

Explainable Coarse-to-Fine Ancient Manuscript Duplicates Discovery

Chongsheng Zhang^{*†}, Shuwen Wu^{*}, Yingqi Chen, Yi Men, Gaojuan Fan

Henan University, Lab for Archaeometry and Cultural Heritage Conservation, China
{Cszhang, ShuwenWu, Wingkeichanchan, Yi.Men, Fangaojuan}@henu.edu.cn

Yubin Jiang

Fudan University, Shanghai, China
jiangyubin@fudan.edu.cn

Matthias Aßenmacher, Christian Heumann

LMU Munich, Munich, Germany
{matthias, chris}@stat.uni-muenchen.de

Abstract

Ancient manuscripts are the primary source of ancient linguistic corpora. However, some of them exhibit duplications due to unintentional repeated publication or deliberate forgery. Notable examples include counterfeit fragments within the Dead Sea Scrolls and the repeated publication of Oracle Bones (OB). Identifying such duplicates is critical for both archaeological curation and historical research. We present **OBD-Finder**, an explainable coarse-to-fine framework for large-scale duplicate discovery in ancient manuscript collections, with a primary focus on Oracle Bones. OBD-Finder integrates unsupervised low-level keypoint matching with high-level character-centric content similarity to progressively filter and rank duplicate candidates, balancing retrieval accuracy with computational efficiency. Designed as a practical, transparent, and domain-expert-friendly tool, the system produces candidate duplicates and supports expert inspection and validation. In real-world deployment, OBD-Finder has enabled domain experts to uncover 63 previously unknown duplicate Oracle Bone pairs that had remained overlooked for decades. Extensive experiments also confirm that OBD-Finder achieves competitive retrieval accuracy while offering substantially superior efficiency and interpretability compared to existing approaches. Code and deployed system are publicly available: <https://github.com/cszhangLMU/OBD-Finder/>.

1 Introduction

Ancient manuscripts are primary sources for the study of early languages and civilizations. Yet many collections contain duplicate materials, arising from unintentional repeated publication, fragment rejoining and re-fragmentation during circulation, or deliberate forgery (Greshko, 2020). Identifying such duplicates is crucial for preventing

redundant research, correcting cataloging errors, and supporting reliable historical interpretation.

Oracle Bone Inscriptions (OBI), originating from the late Shang Dynasty over 3,000 years ago, present a particularly challenging case of this problem. After their rediscovery in 1899, Oracle Bones (OB) were rapidly and widely circulated among collectors and scholars all over the world, often without standardized documentation or provenance tracking. As a consequence, the same fragment was frequently published multiple times across different sources, while some physical fragments were further broken apart or rejoined during handling and circulation (Zhang et al., 2022). Today, more than 160,000 Oracle Bones are documented, with approximately 90% existing as fragments scattered across different collections all over the world. Consequently, manually identifying duplicates has become increasingly impractical, motivating the need for AI-assisted discovery tools.

OBI is carved writing, and its research relies primarily on rubbings and manual copies. Rubbings of the Oracle Bones faithfully capture physical surface texture but are often contaminated by background noise. Manual copies, by contrast, emphasize character structure and semantic content while suppressing background artifacts, offering cleaner input for computational analysis. Figure 1 presents examples for both formats (which are also cases of the new OB duplicates discovered by our model). In our experiments, we focus on manual copies, particularly the *OBI Copy Series* (Huang, 2022), which provides high-quality expert-produced copies for approximately 60,000 fragments. But in real-world deployment, we will test on both rubbing and manual copy images.

