95 are commonly used metrics extended from *Precision* to evaluate the performance of detection models. AP means the average precision with different confidence thresholds under a certain IOU threshold. AP0.5 represents the AP under the IOU threshold of 0.5, which means the correct prediction should overlap with the corresponding GT by more than half. AP0.5:0.95 means the average of the AP at IOU thresholds from 0.5 to 0.95 in a step of 0.05. AP0.5:0.95 considers the average location performance under a loose IOU threshold(e.g., 0.5) and a strict IOU threshold(e.g., 0.95), representing the overall performance of different positioning accuracy requirements.

E. OBI Rubbing Generation

1. Augmentation on Different Models

Fig. 8 shows some generated rubbing images. It can be captured in the shown samples that each placed single OBI image has not deteriorated during the transferring procedure, and they keep their basic information. Also, it can be observed that the generated rubbing images are realistic and reasonable in shape.

Detection models are trained under the circumstances of highly scarce training data and augmentation of generated data to evaluate the proposed rubbing generation framework. Since manual annotation for a single rubbing image is time-consuming for checking professional materials, it is reasonable to define a total of 10-100 annotated images as 'scarce'. In particular, the performance of several models is compared when only using 0.5% of the training data (41 images) and using 0.5% of training data mixed up with 10% of the generated data. The experimental results are shown in Table II, where '-' represents results in cases that are not converged well.

Since only using 0.5% of real training data is in extreme data scarcity, some of these models are difficult to converge in training. The SSD300 [12], deformable DETR [47], and Faster R-CNN [48] are hard to be trained. Hence, these three models perform poorly when only using 0.5% real training data. However, after augmentation of mixing up with generated data, all these models gain considerable performance improvement on AP0.5 and AP0.5:0.95, demonstrating the effectiveness of the proposed rubbing generation framework in providing sufficient training data. Meanwhile, the DETR [32], YOLOv3 [15], YOLOv5s [41], and the proposed detection model perform better when training data is scarce, and they also gain noticeable improvement in performance both on AP0.5 and AP0.5:0.95. This improvement results from the increment and improved variety of augmented training data, which is essential for training deep neural networks. Overall, these models all gain a noticeable performance boost by augmentation of the proposed rubbing generation framework, proving its effectiveness.

TABLE II. Results of Augmentation on Different Models

	AP0.5		AP0.5:0.95	
Models	w/o	w	w/o	w
SSD300 [12]	- (0.210)	0.526	- (0.060)	0.189
DETR [32]	0.639	0.675	0.263	0.306
deformable DETR [47]	- (0.203)	0.611	- (0.049)	0.263
Faster R-CNN [48]	- (0.316)	0.641	- (0.132)	0.309
YOLOv3 [15]	0.603	0.657	0.198	0.271
YOLOv5s [41]	0.610	0.690	0.246	0.311
the proposed model	0.611	0.711	0.238	0.338

w/o: without augmentation. w: with augmentation.

2. Augmentation Quantity Analysis

To further analyze the effectiveness of the rubbing generation framework, the proposed detection model and YOLOv5s are trained under different amounts of generated data for comparison. In particular, three groups of 0.5%, 1.0%, and 1.5% of the original 8,221 training samples are set as baselines, containing 41, 82 and 123 images, respectively. Augmentation is mixing up 2%, 4%, 6%, 8%, and 10% of the 8221 generated rubbing images for training the detection model and real training samples. These two models are also evaluated on the test set of 913 real rubbing images. For each group, the mean values are calculated for the metrics result. The experimental results are shown in Table III, Table IV, Table V, and visualized in Fig. 9.

