Modality Adaptation Representation Learning for Historical Document Image Retrieval

Jing Zhang, Hongxi Wei†, Qing Zhang

Inner Mongolia University

Hohhot, China

Abstract

Word spotting is a fundamental task in Historical Document Image Retrieval (HDIR), categorized into query-by-example (QBE) and query-by-string (QBS) based on the query modality. QBS word spotting, as a cross-modal retrieval task, poses a significant challenge due to the substantial modality gap between image and text representations. Existing methods typically embed both image and text features into a common space and directly constrain their alignment, overlooking the substantial distribution gaps between modalities. Moreover, out-of-vocabulary (OOV) words present an additional challenge, requiring the model to possess strong generalization capability. To address these challenges, we propose a novel bi-directional modality adaptation representation learning framework, termed MARK. Specifically, the model adopts CLIP as the backbone to extract image and text features, providing a strong baseline for cross-modal representations. A bi-directional modality adaptation mechanism based on the Kolmogorov-Arnold network (KAN) is designed to capture both shared and complementary information between image and text representations. Furthermore, Pyramid Histograms of Characters (PHOC) is introduced to enhance fine-grained text representation and overcome the limitations of CLIP's text encoder. Extensive experiments on Monglian and Manchu benchmark datasets, including Kanjur, Geser and Manchu, demonstrate that MARK establishes a new state-of-the-art in HDIR. Our code is available at https://github.com/cszjing/MARK.

Keywords: Historical document image retrieval; CLIP; KAN; PHOC; Modality adaptation

1. Introduction

Word spotting, a fundamental task in Historical Document Image Retrieval (HDIR), has been widely applied to various document image processing tasks, including document visual question answering [9, 16, 29] and historical document image recognition [2, 10, 17]. It was introduced as an alternative to HDIR, especially for historical documents that often suffer from severe degradations such as handwriting deformation and heavy noise. In such cases, the conventional recognition-then-retrieval paradigm leads to poor retrieval performance due to unreliable transcription. Word spotting can be categorized into Query-by-Example (QBE) and Query-by-String (QBS) word spotting according to the query modality. QBE retrieves word instances by using a

†Corresponding author: Hongxi Wei (Email: cswhx@imu.edu.cn; ORCID: 0000-0002-2570-4544)

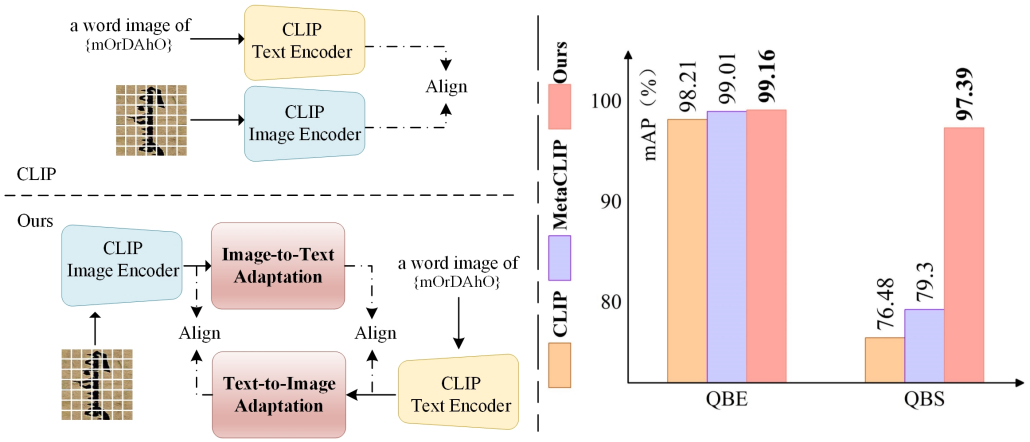


Fig. 1. Comparison with the original CLIP model (top left). We adapt it to the HDIR setting using a bi-directional modality adaptation framework (bottom left). Our MARK model effectively reduces the modality gap and significantly outperforms CLIP on QBS word spotting (right).

given word image as a query, identifying visually similar instances in the document image set. In contrast, QBS retrieves word images based on a provided text string, and returns the corresponding word images that match the given text. Notably, QBS word spotting is regarded as a cross-modal retrieval task, as it involves a significant gap between the image and text.

Early work primarily focused on QBE word spotting by learning low-level image features [6, 15, 21, 23, 30, 31, 38, 39]. With the advancement of deep learning, several learning-based approaches [7, 32–36, 43] have emerged to extract deep visual features, significantly improving word spotting performance. To address both QBE and QBS word spotting, Almazán et al. [2] are the first to introduce Pyramidal Histogram of Characters (PHOC) as a common space to embed both image and text features. Building upon this idea, subsequent works [3, 25, 26, 41] have explored various learning-based methods to directly map images into PHOC representations. However, the substantial modality gap between image and text cannot be adequately addressed through direct mapping, constraining the improvement in retrieval performance.

Moreover, one of the primary challenges faced by word spotting is the out-of-vocabulary (OOV) problem, which is unavoidable in real-world applications. OOV word spotting is generally analogous to zero-shot learning [4], as both require the model to handle previously unseen classes without explicit training examples. As pointed out in [28], most words in natural languages can be decomposed into sequences of characters. While the vocabulary of words is virtually unbounded, the character set of any given language remains relatively small and fixed. This observation implies that, in theory, as long as the complete character set of a language is covered in the training stage, it is possible to spot previously OOV words by modeling. However, despite its importance, there is still a notable lack of benchmark datasets specifically designed for evaluating OOV word spotting. Furthermore, existing studies seldom quantify the proportion of OOV words in test datasets, nor do they thoroughly assess the performance gains on OOV words.

To break this bottleneck, Wei et al. [35] were the first to introduce the Mongolian historical Kanjur dataset, which includes a specific partitioning strategy aimed at evaluating model performance on OOV words and provides baseline results for future research. Building on this

foundation, Zhang et al. [41] proposed a larger Mongolian historical Geser dataset, which further advances the study of OOV word spotting by introducing multiple OOV splits from 20% to 80% levels. Based on these two OOV-specific datasets, this paper further explores methods to enhance model generalization and retrieval performance on OOV words.

Recently, CLIP [20] has demonstrated impressive domain generalization capabilities and has been widely adopted in various vision-language tasks, ranging from Image Classification [1, 18, 19] to Image Retrieval [12, 24, 42]. Its ability to align image and text embeddings in a shared semantic space makes it a promising candidate for cross-modal applications. However, as shown in our preliminary results in Fig. 1, CLIP performs poorly on QBS word spotting. This observation is consistent with the original CLIP paper, which acknowledges its limitations in fine-grained text-related tasks. In this paper, word spotting can be regarded as a fine-grained task where even a single character difference can lead to a negative result. Therefore, effectively adapting CLIP for such fine-grained retrieval remains a significant challenge.