OB duplicates discovery is fundamentally a problem of similarity-based matching. Variations in scribal style, stroke thickness, and copying conventions introduce substantial intra-class visual variation. Furthermore, the scale of the corpus

^{*} Equal contribution

[†] Corresponding author: Chongsheng Zhang

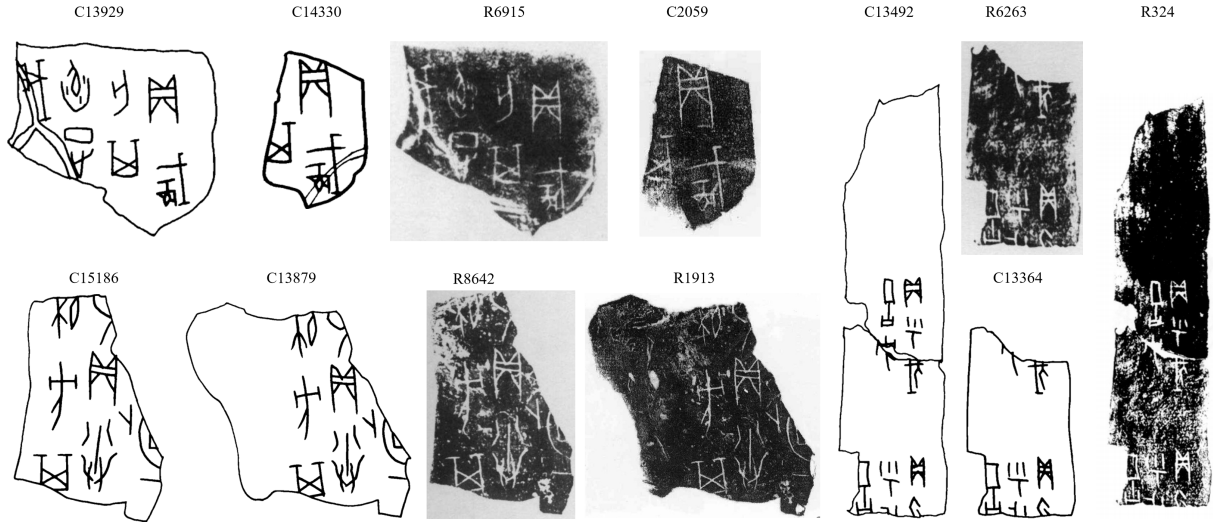


Figure 1: Three groups of new Oracle Bone duplicates discovered by our model, which have been missed by domain experts for decades. For each group of duplicate, we provide both the manual copies and rubbings of the Oracle Bones. We can see that, finding Oracle Bone duplicate is similarity-based matching, rather than exact matching, and there exists both one-to-one (e.g. the bottom left pair) and one-to-many matching (e.g. the top left pair).

gives rise to a large number of false positive candidates, making precision and efficiency jointly critical. These characteristics make OB duplicates discovery well suited to content-based image retrieval (CBIR) (Brown et al., 2020; Kim et al., 2020) or image matching (Jiang et al., 2024; Ren et al., 2025) paradigms, but they also impose stringent requirements on interpretability and efficiency to support expert-driven validation in real research workflows.

To address these real-world challenges, we present **OBD-Finder**, an explainable coarse-to-fine duplicate discovery system designed for large-scale Oracle Bone collections. OBD-Finder progressively integrates unsupervised low-level visual correspondence with high-level character-aware semantic similarity to filter and rank candidate duplicates, producing interpretable matching evidence. Deployed in active research workflows, OBD-Finder has already enabled domain experts to identify 63 previously unrecognized duplicate pairs that had remained overlooked for decades, demonstrating its substantial practical value.

2 Related Work

AI-enabled ancient manuscript fragments retrieval and rejoining have very important real-world applications (Seuret et al., 2020; Zhang et al., 2022, 2024). Our domain-specific Oracle Bone (OB) duplicates discovery problem can be essentially formulated as a image retrieval or image matching task. In the following, we provide a brief overview

of the content-based image retrieval (CBIR) and image matching techniques assessed in this study.