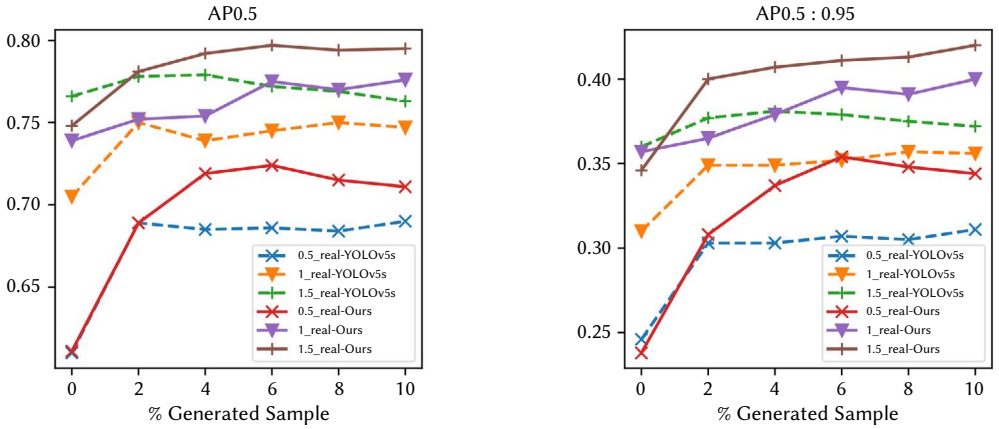


Fig. 9. Results of different portions of mixing real and generated group.

TABLE III. Results of 0.5% Real Group

	AP0.5	AP0.5:0.95
Gen	YOLOv5s / Ours	YOLOv5s / Ours
0%	0.610 / 0.611	0.246 / 0.238
2%	0.689 / 0.689	0.303 / 0.308
4%	0.685 / 0.719	0.303 / 0.337
6%	0.686 / 0.724	0.307 / 0.354
8%	0.684 / 0.715	0.305 / 0.348
10%	0.690 / 0.711	0.311 / 0.344
mean	0.687 / 0.712	0.306 / 0.338

0.5% real training samples is the first experiment group. For the first group of only 0.5% real training samples, the results without any generated rubbing image are worse for both models. Their AP0.5 and AP0.5:0.95 are low, indicating that the models can detect OBI but are not precise. This inferior performance is because the detection model needs at least sufficient training samples. The performance is boosted by using the generated rubbing images and the real data in training. After adding generated rubbing images in training, the AP0.5 and the AP0.5:0.95 increased. This improvement is because the generated rubbing images enrich the training samples with similar shapes and characteristics of the extended noise, which proves the effectiveness of the proposed rubbing generation network.

TABLE IV. Results of 1% Real Group

	AP0.5	AP0.5:0.95
Gen	YOLOv5s / Ours	YOLOv5s / Ours
0%	0.705 / 0.739	0.310 / 0.357
2%	0.750 / 0.752	0.349 / 0.365
4%	0.739 / 0.754	0.349 / 0.379
6%	0.745 / 0.775	0.352 / 0.395
8%	0.750 / 0.770	0.357 / 0.391
10%	0.747 / 0.776	0.356 / 0.400
mean	0.746 / 0.765	0.353 / 0.386

1% real training samples is the second experiment group. For the second group of only 1% real training samples, the results without any generated rubbing image are better than the first group. The AP0.5 and the AP0.5:0.95 are close to the first group for both models because detection models need training samples of a similar distribution to test samples, which means real data is best for improving the performance. The performance is also boosted by using the generated rubbing images and the real training data. After adding generated rubbing images in training, the AP0.5 and the AP0.5:0.95 increased for both models. Although the generated rubbing images are different from real samples in distribution, the enrichment in the training samples is more important than the distortion in distribution under such a scarcity of real data, which also proved the effectiveness of the proposed rubbing generation network.

TABLE V. Results of 1.5% Real Group

	AP0.5	AP0.5:0.95
Gen	YOLOv5s / Ours	YOLOv5s / Ours
0%	0.766 / 0.748	0.360 / 0.346
2%	0.778 / 0.781	0.377 / 0.400
4%	0.779 / 0.792	0.381 / 0.407
6%	0.772 / 0.797	0.379 / 0.411
8%	0.769 / 0.794	0.375 / 0.413
10%	0.763 / 0.795	0.372 / 0.420
mean	0.772 / 0.792	0.377 / 0.410

1.5% real training samples is the third experiment group. For the third group of 1.5% training samples, the results without any generated rubbing image are slightly better than the second group. However, the AP0.5 and the AP0.5:0.95 are high, indicating the importance of sufficient training data. The performance is also boosted prominently in AP0.5:0.95 by using the generated rubbing images and the real data in training. While the improvement is slight compared with the former groups, this may indicate that the influence of the distortion in distribution will become dominant when real training samples are sufficient.

As shown in Fig. 9, both YOLOv5s and the proposed model significantly improve on AP0.5 and AP0.95. Furthermore, it can be found that both models gain considerable improvement after being trained by mixed real data with augmentation of generated rubbing images on both metrics.