To tackle the aforementioned challenges, we propose a bi-directional modality adaptation representation learning framework, named MARK. First, CLIP is employed as the backbone to extract image and text features, enhancing the model's generalization capability to address the challenges posed by OOV words. Then, a bi-directional modality adaptation based on the Kolmogorov-Arnold network (KAN) [14] is designed to capture shared and complementary information between image and text representations, thereby reducing the modality gap. The modality adaptation mechanism consists of an Image-to-Text Adaptation (ITA) and a Text-to-Image Adaptation (TIA). In particular, PHOC encoding is incorporated into the TIA process to enhance fine-grained text representation and address the limitations of CLIP's text encoder.

In summary, our main contributions are as follows:

- We propose MARK, a novel CLIP-based bi-directional modality adaptation representation learning framework that effectively bridges the modality gap and tackles OOV word spotting challenges.
- We design a new bi-directional modality adaptation mechanism based on KAN to capture shared and complementary information between image and text, thereby improving cross-modal translation.
- We introduce PHOC into the text-to-image adaptation process to enhance fine-grained text feature and overcome the limitation of CLIP's text encoder.
- Extensive experiments conducted on three historical document datasets demonstrate that MARK establishes a new state-of-the-art in both QBE and QBS word spotting.

2. Related Work

2.1. Word spotting

Early works on word spotting mainly focus on QBE word spotting, exploring effective representations of word images. Various methods have been proposed to represent the low-level features of word images, including profile features [21, 30, 31, 38, 39], Bag of Visual Words (BoVW) [6, 8, 15, 23]. With the advancement of deep learning, various learning-based methods [7, 33, 35, 36, 43] have been explored for word image representation. Nevertheless, in real-world applications, more flexible querying mechanisms are required. For instance, QBS offers a more user-friendly alternative by allowing users to input queries directly via the keyboard.

Driven by this practical demand, subsequent works have focused on simultaneously addressing QBE and QBS word spotting by learning a common space for both image and text representations. Almazán et al. [2] are the first to propose PHOC, a rule-based representation that encodes text at the character level. Building upon this, several PHOC-based approaches have utilized deep learning techniques to map images to PHOC representations, such as PHOCNet [25], PUNet [3], and HENet [41]. However, these methods rely on manually extracted text features, which motivated subsequent works [10, 17] to jointly embed both images and text into a common space with deep learning methods. Despite these advancements, directly aligning image and text representations in a common space still fails to effectively address the substantial modality gap. Unlike previous methods, this paper explore a modality-adaptive strategy that constructs a flexible bridge between modalities.

Despite advances in representation learning, word spotting remains vulnerable to the out-of-vocabulary (OOV) problem, where unseen words during training hinder robust retrieval. OOV words require the model to generalize to entirely unseen textual instances, making the task analogous to zero-shot learning. In response to the lack of OOV-specific benchmarks, Wei et al.[35] introduced the Mongolian historical Kanjur dataset, which explicitly partitions data to evaluate model performance under OOV conditions. Building upon this, Zhang et al.[41] constructed the larger Geser dataset with progressively structured OOV splits (20%, 40%, 60%, 80%), further promoting research on OOV word spotting. Motivated by these efforts, this work is dedicated to advancing OOV word spotting by enhancing model generalization.

2.2. Contrastive Language-Image Pretraining

CLIP [20], developed by OpenAI, constructs a joint image-text embedding space through contrastive learning on 400 million image-text pairs, thereby achieving strong zero-shot generalization capability. However, CLIP remains limited in handling fine-grained tasks, such as handwritten character recognition, as acknowledged in the original paper. This limitation is also evident in the field of Visual Document Understanding, Li et al. [11] have highlighted that CLIP primarily learns global representations to enable instance-level discrimination, but struggles to extract fine-grained features in text-rich scenarios. Similarly, in scene text retrieval tasks, Zeng et al. [40] introduce dynamic attention shift and text knowledge probing methods to alleviate CLIP's limitations in text perceptual scale, making it more suitable for fine-grained text retrieval.

Furthermore, since CLIP's training data and filtering process have not been disclosed, MetaCLIP [37] adopts a metadata-based data filtering strategy to enhance the transparency and reproducibility of language-image pretraining. Nevertheless, existing CLIP models and their variants still face significant limitations in fine-grained text-related tasks. Word spotting, which requires distinguishing subtle differences at the character level, is also inherently a fine-grained task.

Therefore, whether we can improve word spotting performance by incorporating character information is the main concern of our work. However, unlike the aforementioned PHOC-based works that directly use PHOC as the text representation, this paper focuses on integrating PHOC into the feature extracting from the CLIP text encoder, which not only enables learning deep representations but also incorporates character-level information.

2.3. Modality Adaptation

Modality adaptation aims to bridge the gap between different data modalities, which is particularly useful in cross-modal tasks where direct correspondence between modalities is limited or absent. The work in [22] improves existing domain adaptation techniques to mitigate modality

gaps, effectively encoding shared and complementary knowledge across multimodal data while preserving the discriminative ability of each modality.

A key challenge in modality adaptation lies in learning flexible and expressive mappings between modalities. Apart from classical Multi-Layer Perceptrons (MLP), Kolmogorov-Arnold Networks (KAN) [14] is inspired by the Kolmogorov-Arnold representation theorem and designed to provide flexible and interpretable representations of high dimensional functions. Subsequent researches, such as EfficientKAN[1] and FastKAN[2], have focused on improving the computational efficiency of KAN models. Notably, FastKAN [13] approximates B-splines using Gaussian radial basis functions (RBF), which significantly improves the computational efficiency of KAN.

Motivated by these advancements, our work aims to design a KAN-based modality adaptation mechanism to capture shared and complementary information between image and text.

3. The Proposed Method

Based on the above analysis, we propose MARK, a bi-directional modality adaptation framework that takes advantage of MetaCLIP's strong generalization capability and a KAN-based modality adaptation mechanism, aiming to effectively address the modality gap between image and text. As presented in Fig. 2, MARK first extracts image and text features through MetaCLIP, and then feeds them into two modality adaptation modules, including Image-to-Text Adaptation (ITA) and PHOC-based Text-to-Image Adaptation (PTIA). The ITA module translates image features into the textual feature space to bridge the modality gap. In contrast, the PTIA module translates textual features into the image feature space. Additionally, PTIA enhances text representations by incorporating PHOC, which effectively improves the model's sensitivity to character-level information. More details will be described in the following subsections.

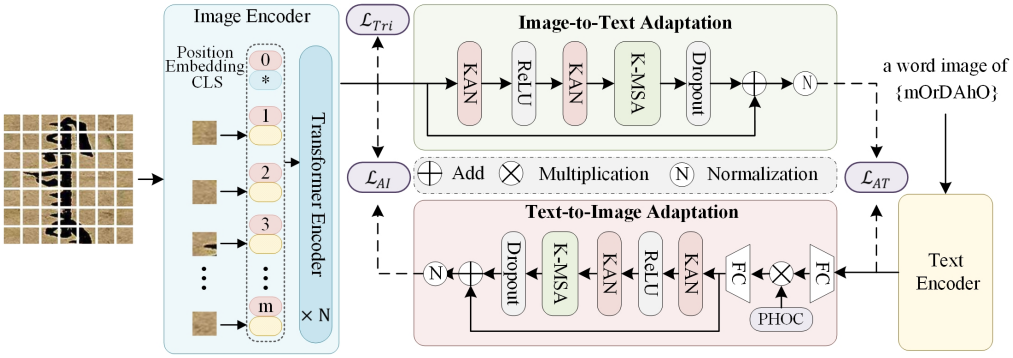


Fig. 2. Overview of our bi-directional modality adaptation framework, MARK, which comprises two key components: Image-to-Text Adaptation (ITA) and PHOC-based Text-to-Image Adaptation (PTIA).