Content-based image retrieval. The Smooth-AP loss (Brown et al., 2020) provides a differentiable approximation for Average Precision to enable end-to-end training for ranking-based CBIR tasks. HashNet (Cao et al., 2017) adopts a sign activation function for binarizing the K-dimensional deep feature representation into K-bit binary hash code for retrieval tasks. In this paper, we use latest transformer-based implementation of HashNet (Dubey et al., 2022). HybridHash (He and Wei, 2024) is a hybrid deep hashing architecture combining self-attention and convolutional layers for enhancing image retrieval performance. Note that, since there also exists one-to-many mapping relationship in OB duplicates, contour-based retrieval methods are not applicable in our task.

Image matching. Image matching aims to establish correspondences between key points in two images depicting the same scene or object. SIFT (Scale-Invariant Feature Transform) (Lowe, 2004) is a classic algorithm for image matching. In recent years, deep learning based techniques have substantially improved the state-of-the-art in this field. LoFTR (Sun et al., 2021) utilizes Transformer self-attention and cross-attention mechanisms to establish coarse-level matches of the keypoints, followed by fine-level match on the cropped local windows for each coarse match. OmniGlue (Jiang et al.,

2024) leverages vision foundation models for boosting generalization to unseen domains. It also uses self- and cross-attention for establishing intra- and inter-image connectivity graphs to enhance correspondence estimation. MINIMA (Ren et al., 2025) is a latest image matching technique that contains a simple data engine for freely generating multiple modalities for images to pre-train modality invariant image matching model, achieving state-of-the-art performance.

Despite their technical advancement and substantial performance improvement, OmniGlue essentially depends on foundation models for accurate image matching, while MINIMA relies on “heavy” pre-training on generated multi-modal data, and both of them have significantly slower inference speed, compared to our proposed framework. Moreover, all the above image matching methods depend on low-level feature matching, but lack high-level semantic guidance. In comparison, in this work we seamlessly integrate unsupervised low-level feature matching and high-level character-centric semantic-aware content matching, offering outstanding retrieval and matching performance, fast inference speed, and strong interpretability for OB duplicates discovery.

3 Methodology

3.1 Framework

As can be seen from Figure 2, we propose a progressive coarse-to-fine Oracle Bone duplicate discovery framework, namely OBD-Finder, which combines unsupervised low-level keypoint matching with high-level, character-centric content-based image matching (in this work, content-based image matching refers to finding the correspondences between images based on their actual visual content, particularly the content of the ancient characters and their spatial relationships, with the goal of discovering the duplicate/repeated Oracle Bone images). Keypoint matching operates at low-level visual feature scale, which can prune out candidates with low degree of match in the initial stage, but lacks explicit semantic supervision and interpretability. Our framework bridges this gap by first grouping the keypoints based on their association with the character regions, then assesses the global matching degree between the two groups of keypoints via character-level visual content similarity computation. This dual matching mechanism enhances Oracle Bone duplicates discovery

accuracy through a progressive coarse-to-fine refinement manner, by effectively and seamlessly integrating both low-level keypoint and high-level character-based semantic cues, resulting in more accurate and semantic-aware image matching. Our framework consists of four subsequent steps:

1. **Feature Extraction.** We perform unsupervised keypoints extraction on the OBs using a pre-trained model (DeTone et al., 2018).
2. **Feature Matching.** We next apply unsupervised keypoints mapping between the two OB images using a pre-trained model (Lindemberger et al., 2023). Candidate with low overall matching degrees will be filtered out.
3. **Coordinate Alignment.** After obtaining the correspondence between the keypoints in feature matching, we apply affine transformations for each image pair, in which we map the coordinates of the image with fewer feature points to the other image.
4. **Character-level Content Similarity.** We localize the OBI characters in each image, using a text detector (Zhou et al., 2017). Given that the coordinate systems of two images are aligned, for each character in the smaller image, we search for the overlapped characters in the counterpart image, next compute the content similarity between them, using a simple Siamese network model (Taigman et al., 2014).