Among them, the improvement is most significant when there is only 0.5% real data. This improvement indicates the importance of sufficient training data. Also, more images generation may lead to better performance, which provides more training data. However, the improvement tendency is not rising for YOLOv5s; this may result from the different distribution to real data, and the difficulty of the generated samples makes YOLOv5s harder to progress further in performance. In contrast, the proposed detection model gains higher performance and a rising tendency on performance. Using a higher resolution and adding the 1/4 scale feature can contribute to the model's progress to learn more effectively when training data is scarce. Among all three groups, the proposed detection model has higher AP0.5 and AP0.5:0.95 mean values than with YOLOv5s, confirming that the improvement of the proposed detection model is effective.

However, there are some cases when more generated samples lead to performance reduction. We attribute these cases to the deviation of real data distribution because OBIs on real rubbing images were not simply curved in grids. Instead, they tend to clump together in clusters of almost the same size. Nevertheless, these results prove that our rubbing generation network effectively alleviates the scarce data problem.

F. Comparison With Competitive Detection Models

We compare the proposed OBI detection model with several competitive object detection models on the OBI-Detection dataset. The experiment is conducted on all 8221 training samples and 913 testing samples. Only AP0.5:0.95 for the highest four models of AP0.5 are recorded.

Results are shown in Table VI. Compared with other models, the proposed OBI detection model is competitive in AP0.5 as well as AP0.5:0.95. This preferable result benefits from the multi-scale feature fusion structure with an enlarged resolution, the well-designed box loss, and the Soft-NMS dealing with overlapping. Compared with the classical two-stage model Faster R-CNN of a VGG16 backbone [11], and Faster R-CNN of a Conformer-FPN backbone [48], the proposed OBI detection model demonstrates the superiority in feature extraction by the well-designed backbone. Compared with classical one-stage models SSD300 [12], and DETR [32], the proposed OBI detection model also outperforms them for the enlarged resolution providing more semantic information and the Soft-NMS better at dealing with overlapping. Though the design of DETR can relieve it from anchor matching and NMS, the weakness in small objects inherited from the learnable fixed-length KEY sequences result in inadequacy for small and dense objects like OBI. Compared with YOLO series models, the proposed model outperforms YOLOv4 [16], resulting from better feature fusion ability from backbone and resolution. While YOLOv3 [15] and YOLOv5s [41] have slightly better performance of AP0.5, the proposed OBI detection model shows superiority in AP0.5:0.95 by improvements, representing a better overall performance of different positioning accuracy requirements. These results prove the effectiveness and superiority of the proposed detection model for the OBI detection task.

TABLE VI. Results of Competitive Models and the Proposed Model

Models	AP0.5	AP0.5:0.95
SSD300	0.734	-
Faster R-CNN(VGG16)	0.778	-
DETR	0.800	-
YOLOv4	0.860	-
Faster R-CNN(Conformer)	0.882	0.501
YOLOv3	0.906	0.535
YOLOv5s	0.906	0.529
the proposed model	0.902	0.581

G. Analysis of Methodology and Model

Function of Each Part. The proposed OBI rubbing image generation framework and the semantic-enhanced detection model can be seen as different parts of an OBI detection method under scarce labeled data regimes. The former generates additional training data, while the latter extracts more helpful information from limited data to boost performance.

Relevance with Semi-supervised Object Detection (SSOD). SSOD models [49], [50] share the same assumption on limited labeled training data availability(typically 1%, 2%, 5%, and 10%). The main difference is that they are trained with abundant unlabeled data by pseudo labels, which means they need to collect a massive amount of unlabeled data. In comparison, our method generates pseudo-training data without collecting additional unlabeled data.

Distribution Difference. In order to simplify the method, we take the uniform distribution as the prior of OBI positions, but the real distribution of OBI is complex. We visualized the OBI centers' position and frequency to show the difference in Fig. 10. In generated data, the OBI centers nearly fall into the square area with equal frequency. However, the real distribution is a narrow band region in the center, and the frequency of OBI fall in each grid position is biased. We observed this difference and regarded it as a distribution distortion.

Deceleration in the Boost of Performance. As shown in in Fig. 9, the more real labeled data is available, the more muted the growth in performance is by adding more generated data. We attribute this to the distribution distortion of generated data to real data in the test set. Similarly, in SSOD, an observation is that the performance boost slows down when the pseudo label of unlabeled data is distorted . These similar observations are caused by the distortion in distribution.