3.1. Image-to-Text Adaptation

The Image-to-Text Adaptation (ITA) module is designed to bridge the modality gap by translating image representations into the textual feature space. ITA is based on FastKAN [13], which

[1] ff
[2] ff

approximates the 3-order B-spline basis using radial basis functions (RBFs) with Gaussian kernels. Mathematically, an RBF network with N centers can be represented as:

$$F_{KAN}(x) = \sum_{i=1}^{N} w_i \phi\left(\|x - c_i\|\right) \tag{1}$$

where w_i are the adjustable weights or coefficients and ϕ is the radial basis function, which depends on the distance between the input x and a center c_i. The Gaussian function is one of the most widely used RBFs, defined as:

$$\phi(r) = \exp\left(-\frac{r^2}{2h^2}\right) \tag{2}$$

where r is the radial distance, and h is a parameter that controls the width or spread of the function.

As depicted in Fig. 2, ITA consists of two KANs with an intermediate ReLU activation, followed by a KAN-based Multi-Head Self-Attention (K-MSA) mechanism, dropout, and residual connections with layer normalization. The K-MSA (Fig. 3) differs from traditional Multi-Head Self-Attention by introducing a gating mechanism that operates continuously through a sigmoid function for adaptive attention across different dimensions. Given query (Q), key (K), value (V) and gate (G), the attention computation follows:

$$Att = Softmax\left(\frac{F_{KAN}(Q) \cdot F_{KAN}(K)^T}{\sqrt{d}}\right) \cdot F_{KAN}(V) \tag{3}$$

$$Out = Att \cdot \sigma\left(F_{KAN}(G)\right) \tag{4}$$

where σ is the sigmoid activation function. Note that since this is a self-attention mechanism, Q, K, V and G are all derived from the same input feature.

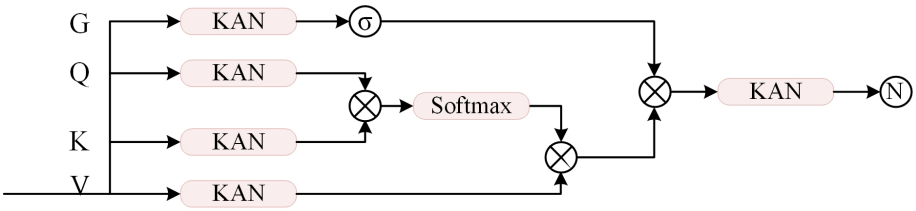


Fig. 3. Architecture of the proposed K-MSA block, incorporating a gating mechanism for adaptive attention across different dimensions.

3.2. Text-to-Image Adaptation

The Text-to-Image Adaptation module extends ITA, which introduces PHOC [2] to capture character-level information to compensate for CLIP's limitations. PHOC is a rule-based method for character-level text representation. Taking "mOrDAhO" as an example, its PHOC representation at levels 1, 2, and 3 is depicted in Fig. 4. PHOC encodes whether a character is present in a specific segment of the word. For example, at the 2nd level, PHOC captures whether a

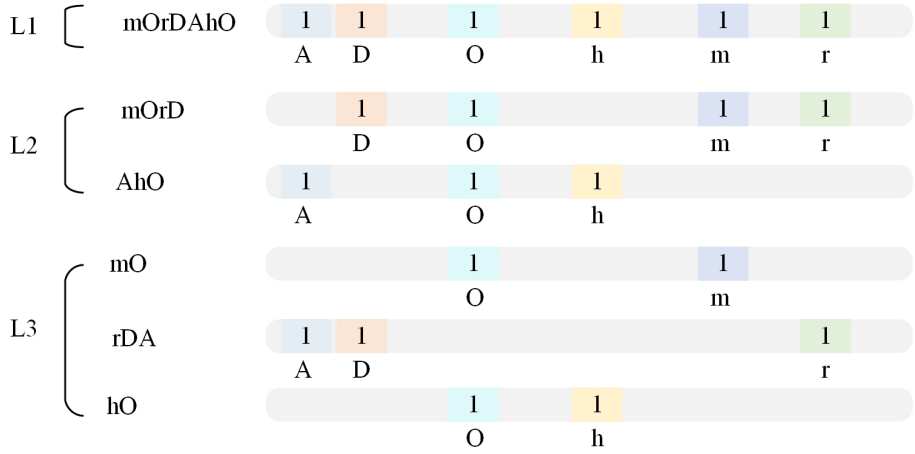


Fig. 4. An example for PHOC representation at levels 1, 2, and 3.

character appears in the first or second half of the word. The final PHOC vector is obtained by concatenating the features from each segmentation level.

Due to the pyramid structure of PHOC, text is encoded at multiple levels of spatial partitioning. In the Mongolian Kanjur dataset, for example, two word labels "GjjjjjnjE" and "GjjjjnjE" differ by only one additional character 'j' in the former. When applying a 2–3 level PHOC encoding, these two labels produce identical PHOC vectors, making it impossible to distinguish between them. Even with a 2–4 level split, the representations remain the same. However, when using a 2–5 level split, the resulting PHOC vectors become distinguishable. This demonstrates that finer-grained splits can capture subtle differences in character sequences and reduce representational ambiguity. Consistent with [41], we skip a global representation with 2, 3, 4 and 5 splits and the 50 most frequent bigrams at the second level split to represent text. As a consequence, the final length of PHOC is :

$$L = (2 + 3 + 4 + 5) \times l + 2 \times 50 \tag{5}$$

where l is the length of the character set in a given language.

3.3. Training Objective

To ensure the image features extracted by MetaCLIP learn discriminative image representations, the image triplet loss is employed to minimize the distance between an image anchor f_I and a positive image sample f_I^+, while maximizing the distance between the anchor and a negative image sample f_I^-:

$$\mathcal{L}_{Tri} = max\,(0,\, d(f_I, f_I^+) - d(f_I, f_I^-) + \alpha) \tag{6}$$

where α is the margin parameter. $d(\cdot)$ represents cosine distance. Given two vectors f_1 and f_2, the cosine distance $d(f_1, f_2)$ is defined as $1 - \frac{f_1 \cdot f_2}{\|f_1\|\|f_2\|}$.

Mean squared error (MSE) loss ensures that the features f_{AI} extracted through Image-to-Text Adaptation align more closely with text features f_T extracted by MetaCLIP text encoder, while

features f_{AT} obtained through Text-to-Image Adaptation become more similar to image features f_I extracted by MetaCLIP image encoder.

$$\mathcal{L}_I = \frac{1}{N}\sum_{i=1}^{N}\left\|f_I^i - f_{AT}^i\right\|^2 \tag{7}$$

$$\mathcal{L}_T = \frac{1}{N}\sum_{i=1}^{N}\left\|f_{AI}^i - f_T^i\right\|^2 \tag{8}$$

where N represents the dimension of the text or image representation vector.