3.2 Key Features

OBD-Finder for ancient manuscript duplicates discovery has four primary characteristics:

1. It is a progressive coarse-to-fine framework that seamlessly proceeds from low-level keypoints matching to high-level semantic-aware content similarity computation, resulting in very accurate OB duplicates discovery.
2. It is a transparent framework with strong interpretability.
3. It is unsupervised and almost training-free, only requiring little annotation.
4. It is highly efficient, compared to state-of-the-art image matching methods, which we will demonstrate in the empirical studies.

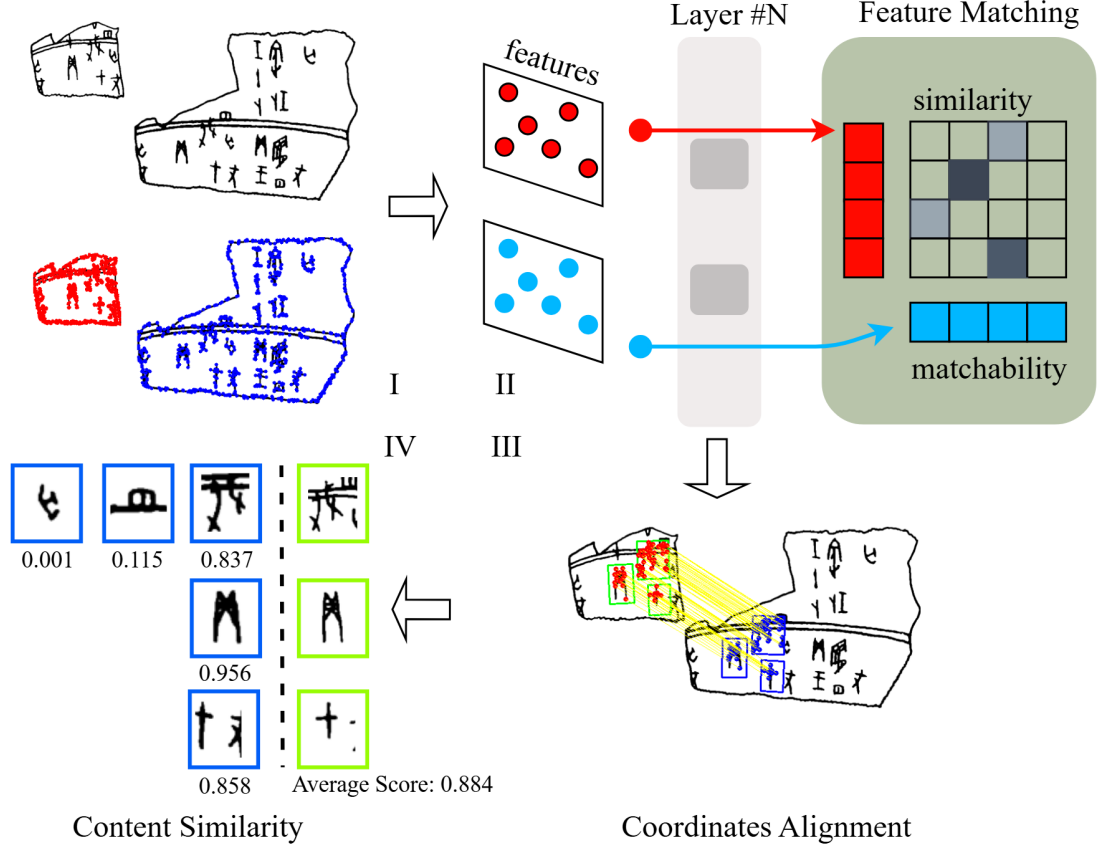


Figure 2: The overall framework of OBD-Finder for Oracle Bone duplicates discovery.

4 Experiments

4.1 Experimental Setup

Dataset. We use the 60,000 manual Oracle Bone copies from the Oracle Bone Inscription Copy Series (Huang, 2022) as the retrieval database, which have been divided into 25 OB groups by the OBI community based on dating and writing style. Three OBI domain experts who are co-authors of this work (Prof. Yi Men and Prof. Yubin Jiang, and Ms. Yingqi Chen) have manually collected 150 pairs of OB duplicates published in the OBI literature, which belong to different OB groups and are used as the query images (with ground-truths). For each pair of images, we perform OB duplicate checking in the corresponding OBI group. For OBI character localization using EAST (Zhou et al., 2017) and character-level content matching using Siamese network (Taigman et al., 2014), OBI domain experts have manually annotated the character regions and categorization for 500 OBs.