H. Ablation Study

1. Ablation of Box Loss

Ablation experiments on the proposed OBI detection model are conducted to verify the effectiveness of different $\mathcal{L}_{Box}$ losses in the proposed OBI detection model, including IOU, GIOU, DIOU, and CIOU, without other improvements. In addition, the OBI-detection dataset is randomly divided into 50% for training and 50% for testing as experimental groups.

Table VII shows the results of different box losses. The results in AP0.5 of the four loss methods are very close. When the training data is sufficient (9:1), IOU, GIOU, and DIOU all have an AP0.5 of 0.908, and CIOU has an AP0.5 of 0.906. While IOU, GIOU, and DIOU have an AP0.5 of 0.884, and the CIOU has an AP0.5 of 0.885 when the training data is less (5:5). This closeness results from the fact that the variation range of OBI aspect ratio variation is tiny. Moreover, a tiny change in the bounding box offset will not reflect the effect of other losses, so the CIOU is finally adopted as the box loss in the proposed detection model because the design of CIOU is based on improving IOU, GIOU, and DIOU.

TABLE VII. Results of Different Box Losses With Different Train: Test Splits

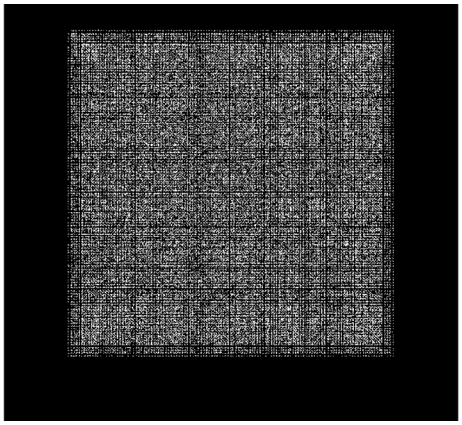
Train:Test	$\mathcal{L}_{Box}$	AP0.5	Train:Test	$\mathcal{L}_{Box}$	AP0.5
9:1	IOU	0.908	5:5	IOU	0.884
9:1	GIOU	0.908	5:5	GIOU	0.884
9:1	DIOU	0.908	5:5	DIOU	0.884
9:1	CIOU	0.906	5:5	CIOU	0.885

2. Ablation of 1/4 Feature, Resolution and Soft-NMS

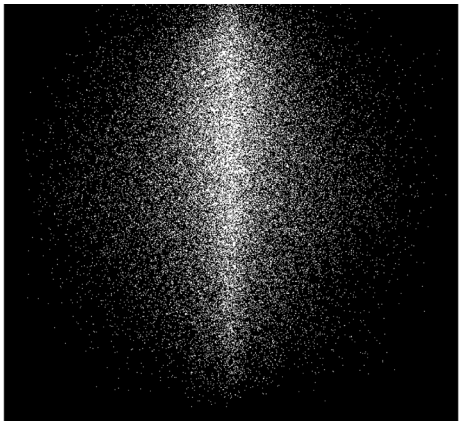
To verify the effectiveness of the other three improvements: (1) adding 1/4 feature (1/4 Feat), (2) using higher resolution (Higher Res), (3) Soft-NMS, ablation experiments are conducted on the OBI-detection dataset. The results are shown in Table VIII. The baseline without any improvement equals the original YOLOv5s model.

TABLE VIII. Results of Ablation Study of 1/4 Feature, Resolution, and Soft-NMS

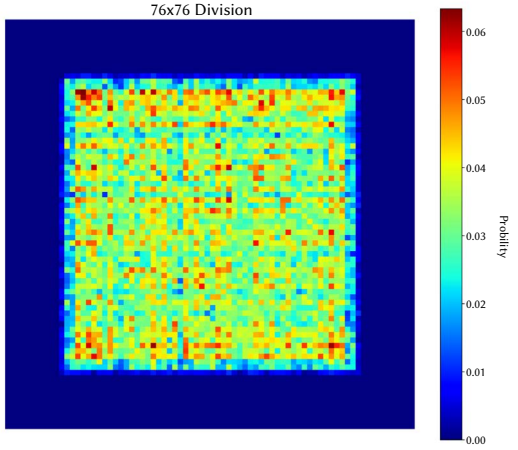
1/4 Feat	Higher Res	Soft-NMS	AP0.5	AP0.5:0.95
			0.906	0.529
		✓	0.900	0.571
	✓		0.909	0.534
✓			0.901	0.526
	✓	✓	0.906	0.580
✓		✓	0.907	0.530
✓	✓		**0.912**	0.538
✓	✓	✓	0.902	**0.581**