Finally, the total loss is a weighted combination:

$$\mathcal{L}_{total} = \mathcal{L}_{Tri} + \lambda_1\mathcal{L}_I + \lambda_2\mathcal{L}_T \tag{9}$$

where λ_1, λ_2 are weighting factors. This combines loss function enables the model to learn robust discriminative features and encode shared and complementary knowledge among multimodal data to facilitate modality translation.

4. Experimental Results

4.1. Datasets and Evaluation Criteria

To evaluate our model's generalization ability for OOV word spotting, we use two historical Mongolian document image datasets: Kanjur [32] and Geser [41]. We specifically choose Mongolian as the target language due to its agglutinative morphology, where a large number of word forms are systematically derived by attaching suffixes to root words. With around 60,000 root words and hundreds of suffixes, the vocabulary size can easily exceed one million. This makes it infeasible for training sets to cover all word forms, resulting in frequent OOV problem, thus making Mongolian particularly well-suited for evaluating OOV word spotting. In our experiments, both datasets are randomly divided into training and testing sets using an OOV scheme.

The Kanjur dataset is a collection of scanned images from the woodblock-printed Mongolian Kanjur, comprising a total of 100 pages in TIF format and 20,948 word images at 600 dpi. The text annotations are manually transcribed via glyph encoding, with a vocabulary size of 1,423 and 48 unique characters (l=48). Following [32], Kanjur dataset is roughly divided into four folds according to word categories with no overlap. Three folds contain 356 words and the other fold contains 355 words. The detailed partition is listed in Table 1. By substituting $l = 48$ into equation (5), the final dimension of the PHOC representation is $(2 + 3 + 4 + 5) \times 48 + 2 \times 50 = 772$ for the Kanjur dataset.

Table 1. Partition of the Kanjur dataset for 4-fold cross-validation.

Fold	Vocabulary	Word Images	Queries
Fold1	356	5580	5494
Fold2	356	5629	5555
Fold3	356	5219	5144
Fold4	355	4520	4424
Total	1423	20948	20617

The Geser dataset contains two versions of historical Mongolian documents: the Beijing woodblock printing and the Longfu Temple manuscript, comprising a total of 896 pages of document images in JPEG format at 300 dpi. It is a large-scale dataset consisting of 143,508 word images annotated with middle coding, yielding a vocabulary size of 10,951 and 36 distinct characters ($l = 36$). In [41], the dataset is randomly divided into training and testing sets, with training set proportions of 20%, 40%, 60% and 80%. To ensure a strict OOV setting, there is no overlap in word categories between the training and testing sets. The detailed partition is listed in Table 2. By substituting $l = 36$ into equation (5), the length of PHOC representations for the Geser dataset is $(2 + 3 + 4 + 5) \times 36 + 2 \times 50 = 604$.

Table 2. Partition of the Geser dataset for training and test sets.

Training Set Proportion	Vocabulary (Training/Test)	Word images (Training/Test)
20%	2,190/8,761	28,265/115,243
40%	4,380/6,571	56,602/86,906
60%	6,570/4,381	86,438/57,070
80%	8,760/2,191	112,371/31,137

In addition, to evaluate the generalizability of our method, we conducted experiments on the Manchu language, which shares similar linguistic characteristics with Mongolian. The dataset we used is a recently released Manchu text image dataset by Sun et al. [27], originally designed for OCR tasks. It is currently the largest publicly available Manchu dataset, comprising 2,428 unique words, each with 10 high-resolution images. The data originates from Manchu ancient books dating from 1733 to 1867, collected within the present-day territory of China, forming a balanced and well-structured dataset. Similarly, we reorganized this dataset into four folds based on vocabulary, ensuring that no overlap exists between any two folds. The details of the data split are provided in Table 3. Given the balanced distribution of samples, each fold contains 607 unique words and 6,070 image samples. According to statistical analysis, the dataset includes 26 distinct characters ($l = 26$). Substituting $l = 26$ into equation (5), the length of the PHOC representation for the Manchu dataset is $(2 + 3 + 4 + 5) \times 26 + 2 \times 50 = 464$.

Table 3. Partition of the Manchu dataset for each fold.

Fold	Vocabulary (Training/Test)	Word Images (Training/Test)
F1	1,821 / 607	18,210 / 6,070
F2	1,821 / 607	18,210 / 6,070
F3	1,821 / 607	18,210 / 6,070
F4	1,821 / 607	18,210 / 6,070

In line with other word spotting tasks [3, 10, 41], we employ mean average precision (mAP) as the primary evaluation metric. Let AP_q be the average precision for the q^{th} query word image:

$$\mathrm{AP}_q = \frac{1}{R_q} \sum_{k=1}^{N_q} \mathrm{P}(k) \times rel(k) \tag{10}$$

where N_q is the number of the selected word images that are related to the query, $rel(k)$ is 1 if the

rank k in the retrival list is a relevant word image and 0 otherwise, $P(k)$ is the precision at cut-off k in the list, and R_q the number of relevant word images for the query word image. Denoting Q as the total number of queries, the mAP can be defined as:

$$\text{mAP} = \frac{1}{\text{Q}} \sum_{q=1}^{\text{Q}} \text{AP}_\text{q} \tag{11}$$

4.2. Implementation Details

After conducting a statistical analysis on the dataset, we selected the 2-, 3-, 4-, and 5-level splits for PHOC. As shown in Table 4, when using 2–3 level splits to generate PHOC representations for all words, we found 6 duplicate PHOC representations. With 2–4 level splits, there were 2 duplicate PHOC representations. However, when using 2–5 level splits, no duplicate PHOC representations were found. Therefore, we adopted the 2–5 level splits for PHOC.

Table 4. Statistics of duplicate PHOC representations under different level splits.

PHOC Levels	Duplicate PHOCs	Affected Words
2, 3	6	12
2, 3, 4	2	4
2, 3, 4, 5	0	0

For QBE word spotting, word images appear only once are discarded, which means that these word images are not selected as queries. But they still appear as distractors in other retrieval lists. In addition, query instances are removed from the retrieval list and not counted. Besides, the retrieval results are sorted according to the cosine distance metric.

Our method is developed using the PyTorch framework, and runs on Tesla V100s GPUs. We use the Adam optimizer setting learning rates to 1e-6 for MetaCLIP's image encoder and text encoder, 1e-4 for the bi-directional modality adaptation framework, with a batch size of 64. We employ MetaCLIP with Vision Transformer (ViT) [5] as its backbone and utilize ViT-B-32-quickgelu as its pre-trained weights. The feature vectors encoded by MetaCLIP have a dimensionality of 512. Weighting factors λ_1 and λ_2 are set to 1.