Evaluation Metrics. We use two evaluation metrics: Recall@K and a simplified mean reciprocal rank measure MRR@K. Recall@K measures how many relevant items were successfully retrieved in

the Top-K results, but doesn’t consider their rankings. To address this limitation, we also adopt a modified version of MRR@K, since our dataset only contains one correct match per query, we simply average the rank position of the correct answer in the Top-K results across all query images, abbreviated as Rank@K.

4.2 Results

We first compare our model with content-based image retrieval (CBIR) methods. As reported in Table 1, OBD-Finder consistently obtains the best recall performance than the CBIR methods.

We next compare our model with state-of-the-art image matching techniques. As shown in Table 2, OBD-Finder exhibits recall performance on par with latest image matching methods. In Figure 3, we showcase image matching results of different methods. Moreover, as reported in Tables 3, our model obtains the best Rank@K scores in the Top-5 and Top-15 retrieval results, demonstrating that it excels at prioritizing correct matches.

In Tables 4, we observe that OBD-Finder has significantly faster inference speed than state-of-the-art image matching algorithms (40 times faster),

Table 1: Comparing OBD-Finder with CBIR methods (Recall@K performance).

Method	Recall@1	Recall@5	Recall@10	Recall@15	Recall@20
Smooth-AP (Brown et al., 2020)	34.1	62.9	73.8	78.8	82.8
Proxy-Anchor (Kim et al., 2020)	71.5	77.0	84.1	86.6	91.9
HashNet (Dubey et al., 2022)	57.6	63.8	69.3	78.4	85.8
HybridHash (He and Wei, 2024)	60.8	66.8	73.2	79.3	88.3
Ours (OBD-Finder)	80.0	85.3	90.4	94.3	98.0

Table 2: Comparing OBD-Finder with image matching methods (Recall@K).

Method	Recall@1	Recall@5	Recall@10	Recall@15	Recall@20	Recall@25
SIFT (Lowe, 2004)	33.3	40.0	46.8	53.3	66.6	73.2
LoFTR (Sun et al., 2021)	73.6	75.4	81.3	86.0	92.3	98.3
OmniGlue (Jiang et al., 2024)	82.5	86.0	91.2	95.6	98.2	100
MINIMA (Ren et al., 2025)	84.4	90.5	94.7	98.2	100	100
Ours (OBD-Finder)	80	85.3	90.4	94.3	98	100