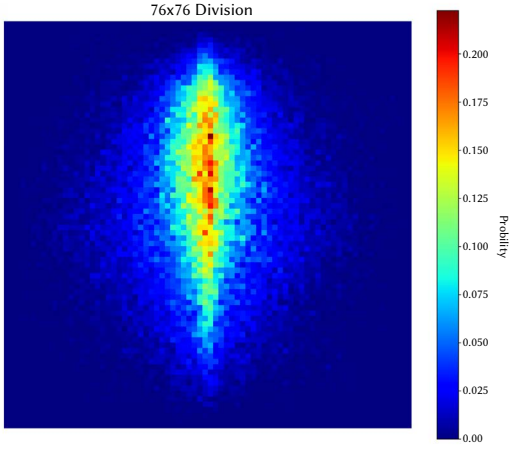
(a) Position of OBI centers in generated rubbing images



(b) Position of OBI centers in real rubbing images



(c) Frequency of OBI centers in generated rubbing images



(d) Frequency of OBI centers in real rubbing images

Fig. 10. Visualization of OBI center position and frequency. In (a) and (b), each OBI center is shown as a white point. In (c) and (d), the frequency of OBI center falls into equally divided grid cells is shown via the heatmap.

The combination of higher resolution with and 1/4 feature map is significant. However, when adding only the 1/4 feature map into the network structure without higher resolution, the poor predictions can deteriorate the performance due to insufficient semantic information in the 1/4 scale feature map. On the one hand, when only increasing the input resolution, the performance gets slightly better for enlarged resolution provides more information per position in feature maps. Therefore, the AP0.5 metric can improve by 0.6% by using higher input resolution and a 1/4 feature map for simultaneous detection without Soft-NMS.

Soft-NMS is also critical for improving AP0.5:0.95 but may result in a decline in AP0.5. Soft-NMS can effectively improve AP0.5:0.95. However, in some cases, due to some reserved inaccurate redundancy, Soft-NMS might lead to a decline in AP0.5. With increasing the input resolution, the improvement of Soft-NMS in AP0.5:0.95 after adding the feature map is more significant than that without adding the feature map. This phenomenon is because an enlarged input resolution is sufficient for the 1/4 feature to provide good predictions and leads to better performance.

Using higher input resolution and 1/4 feature map for fusion simultaneously, Soft-NMS can improve the AP0.5:0.95 metric from 0.529 to 0.581, which is the highest improvement result from better semantic information for fusion along with reducing overlapping. This result also proves the effectiveness of our OBI detection model.

3. Ablation of the $\mathcal{L}_{rubbing}$ Loss

To verify the rubbing generation network, the $\mathcal{L}_{rubbing}$ loss is also evaluated by substituting it with the L1 loss used in the rubbing generation framework. The loss records are shown in Fig. 11. It can be observed that the L1 loss calculated from the generated rubbing image and the real image is unstable, and the adversarial training cannot converge. This difficulty of convergence results from the noise on a real rubbing image caused by the random damage it had undergone. The noises on real rubbing images are not pixel-level stable. Hence, a pixel-level loss is sensitive to extensive random noise. By using the $\mathcal{L}_{rubbing}$ loss, the rubbing adversarial network is feasible in convergence and generate rubbing images with noise to a certain extent, which proves the effectiveness of $\mathcal{L}_{rubbing}$.

V. Conclusion

Recently, advancement in Oracle Bone Inscription (OBI) detection has emerged to assist OBI researchers. Nevertheless, the costly annotation of such noisy and complex data requires expertise, hindering its development. This paper proposes an OBI rubbing generation framework aiming to alleviate the scarcity of OBI detection training data with an OBI detection model aiming to provide better performance on OBI detection. The OBI rubbing generation framework is based on geometric algorithms and an adversarial network. First, it generates a rubbing base with controllable OBI placement only using single OBI images by geometric algorithms that construct multi-level mesh and morphological operations such as erosion on the mesh. Then the rubbing adversarial network serves as style transferring to obtain a more realistic rubbing image. The proposed OBI detection