4.3. Results and Analysis

For Kanjur dataset, we perform 4-fold cross-validation and report the average mAP performance. As shown in Table 5, we compare our method with existing methods from low-level feature representation to deep feature representation, such as Bag of Visual Words (BoVW) [8], Average Visual Word Embedding (AVWE) [33], and its variant AVWE-SC [34], CNN (FC1) [35], RNN [33], Sequence-to-Sequence models Seq2seq (F2H1) [32], HENet [41], PUNet [3], etc. Additionally, we reproduce and fintune CLIP [20] and MetaCLIP [37] on Kanjur dataset for comparison. Based on the results in Table 5, we draw the subsequent observations:

First, for QBE word spotting, our MARK model outperforms all existing methods by achieving the highest average mAP of 99.16% across 4 folds. Compared to the previous state-of-the-art method HENet (96.47%), our approach achieves a notable improvement of 2.69%. This performance gain can be attributed to the use of CLIP as the backbone for image feature extraction, which provides strong generalization capabilities. Furthermore, our model surpasses both CLIP

Table 5. Comparative results of our MARK against existing methods on the Kanjur Dataset.

Method	Query	F1 (%)	F2 (%)	F3 (%)	F4 (%)	Avg (%)
BoVW [8]	QBE	40.40	51.43	42.25	43.62	44.43
AVWE [33]		59.57	66.83	60.76	58.63	61.45
RNN [33]		54.54	62.05	53.30	60.82	57.68
AVWE-SC [34]		63.25	65.06	63.99	64.71	64.25
AVWE+RNN [33]		75.05	79.43	75.39	75.83	76.43
Seq2seq (F2H1) [32]		80.16	85.08	79.08	84.59	82.23
CNN (FC1) [35]		84.48	85.98	80.03	83.27	83.44
Seq2seq+CNN[35]		88.02	90.35	84.62	87.21	87.55
PUNet [3]		95.57	95.02	95.63	93.19	94.85
EENet [41]		96.63	96.53	96.89	95.58	96.41
HENet [41]		96.70	96.63	96.92	95.62	96.47
CLIP* [20]		97.99	98.52	98.36	97.99	98.21
MetaCLIP* [37]		98.90	99.17	99.20	98.76	99.01
MARK (Ours)		**99.10**	**99.40**	**99.27**	**98.89**	**99.16**
PUNet [3]	QBS	88.46	85.67	91.77	90.11	89.00
EENet [41]		88.90	94.93	92.61	92.39	92.21
HENet [41]		89.19	95.13	93.21	92.51	92.51
CLIP* [20]		70.62	71.17	79.32	84.82	76.48
MetaCLIP* [37]		77.73	75.50	86.06	77.89	79.30
MARK (Ours)		**98.58**	**98.01**	**98.71**	**94.27**	**97.39**

* Indicates our reproduction, training on the Kanjur dataset using the loss function mentioned in Section 3.3.

and MetaCLIP , which are originally designed for natural image. This demonstrates the effectiveness of our bi-directional modality adaptation module, which successfully bridges the modality gap by capturing shared and complementary information between image and text.

Second, for QBS word spotting, our MARK model again demonstrates superior performance by achieving an average mAP of 97.39%, significantly surpassing all baseline methods. Compared to the SOTA method HENet (92.51%), MARK achieves an absolute gain of 4.88%. Although CLIP and MetaCLIP demonstrates strong performance for QBE word spotting, they show limited capability in QBS word spotting, even underperforming compared to some CNN-based methods. Notably, compared to CLIP, our model achieves a 20.91% higher four-fold average mAP, while outperforming MetaCLIP by 18.09%. This substantial improvement can be attributed to the introduction of PHOC into the text-to-image adaptation process, which enhances fine-grained text representations and mitigates the limitations of CLIP's text encoder.

Overall, our proposed MARK outperforms all the aforementioned methods in both QBE and QBS OOV word spotting tasks. It can be attributed to two key factors: a) We leverage MetaCLIP as the backbone for image and text feature extraction, benefiting from its strong generalization capabilities. b) We introduce PHOC, a character-level textual representation, which enhances the model's ability to generalize to unseen categories by capturing fine-grained character-level features, particularly when the training data includes a diverse range of character combinations.

For Geser dataset, we report the mAP performance across different training set proportions (20%, 40%, 60%, and 80%) to evaluate the model's generalization performance for OOV with varying amounts of training data. Table 6 presents the comparative results of MARK against existing methods (PUNet [3], EENet [41], and HENet [41]). Additionally, we also reproduce and finetune CLIP [20] and MetaCLIP [37] on the Geser dataset for comparison. Experimental results demonstrate that the proposed MARK achieves a remarkable improvement in QBE and QBS word spotting. It is worth noting that even when the training set constitutes only 20%, our method still achieves strong performance, with QBE and QBS reaching 85.19% and 81.87%, respectively, significantly outperforming the current SOTA model, HENet, by 11.74% for QBE and 22.49% for QBS. As the training set proportion increases, MARK's performance shows substantial improvements, achieving 95.65% for QBE and 93.18% for QBS with 80% of the training data. Consistent with the results observed on the Kanjur dataset, MARK achieves a remarkable improvement in QBS word spotting on the Geser dataset, further underscoring the effectiveness of MARK in enhancing cross-modal retrieval tasks.

Table 6. Comparative results of our MARK against existing methods on the Geser Dataset.

Method	Training Set Proportion							
	20%		40%		60%		80%	
	QBE (%)	QBS (%)	QBE (%)	QBS (%)	QBE (%)	QBS (%)	QBE (%)	QBS (%)
PUNet [3]	70.85	57.42	84.33	70.37	86.53	71.57	89.42	82.96
EENet [41]	73.35	58.65	86.13	72.07	88.57	74.16	93.91	91.38
HENet [41]	73.45	59.38	86.24	73.01	88.67	74.75	94.18	92.83
CLIP* [20]	82.03	42.21	89.95	65.99	90.98	71.66	94.70	83.28
MetaCLIP* [37]	82.85	79.62	90.33	84.54	91.74	85.96	94.32	91.10
MARK (Ours)	**85.19**	**81.87**	**91.59**	**86.84**	**92.64**	**88.17**	**95.65**	**93.18**

* Indicates our reproduction, training on the Geser dataset using the loss function mentioned in Section 3.3.

To qualitatively evaluate retrieval performance, we visualize top-10 retrieval results on the Geser dataset in Fig.5, comparing our method with MetaCLIP. In the figure, green checkmarks indicate the retrievals that are correctly matched to the query, while red crosses point to instances where retrieval errors occurred. Fig.5 (a) shows QBE word spotting results. In the first two rows, our method retrieves a greater number of relevant examples than MetaCLIP. In the last two rows, although the number of returned retrieved samples is the same, our method consistently ranks the relevant samples higher, resulting in better retrieval performance. Fig.5 (b) presents QBS word spotting results. In the first two rows, our method retrieves only relevant samples in the top 10, while MetaCLIP returns many incorrect ones. In the last two rows, our method retrieves a larger number of correct results, demonstrating the contribution of PHOC, which captures detailed character information and compensates for the limitations of MetaCLIP in handling fine-grained textual features.