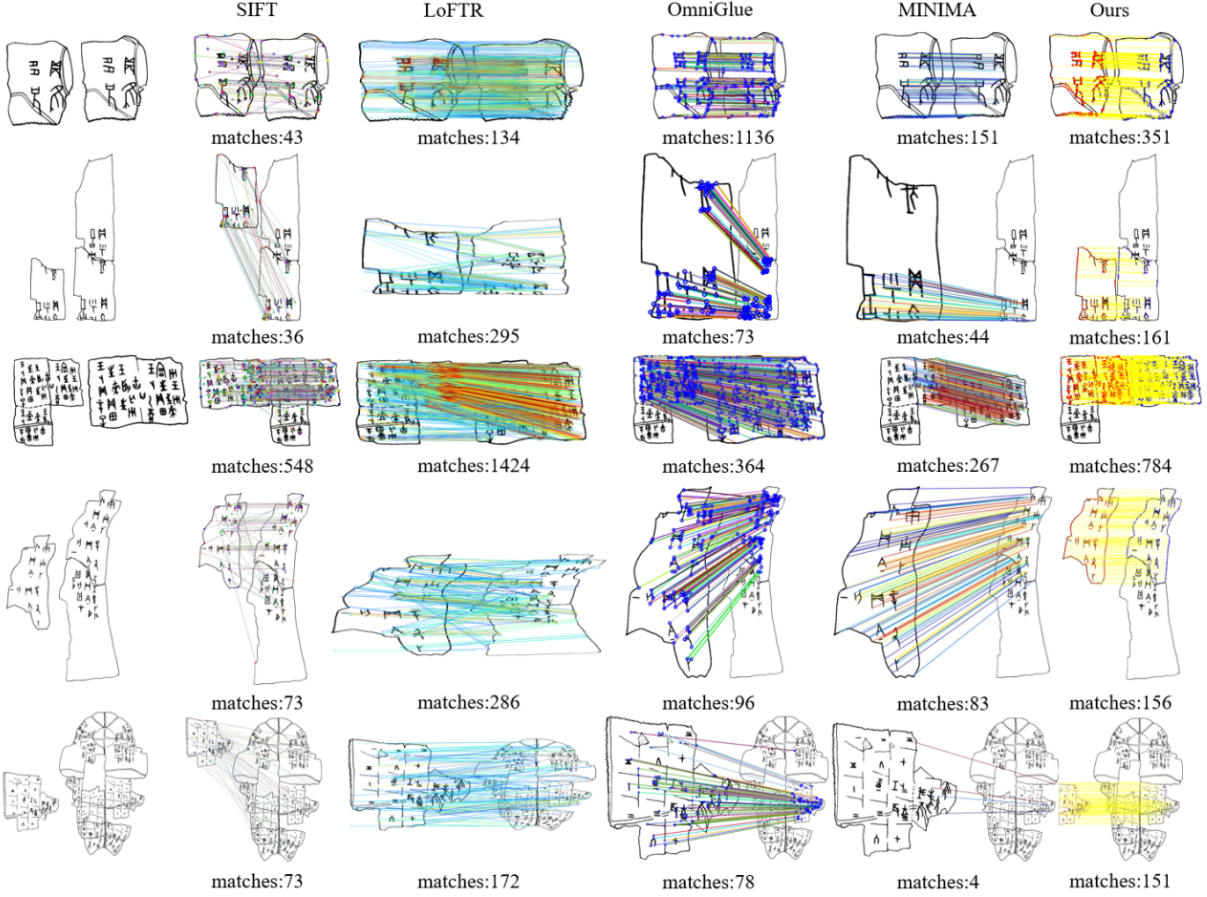


Figure 3: Case studies for image matching results of different methods.

Table 3: Comparing OBD-Finder with image matching methods (Rank@K).

Method	Rank@5	Rank@10	Rank@15	Rank@20	Rank@25
SIFT (Lowe, 2004)	1.6	2.70	4.16	6.15	7.74
LoFTR (Sun et al., 2021)	1.09	1.88	2.74	3.76	5.22
OmniGlue (Jiang et al., 2024)	1.13	1.51	2.05	2.52	2.89
MINIMA (Ren et al., 2025)	1.36	1.66	2.07	2.38	2.38
Ours (OBD-Finder)	1.06	1.63	2.00	2.61	2.93

Table 4: Comparisons on Inference speed, FPS, and GPU memory consumption.

Method	Recall@20	Inf. Speed (s/pair)	FPS (pair/s)	GPU (MiB)
SIFT (Lowe, 2004)	66.6	0.017	59	N/A
LoFTR (Sun et al., 2021)	92.3	3.6	0.28	3502
OmniGlue (Jiang et al., 2024)	98.2	45	0.02	23612
MINIMA (Ren et al., 2025)	100	1	1	14890
Ours (OBD-Finder)	98	0.021	47.62	5215