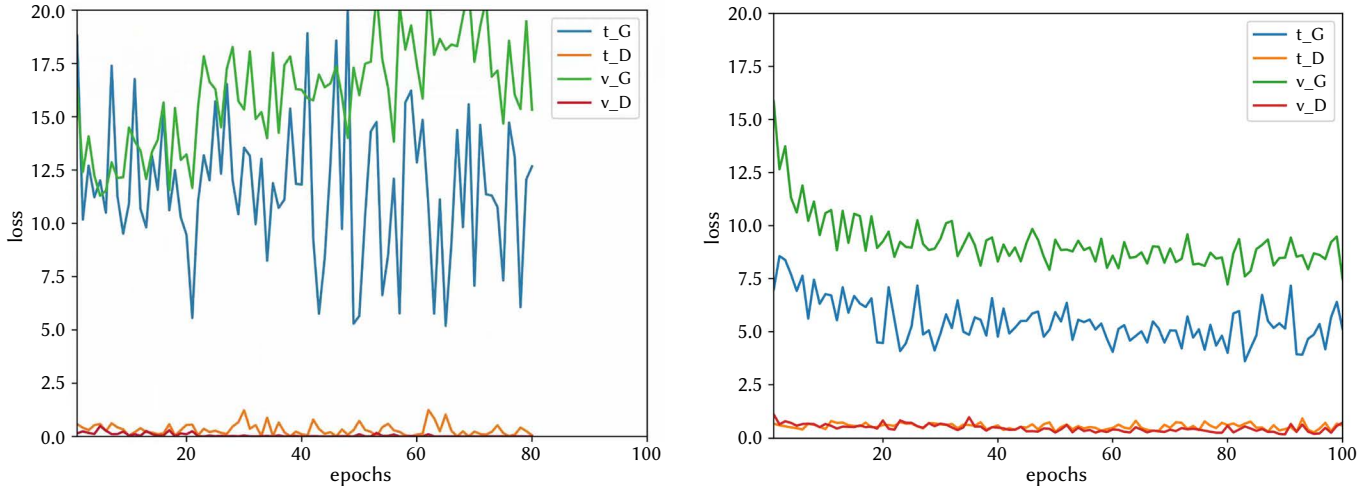


Fig. 11. Training records of rubbing adversarial network. (a) Using L1 loss instead of $\mathcal{L}_{rubbing}$ loss. (b) Using $\mathcal{L}_{rubbing}$ loss.

model with effective improving methods is capable of providing more semantic information for tiny OBI and deal with overlapping softly. The goal of the proposed method is to generate additional training data under a uniform position prior without collecting extra unlabeled data. Besides, it extracts semantic-enhanced information from limited data. Experiments of augmentation show the performance boost on several popular detection models, demonstrating the effectiveness of the proposed OBI rubbing generation framework. The proposed model achieves the best performance compared with other popular detection models. Visualization results verified the distribution difference between real and generated data.

Experiments demonstrate that the proposed method improves the performance of detection models when training data is scarce, while the proposed OBI detection model outperforms several competitive candidates. The proposed methods allow researchers to detect OBI precisely only with scarce training data. Even the simple uniform prior is adequate with limited data, considering how to design the prior of generating OBI rubbing is a promising future research.

Acknowledgment

The research was supported by the National Natural Science Foundation of China under Grant No.: 61976132 and 61991410.

This work is supported by Shanghai Technical Service Center of Science and Engineering Computing, Shanghai University.

References

[1] R. Pramanik, S. Bag, "Segmentation-based recognition system for handwritten Bangla and Devanagari words using conventional classification and transfer learning," *IET Image Processing*, vol. 14, no. 5, pp. 959–972, 2020, doi:10.1049/iet-ipr.2019.0208.

[2] R. Parashivamurthy, C. Naveena, Y. H. Sharath Kumar, "Sift and hog features for the retrieval of ancient kannada epigraphs," *IET Image Processing*, vol. 14, no. 17, pp. 4657–4662, 2020, doi:10.1049/iet-ipr.2020.0715.

[3] G. Li, J. Zhang, D. Chen, "F2pnet: font-to-painting translation by adversarial learning," *IET Image Processing*, vol. 14, no. 13, pp. 3243–3253, 2020, doi.org/10.1049/iet-ipr.2019.0476.

[4] K. Takashima, "Towards a more rigorous methodology of deciphering oracle-bone inscriptions," *T'oung Pao*, vol. 86, no. 5, pp. 363–399, 2000, doi.org/10.1163/15685320051072753.

[5] L. Meng, "Two-stage recognition for oracle bone inscriptions," in *International Conference on Image Analysis and Processing*, Catania, Italy, 2017, pp. 672– 682, Springer, doi:10.1007/978-3-319-68548-9_61.