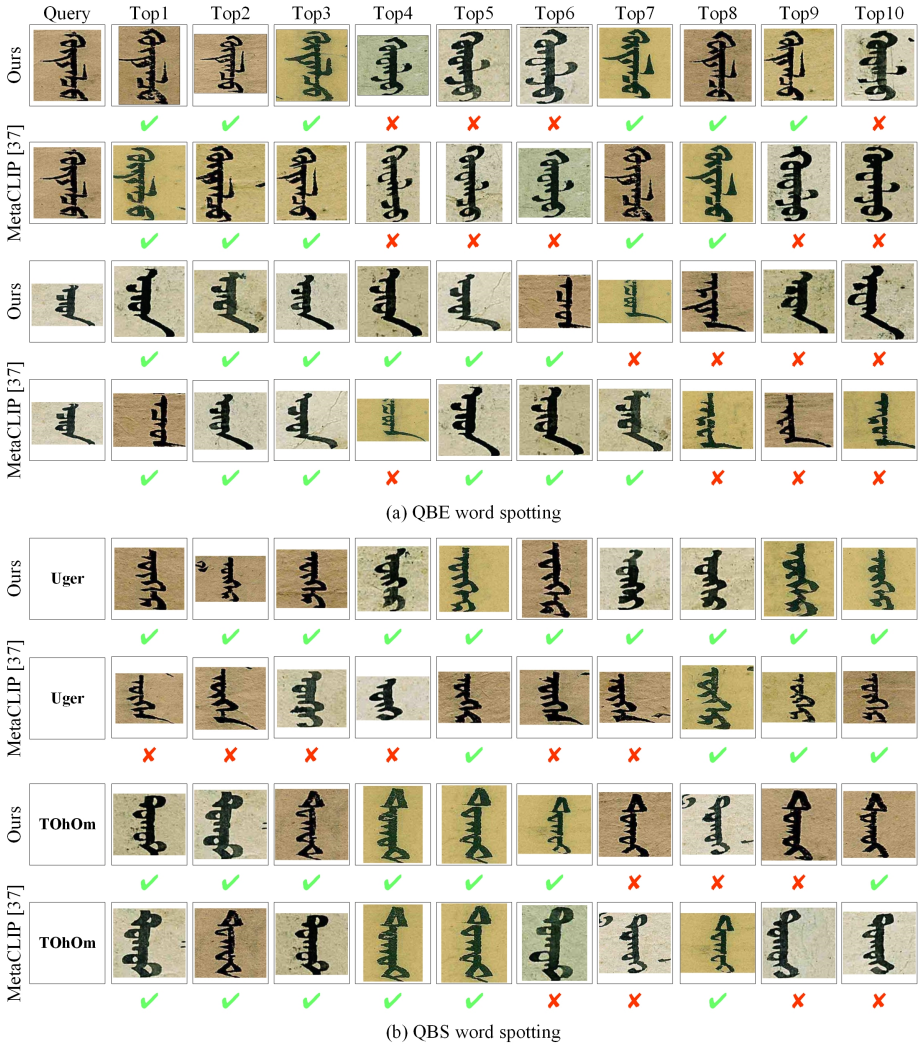


Fig. 5. Retrieval results on Geser dataset for both QBE and QBS word spotting.

For Manchu dataset, we conducted 4-fold cross-validation following the data splitting strategy described in Section 4.1. The experimental results for each fold, along with the average performance across 4 folds, are presented in Table 7. For the QBE word spotting task, we achieved an mAP of 99.69% on the first fold, 99.63% on the second fold, 99.72% on the third fold, and 99.80% on the fourth fold, with an average mAP of 99.71%. For the QBS word spotting task, the mAP reached 98.27% on the first fold, 97.91% on the second fold, 98.38% on the third fold, and 98.59% on the fourth fold, resulting in an average mAP of 98.29%. The consistently high mAP scores across all four folds in both QBE and QBS tasks demonstrate the robustness and stability of our method. These results clearly confirm the potential of our approach to generalize well to other low-resource, morphologically rich languages beyond Mongolian.

Table 7. Four-fold cross-validation results of our MARK on the Manchu dataset.

Fold	QBE (%)	QBS (%)
F1	99.69	98.27
F2	99.63	97.91
F3	99.72	98.38
F4	99.80	98.59
AVG	99.71	98.29

4.4. Ablation Experiments

4.4.1. Ablation Study on Modality Adaption.

The ablation study presented in Table 8 reveals the impact of the Image-to-Text Adaptation module (denoted as ITA) and the Text-to-Image Adaptation module (denoted as PTIA) on the average mAP over four folds of the Kanjur dataset for QBE and QBS word spotting. The first row of Table 8 shows the performance of fine-tuning MetaCLIP without any adaptation modules. The QBS performance is notably poor, indicating that MetaCLIP alone is not well-suited for QBS word spotting, which requires fine-grained character-level information. After incorporating either the ITA or PTIA module, both QBE and QBS performance improve, with a particularly significant improvement in QBS. This highlights the contribution of the PTIA module, which leverages PHOC to represent statistical character information, thus compensating for the limitations of MetaCLIP in character-level representation. Finally, when both ITA and PTIA modules are combined, the model achieves the highest performance, with QBE up to 99.16% and QBS up to 97.39%, surpassing the previous state-of-the-art results and demonstrating the effectiveness of the proposed method.

Table 8. Ablation study on modality adaption modules for Kanjur dataset.

ITA	PTIA	QBE (%)	QBS (%)
○	○	99.01	79.30
√	○	99.07	86.91
○	√	99.03	97.11
√	√	**99.16**	**97.39**

4.4.2. Ablation Study on PHOC.

An ablation study on PHOC is conducted to evaluate its impact on model performance. The models with and without PHOC are denoted as w-PHOC and wo-PHOC, respectively. As shown in Fig. 6, the model with PHOC (blue) outperforms the model without PHOC (red) in both QBE and QBS word spotting on each fold of the Kanjur dataset. The improvement in QBS is particularly pronounced, which can be attributed to PHOC's capability to capture character-level information. Not only does it compensate for CLIP's text encoder limitations in fine-grained character perception, but it also strengthens the PTIA module's ability to capture shared and complementary information between text and image representations.

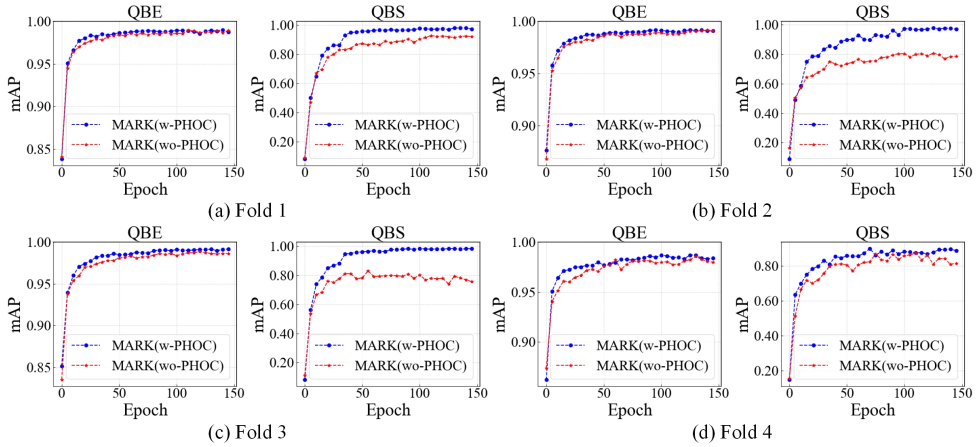


Fig. 6. Ablation study on PHOC for Kanjur dataset across four folds. The performance with PHOC (blue) consistently outperforms the model without PHOC(red), particularly in QBS word spotting.