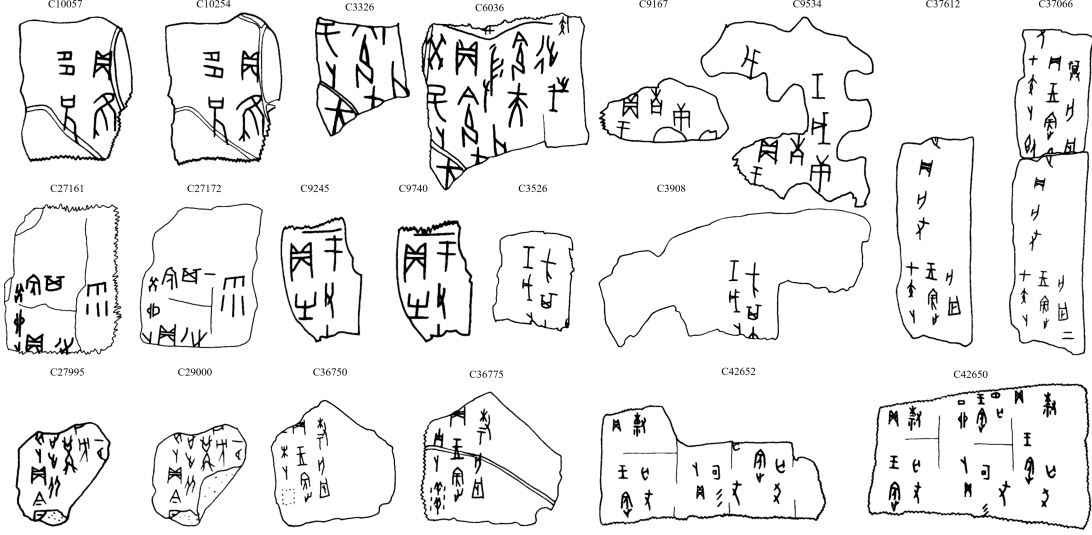


Figure 4: Ten groups of new Oracle Bone duplicates discovered by OBD-Finder.

but with substantially less GPU consumption (1/3 of their GPU usages). Therefore, our model is accurate and efficient at prioritizing correct matches.

5 Real-World Deployment

Our OBD-Finder system is available online¹. Given a query Oracle Bone image, the system retrieves candidate duplicates from a specified collection and displays ranked results along with interpretable matching evidence.

In real-world deployment, OBD-Finder has successfully discovered 63 pairs of new Oracle Bone manual duplicates, all of which have been verified by the OBI community. Besides Figure 1, in Figure 4, we present ten more pairs of new Oracle Bone duplicates that we have discovered in real-world deployment, including five pairs of one-to-one matching and another five pairs of one-to-many matching. The complete list of 63 newly discovered Oracle Bone duplicate pairs are publicly available². They have been shared with the OBI research community and received very positive feedback.

¹<https://cszhangai-obd-finder2.hf.space/>.

²<https://drive.google.com/drive/folders/1mWgvv76nDDTCEUpeSWz-pU2dMymcevP?usp=sharing>.

Moreover, OBD-Finder also supports duplicates discovery on Oracle Bone rubbing images. Very recently, using the rubbing images from *Oracle Bone Collection Supplement*, OBD-Finder has successfully found 165 pairs of duplicates^{3 4}.

6 Conclusion

For real-world Oracle Bone duplicates discovery, we presented OBD-Finder, which integrates unsupervised keypoint matching with character-centric semantic similarity, and achieves accurate, efficient, and interpretable duplicate retrieval. Real-world deployment has enabled domain experts to identify 63 previously unknown Oracle Bone duplicate pairs that had remained overlooked for decades.

³<https://www.xianqin.org/blog/archives/24848.html>. Notice that, these 165 groups might contain many existing duplicates results. Since the complete list of existing duplicates results on this collection is not available, our OBI domain experts are checking different books and articles to figure out how many of these 165 groups are new duplicates; till now, 23 of them have been confirmed as new duplicates.

⁴https://drive.google.com/drive/folders/1DWNiBA_I3BVzQH45WFG0hLC0jLcp_1q?usp=sharing.