[6] W. Han, X. Ren, H. Lin, Y. Fu, X. Xue, "Self-supervised learning of orc-bert augmentator for recognizing few- shot oracle characters," in *Proceedings of the Asian Conference on Computer Vision*, Kyoto, Japan, 2020, pp. 652–668, doi.org/10.1007/978-3-030-69544-6_39.

[7] Q. Jiao, Y. Jin, Y. Liu, S. Han, G. Liu, N. Wang, B. Li, F. Gao, "Module structure detection of oracle characters with similar semantics," *Alexandria Engineering Journal*, vol. 60, no. 5, pp. 4819–4828, 2021, doi.org/10.1016/j.aej.2021.03.072.

[8] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Columbus, OH, USA, 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[9] A. Laishram, K. Thongam, "Automatic classification of oral pathologies using orthopantomogram radiography images based on convolutional neural network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 69-77, 2022, doi: 10.9781/ijimai.2021.10.009.

[10] M. Adimoolam, S. Mohan, G. Srivastava, *et al.*, "A novel technique to detect and track multiple objects in dynamic video surveillance systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 112-120, 2022, doi:10.9781/ijimai.2022.01.002.

[11] S. Ren, K. He, R. Girshick, J. Sun, "Faster r- cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015, doi: 10.1109/TPAMI.2016.2577031.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, Amsterdam, The Netherlands, 2016, pp. 21–37, Springer, doi: 10.1007/978-3-319-46448-0_2.

[13] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 2016, pp. 779– 788, doi:10.1109/CVPR.2016.91.

[14] J. Redmon, A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA, 2017, pp. 7263–7271, doi:10.1109/CVPR.2017.690.

[15] J. Redmon, A. Farhadi, "Yolov3: An incremental improvement," 2018. [Online]. Available: https://arxiv.org/abs/1804.02767.

[16] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020. [Online]. Available: https://arxiv.org/abs/2004.10934.

[17] J. Xing, G. Liu, J. Xiong, "Oracle bone inscription detection: A survey of oracle bone inscription detection based on deep learning algorithm," in *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, Sanya, China, 2019, pp. 1–8, doi:10.1145/3371425.3371434.

[18] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez- Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, Vancouver, BC, Canada, 2018, doi:10.48550/arXiv.1710.09412.

[19] C. Shorten, T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019, doi: 10.1186/s40537-019-0197-0.

[20] T. DeVries, G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017. [Online]. Available: https://arxiv.org/abs/1708.04552.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014, doi: 10.1145/3422622.

[22] X. Yue, H. Li, Y. Fujikawa, L. Meng, "Dynamic dataset augmentation for deep learning-based oracle bone inscriptions recognition," *ACM Journal on Computing and Cultural Heritage*, vol. 15, no. 4, pp. 1–20, 2022, doi: 10.1145/3532868.

[23] G. Li, L. Wen, Z. Huang, R. Xia, Y. Pang, "Data augmentation and shadow image classification for shadow detection," *IET Image Processing*, vol. 16, no. 3, pp. 717–728, 2022, doi: 10.1049/ipr2.12377.

[24] K. He, X. Zhang, S. Ren, J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015, doi: 10.1109/TPAMI.2015.2389824.

[25] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, Santiago, Chile, 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.

[26] Z. Cai, N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, USA, 2018, pp. 6154–6162, doi: 10.1109/CVPR.2018.00644.

[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA, 2017, pp. 2117–2125, doi: 10.1109/CVPR.2017.106.

[28] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, Seattle, WA, USA, 2020, pp. 390– 391, doi: 10.1109/CVPRW50498.2020.00203.

[29] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, USA, 2018, pp. 8759– 8768, doi: 10.1109/CVPR.2018.00913.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.

[32] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, Glasgow, UK, 2020, pp. 213–229, Springer, doi: 10.1007/978-3-030-58452-8_13.

[33] L. Meng, B. Lyu, Z. Zhang, C. Aravinda, N. Kamitoku, K. Yamazaki, "Oracle bone inscription detector based on ssd," in *International Conference on Image Analysis and Processing*, 2019, pp. 126–136, Springer, doi:10.1007/978-3-030-30754-7_13.

[34] Y. Fujikawa, H. Li, X. Yue, C. Aravinda, G. A. Prabhu, L. Meng, "Recognition of oracle bone inscriptions by using two deep learning models," *International Journal of Digital Humanities*, pp. 1–15, 2022.