4.4.3. Comparison Between KAN and MLP.

To assess the impact of different architectures, we replace the hidden layers of the ITA and PTIA modules with either KAN or MLP, referred to as w-KAN and w-MLP, respectively. Fig. 7 presents the comparative results of KAN and MLP in the modality adaptation module for QBE and QBS word spotting cross four folds. The KAN-based model (in blue) consistently surpasses the MLP-based model (in red) in both tasks, with a more substantial improvement in QBS word spotting. This can be attributed to KAN's ability to capture complex feature relationships through its Radial Basis Function (RBF)-based structure. This suggests that KAN enables more effective feature alignment and enhances the learning of fine-grained text representations. Moreover, we observe that the MLP-based model exhibits greater training instability and fluctuations compared to the KAN-based model, which demonstrates more stable convergence.

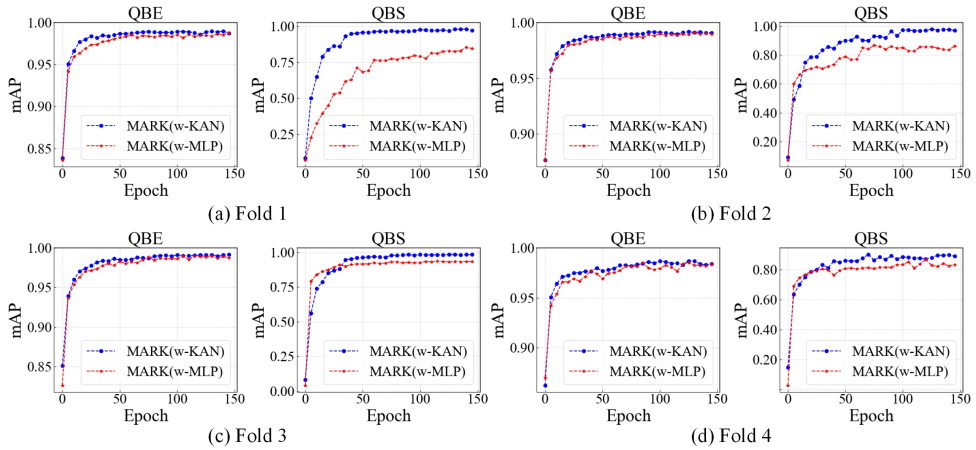


Fig. 7. Performance comparison between KAN and MLP on the Kanjur dataset cross four folds. The model using KAN-based (blue) consistently outperforms the MLP-based model (red), particularly in QBS word spotting.

5. Conclusion

In this paper, we propose MARK, a bi-directional modality adaptation framework designed to address the critical challenge of modality gaps between text and images in OOV word spotting for historical document image retrieval. Compared to existing works, we design the Image-to-Text Adaptation (ITA) and PHOC-based Text-to-Image Adaptation (PTIA) modules based on KAN, incorporating a KAN-based multi-head self-attention mechanism with a gating mechanism for adaptive attention across different dimensions. Experimental results demonstrate that KAN offers significant advantages over MLP by improving fine-grained text representation learning, capturing cross-modal relationships more effectively, and ensuring more stable training and convergence. Notably, QBS word spotting exhibits greater gains than QBE, revealing the strength of KAN in handling text-driven cross-modal tasks and offering valuable insights for future research beyond historical documents.

Furthermore, to mitigate CLIP's limitations in fine-grained tasks, we introduce PHOC, a character-level word representation, to enhance the text-to-image adaptation process. Experimental results also show that incorporating traditional rules, such as character-level information, substantially improves CLIP's performance in word spotting tasks. In summary, compared to the state-of-the-art methods, our MARK model establishes a new state-of-the-art for QBE and QBS on both the Kanjur and Geser datasets. In addition, we further validate the effectiveness and generalizability of our method on the Manchu dataset, which shares similar linguistic characteristics with Mongolian. The consistently high performance across QBE and QBS word spotting confirms the extensibility of MARK to low-resource, morphologically rich languages with complex writing systems. This work paves the way for further advancements in text-centric tasks, inspiring new research directions and innovations in the field.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (62466040), the Project for Science and Technology of Inner Mongolia Autonomous Region (2025YFHH0079, 2019GG281), the Natural Science Foundation of Inner Mongolia Autonomous Region (2024MS06029), and the Research Foundation for Young Scholars of Inner Mongolia University (10000-A25206024).

References

[1] Abdelfattah, R., Guo, Q., Li, X., Wang, X., Wang, S.: Cdul: Clip-driven unsupervised learning for multi-label image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1348–1357 (2023)

[2] Almazán, J., Gordo, A., Fornés, A., Valveny, E.: Word spotting and recognition with embedded attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(12), 2552–2566 (2014)

[3] Boudraa, O., Michelucci, D., Hidouci, W.K.: Punet: Novel and efficient deep neural network architecture for handwritten documents word spotting. Pattern Recognition Letters **155**, 19–26 (2022)

[4] Cao, W., Wu, Y., Sun, Y., Zhang, H., Ren, J., Gu, D., Wang, X.: A review on multimodal zero-shot learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **13**(2), e1488 (2023)

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)

[6] Dovgalecs, V., Burnett, A., Tranouez, P., Nicolas, S., Heutte, L.: Spot it! finding words and patterns in historical documents. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 1039–1043. IEEE (2013)