Acknowledgments

This work was partially supported by the Major Program of National Social Science Foundation (No.23&ZD309), the National Natural Science Foundation of China (62676141, 62250410371), the MOE Liberal Arts and Social Sciences Foundation (No.23YJAZH210), and Henan Provincial Center for Outstanding Overseas Scientists (No.GZS2025004). Matthias Aßenmacher received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure-NFDI 27/1-460037581-BERD@NFDI. We thank the suggestions and help from Prof. João Gama as well. Last but not least, we thank all the anonymous reviewers, area chairs and program committee members for their constructive feedback throughout the review process.

Ethics Statement

We affirm that our research adheres to the [ACL Ethics Policy](#). This work uses publicly available datasets and contains no personally identifiable information. We declare that there are no conflicts of interest that could potentially influence the outcomes, interpretations, or conclusions of this research. All funding sources supporting this study are acknowledged in the acknowledgments section. We have diligently documented our methodology, experiments, and results, and we commit to sharing our code, data, and other relevant resources to enhance reproducibility.

References

- Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. 2020. Smooth-AP: Smoothing the path towards large-scale image retrieval. In *ECCV 2020*, pages 677–694.
- Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S. Yu. 2017. HashNet: Deep learning to hash by continuation. In *ICCV 2017*, pages 5609–5618.
- Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2018. SuperPoint: Self-supervised interest point detection and description. In *CVPR Workshops 2018*, pages 224–236.
- Shiv Ram Dubey, Satish Kumar Singh, and Wei-Ta Chu. 2022. Vision transformer hashing for image retrieval. In *ICME 2022*, pages 1–6.
- Michael Greshko. 2020. [Dead sea scrolls at the museum of the bible are all forgeries](#).
- Chao He and Hongxi Wei. 2024. HybridHash: Hybrid convolutional and self-attention deep hashing for image retrieval. In *ICMR 2024*, pages 824–832.
- Tianshu Huang. 2022. *Oracle Bone Inscription Copy Series*. Peking University Press.
- Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and André Araújo. 2024. OmniGlue: Generalizable feature matching with foundation model guidance. In *CVPR 2024*, pages 19865–19875.
- Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. 2020. Proxy anchor loss for deep metric learning. In *CVPR 2020*, pages 3235–3244.
- Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. 2023. LightGlue: Local feature matching at light speed. In *ICCV 2023*, pages 17581–17592.
- David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110.
- Jiangwei Ren, Xingyu Jiang, Zizhuo Li, Dingkan Liang, Xin Zhou, and Xiang Bai. 2025. MINIMA: Modality invariant image matching. In *CVPR 2025*, pages 23059–23068.
- Mathias Seuret, Anguelos Nicolaou, Andreas Maier, Vincent Christlein, and Dominique Stutzmann. 2020. ICFHR 2020 competition on image retrieval for historical handwritten fragments. In *ICFHR 2020*, pages 216–221.
- Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. 2021. LoFTR: Detector-free local feature matching with transformers. In *CVPR 2021*, pages 8922–8931.
- Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Lior Wolf. 2014. DeepFace: Closing the gap to human-level performance in face verification. In *CVPR 2014*, pages 1701–1708.
- Chongsheng Zhang, Bin Wang, Ke Chen, Ruixing Zong, Bofeng Mo, Yi Men, George Almpanidis, Shanxiong Chen, and Xiangliang Zhang. 2022. Data-driven oracle bone rejoining: A dataset and practical self-supervised learning scheme. In *SIGKDD 2022*, pages 4482–4492.
- Yuqing Zhang, Hangqi Li, Shengyu Zhang, Runzhong Wang, Baoyi He, Huaiyong Dou, Junchi Yan, Yongquan Zhang, and Fei Wu. 2024. LLMCO4MR: LLMs-aided neural combinatorial optimization for ancient manuscript restoration from fragments with case studies on dunhuang. In *ECCV 2024*, pages 253–269.
- Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. EAST: an efficient and accurate scene text detector. In *CVPR 2017*, pages 2642–2651.