[35] G. Xu, "Research on Oracle Bone Radical Detection Based in Deep Learning of Semantic Analysis," M.S. thesis, Jiangxi Science and Technology Normal University, Nanchang, China, 2020.

[36] F. Liu, H. Li, J. Ma, S. Yan, P. Jin, "Research of automatic detection and recognition of oracle rubbings based on mask-rcnn," *Data Analysis and Knowledge Discovery*, vol. 5, no. 12, pp. 88–97, 2022, doi: 10.54097/0k5qen34.

[37] J. Xing, "Research of Oracle Bone Inscription Detection Based on Deep Convolution Neural Network," M.S. thesis, School of Computer Science and Engineering, Zhengzhou University, Zhengzhou, China, 2020.

[38] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, "Unpaired image- to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, Venice, Italy, 2017, pp. 2223–2232, doi:10.1109/ICCV.2017.244.

[39] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA, 2017, pp. 1125– 1134, doi: 10.1109/CVPR.2017.632.

[40] O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241, Springer, doi:10.1007/978-3-319-24574-4_28.

[41] G. J. et al., "ultralytics/yolov5: v6.0," 2021. [Online]. Available: https://github.com/ultralytics/yolov5.

[42] X. Zhu, B. Liang, D. Fu, G. Huang, F. Yang, W. Li, "Airport small object detection based on feature enhancement," *IET Image Processing*, vol. 16, no. 11, pp. 2863–2874, 2022, doi: 10.1049/ipr2.12387.

[43] J. Yu, Y. Jiang, Z. Wang, Z. Cao, T. Huang, "Unitbox: An advanced object detection network'," in *Proceedings of the 24th ACM international conference on Multimedia*, Amsterdam, The Netherlands, 2016, pp. 516–520, doi: 10.1145/2964284.296727.

[44] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Long Beach, CA, USA, 2019, pp. 658–666, doi: 10.1109/CVPR.2019.00075.

[45] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, New York City, NY, USA, 2020, pp. 12993–13000, doi: 10.1609/aaai.v34i07.6999.

[46] N. Bodla, B. Singh, R. Chellappa, L. S. Davis, "Soft- nms–improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, Venice, Italy, 2017, pp. 5561–5569, doi:10.1109/ICCV.2017.593.

[47] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, "Deformable detr: Deformable transformers for end- to-end object detection," in *International Conference on Learning Representations*, Virtual Event, Austria, 2021, doi: 10.48550/arXiv.2010.04159.

[48] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, Q. Ye, "Conformer: Local features coupling global representations for visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, 2021, pp. 367–376, doi: 10.1109/TPAMI.2023.3243048.

[49] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, T. Pfister, "A simple semi-supervised learning framework for object detection," 2020. [Online]. Available: https://arxiv.org/abs/2005.04757.

[50] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, P. Vajda, "Unbiased teacher for semi-supervised object detection," in *International Conference on Learning Representations*, Virtual Event, Austria, 2021.

Xiuan Wan

Xiuan Wan received the B.S. degree from the school of computer engineering and science, Shanghai University in 2022. He is currently pursuing the M.S. degree in the school of computer engineering and science, Shanghai University. His research interest is object detection.

Yuchun Fang

Yuchun Fang received the B.S. degree from the Central University of Nationalities in 1996, the M.S. degree from the Beijing Polytechnique University in 1999, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2003. She is currently a Full Professor with the School of Computer Engineering and Science, Shanghai University. From 2003 to 2004, she was a post-doctoral researcher at the France National Research Institute on Information and Automation (INRIA). Her current research interests include pattern recognition and image processing.

Jiahua Wu

Jiahua Wu received the B.S. degree from the school of computer engineering and science, Shanghai University in 2022. He is currently pursuing the M.S. degree in the school of computer engineering and science, Shanghai University. His research interest is domain adaptation.

Shouyong Pan

Shouyong Pan received the B.A. degree from Jilin University(archaeology, 1989), the M.A. degree from Nankai University (museology and history, 1993), and the Ph.D. degree from the Minzu University of China (ethnology, 1999). He is currently a distinguished Professor of anthropology and museology(Weichang Scholar), and director of university library, Shanghai University. From 2002 to 2004, he was a Harvard-yenching scholar at Harvard University, and from 2013 to 2014, he was a Fulbright scholar at George Washington University. His current research interests include Chinese culture, anthropology and cultural heritage studies.