[7] Frinken, V., Fischer, A., Manmatha, R., Bunke, H.: A novel word spotting method based on recurrent neural networks. IEEE transactions on pattern analysis and machine intelligence **34**(2), 211–224 (2011)
[8] Guo, X., Wei, H., Su, X.: A case study of bovw for keyword spotting on historical mongolian document images. In: International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. pp. 374–378. IEEE (2016)
[9] Kang, L., Tito, R., Valveny, E., Karatzas, D.: Multi-page document visual question answering using self-attention scoring mechanism. In: International Conference on Document Analysis and Recognition. pp. 219–232. Springer (2024)
[10] Krishnan, P., Dutta, K., Jawahar, C.: Hwnet v3: a joint embedding framework for recognition and retrieval of handwritten text. International Journal on Document Analysis and Recognition **26**(4), 401–417 (2023)
[11] Li, X., Wu, Y., Jiang, X., Guo, Z., Gong, M., Cao, H., Liu, Y., Jiang, D., Sun, X.: Enhancing visual document understanding with contrastive learning in large visual-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15546–15555 (2024)
[12] Li, X., Yu, J., Li, Z., Lu, H., Yuan, R.: Dr. clip: Clip-driven universal framework for zero-shot sketch image retrieval. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 9554–9562 (2024)
[13] Li, Z.: Kolmogorov-arnold networks are radial basis function networks. arXiv preprint arXiv:2405.06721 (2024)
[14] Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T.Y., Tegmark, M.: Kan: Kolmogorov-arnold networks. arXiv preprint arXiv:2404.19756 (2024)
[15] Lladós, J., Rusinol, M., Fornés, A., Fernández, D., Dutta, A.: On the influence of word representations for handwritten word spotting in historical documents. International journal of pattern recognition and artificial intelligence **26**(05), 1263002 (2012)
[16] Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2200–2209 (2021)
[17] Mhiri, M., Desrosiers, C., Cheriet, M.: Word spotting and recognition via a joint deep embedding of image and text. Pattern Recognition **88**, 312–320 (2019)
[18] Novack, Z., McAuley, J., Lipton, Z.C., Garg, S.: Chils: Zero-shot image classification with hierarchical label sets. In: International Conference on Machine Learning. pp. 26342–26362. PMLR (2023)
[19] Peng, F., Yang, X., Xiao, L., Wang, Y., Xu, C.: Sgva-clip: Semantic-guided visual adapting of vision-language models for few-shot image classification. IEEE Transactions on Multimedia **26**, 3469–3480 (2023)
[20] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
[21] Rath, T.M., Manmatha, R.: Word spotting for historical documents. International Journal of Document Analysis and Recognition (IJDAR) **9**, 139–152 (2007)
[22] Razzaghi, P., Abbasi, K., Shirazi, M., Shabani, N.: Modality adaptation in multimodal data. Expert Systems with Applications **179**, 115126 (2021)
[23] Rusiñol, M., Aldavert, D., Toledo, R., Lladós, J.: Efficient segmentation-free keyword spotting in historical document collections. Pattern recognition **48**(2), 545–555 (2015)
[24] Sain, A., Bhunia, A.K., Chowdhury, P.N., Koley, S., Xiang, T., Song, Y.Z.: Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2765–2775 (2023)
[25] Sudholt, S., Fink, G.A.: Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In: International Conference on Frontiers in Handwriting Recognition. pp. 277–282. IEEE (2016)
[26] Sudholt, S., Fink, G.A.: Evaluating word string embeddings and loss functions for cnn-based word spotting. In: 2017 14th iapr international conference on document analysis and recognition (icdar). vol. 1, pp. 493–498. IEEE (2017)
[27] Sun, H., Tao, W., Bi, X.: A dataset of Manchu ancient book word images for OCR tasks, China, 1733–1867. (May 2025). https://doi.org/10.57760/sciencedb.25676
[28] Toselli, A.H., Puigcerver, J., Vidal, E.: Probabilistic indexing for information search and retrieval in large collections of handwritten text images, vol. 49. Springer Nature (2024)
[29] Tüselmann, O., Müller, F., Wolf, F., Fink, G.A.: Recognition-free question answering on handwritten document collections. In: International Conference on Frontiers in Handwriting Recognition. pp. 259–273. Springer (2022)
[30] Wei, H., Gao, G.: A keyword retrieval system for historical mongolian document images. International Journal on Document Analysis and Recognition **17**, 33–45 (2014)
[31] Wei, H., Gao, G., Su, X.: A multiple instances approach to improving keyword spotting on historical mongolian document images. In: International Conference on Document Analysis and Recognition. pp. 121–125. IEEE (2015)
[32] Wei, H., Kang, Y., Zhang, H.: Word image representation based on sequence to sequence model with attention mechanism for out-of-vocabulary keyword spotting. In: IEEE International Conference on High Performance Computing and Communications; IEEE International Conference on Smart City; IEEE International Conference on Data

Science and Systems. pp. 2224–2231. IEEE (2019)
[33] Wei, H., Zhang, H., Gao, G.: Representing word image using visual word embeddings and rnn for keyword spotting on historical document images. In: IEEE International Conference on Multimedia and Expo. pp. 1368–1373. IEEE (2017)
[34] Wei, H., Zhang, H., Gao, G.: Word image representation based on visual embeddings and spatial constraints for keyword spotting on historical documents. In: 2018 24th International Conference on Pattern Recognition. pp. 3616–3621. IEEE (2018)
[35] Wei, H., Zhang, J., Liu, K.: A hybrid representation of word images for keyword spotting. In: International Conference on Neural Information Processing. pp. 3–10. Springer (2020)
[36] Wei, H., Zhang, J., Zhang, H.: Deep features representation of word image for keyword spotting in historical mongolian document images. In: IEEE International Conference on Tools with Artificial Intelligence. pp. 413–417. IEEE (2020)
[37] Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data. In: 12th International Conference on Learning Representations, ICLR 2024 (2024)
[38] Zagoris, K., Pratikakis, I., Gatos, B.: Segmentation-based historical handwritten word spotting using document-specific local features. In: 2014 14th International Conference on Frontiers in Handwriting Recognition. pp. 9–14. IEEE (2014)
[39] Zagoris, K., Pratikakis, I., Gatos, B.: Unsupervised word spotting in historical handwritten document images using document-oriented local features. IEEE Transactions on Image Processing **26**(8), 4032–4041 (2017)
[40] Zeng, G., Zhang, Y., Wei, J., Yang, D., Zhang, P., Gao, Y., Qin, X., Zhou, Y.: Focus, distinguish, and prompt: Unleashing clip for efficient and flexible scene text retrieval. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 2525–2534 (2024)
[41] Zhang, J., Wei, H., Zhang, Q., Chen, X., Ma, J.: Henet: Hyperbolic-based encoder-decoder network for word spotting in historical mongolian documents. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3580–3584. IEEE (2024)
[42] Zhang, Q., Zhang, J., Bao, F., Su, X., Gao, G.: Structural-aware disentangled learning with clip for hyperbolic zero-shot sketch-based image retrieval. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
[43] Zhong, Z., Pan, W., Jin, L., Mouchère, H., Viard-Gaudin, C.: Spottingnet: learning the similarity of word images with convolutional neural network for word spotting in handwritten historical documents. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 295–300. IEEE (2016)

Author Biographies



Jing Zhang received the B.E. degree from Inner Mongolia University, Hohhot, China, in 2018. She is currently a Ph.D. candidate in the College of Computer Science, Inner Mongolia University. Her research work is focused on pattern recognition, image retrieval, and word spotting.
Contact: 22009008@mail.imu.edu.cn



Hongxi Wei received the Ph.D. degree from Inner Mongolia University (China) in 2013. He is currently a professor at the College of Computer Science, Inner Mongolia University. His research interests include pattern recognition, computer vision, and intelligent information processing. He has authored more than 50 papers in peer-reviewed journals and international conferences.
Contact: cswhx@imu.edu.cn



Qing Zhang received the Ph.D. degree from Inner Mongolia University (China) in 2024. He is currently a lecturer at the College of Computer Science, Inner Mongolia University. His research interests include pattern recognition, computer vision, and image retrieval. He has authored more than 10 papers in peer-reviewed journals and international conferences since 2017.
Contact: cszq@imu.edu.cn