Writer Retrieval at Scale

Tim Raven ✉ [0000−0002−9142−7864], Tim Hallyburton[0009−0008−0944−9193], and Gernot A. Fink[0000−0002−7446−7813]

TU Dortmund University, Dortmund, Germany
{tim.raven, tim.hallyburton, gernot.fink}@tu-dortmund.de

Abstract. We introduce FormWR, a challenging large-scale dataset for writer retrieval and a supervised end-to-end full-image retrieval method using a novel learnable feature aggregation module, X-VLAD. FormWR contains almost 400k pages of application-for-assistance forms attributed to almost 100k distinct writer proxies. Documents are dominated by printed templates, and further reflect real-world uncertainty through multi-hand contamination and pages with sparse handwriting. Our model jointly trains a feature extractor and aggregator. It is pretrained on sets of 32×32 handwriting-centered patches and then fine-tuned on full-page images. We report a comprehensive large-scale evaluation on FormWR. Further, we demonstrate transfer to fragment-level retrieval on the HisFragIR20 competition benchmark. We achieve new state-of-the-art performance with 97.9% Top-1 accuracy and 78.8% mAP, improving both metrics by large margins (+12.7% Top-1 and +21.6% mAP). Dataset and code are available at `https://github.com/Traven16/writer-retrieval-at-scale`.

Keywords: Writer Retrieval · Large-scale Benchmark · Metric learning

1 Introduction

Writer retrieval is the task of searching a collection (*gallery*) of handwritten documents for those authored by the same writer as a given query sample. It supports applications in historical research and forensic document analysis, including tracing individuals across epochs, verifying authorship, and organizing large document collections.

In related retrieval and identification tasks (e.g., face recognition, speaker verification, and visual place recognition), substantial progress over the last decade has been driven by supervised embedding learning (metric learning or margin-based classification objectives) [9, 37, 41] and, for inputs represented as sets or sequences of local descriptors, by learned pooling/aggregation mechanisms [1, 26, 40]. In contrast, recent writer-retrieval systems are predominantly label-free, and supervised variants are often reported only as baselines that trail unsupervised approaches [5, 30, 36].

We argue that a key bottleneck is dataset scale. For instance, Historical-WI [11], one of the most widely used benchmarks, contains 1,182 training pages attributed to 394 writers (three pages per writer). Such regimes encourage

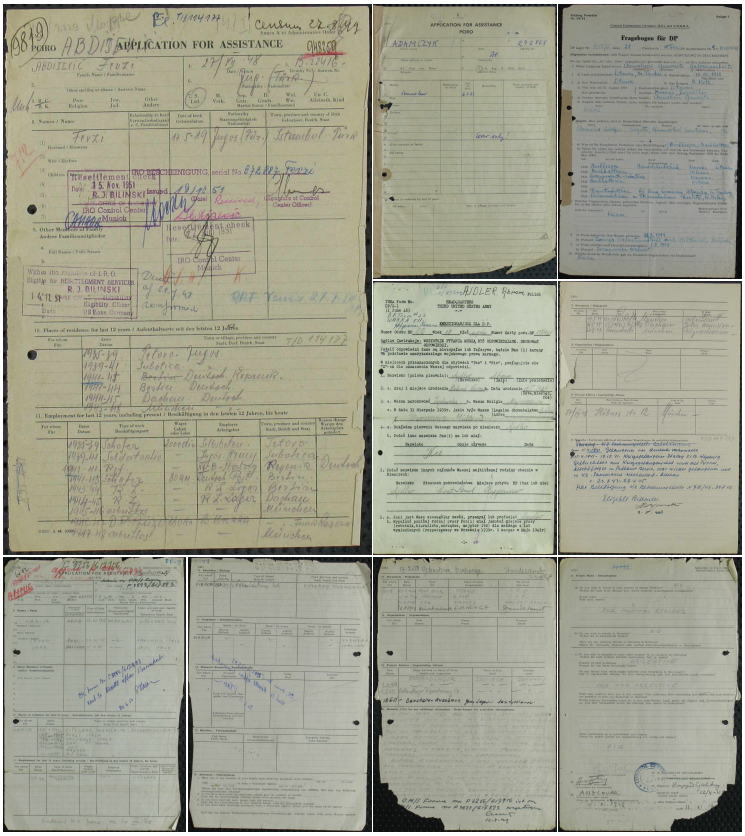


Fig. 1: Sample pages from FormWR. Zoom in for details.

methods that rely on strong inductive biases or pretraining, but make it difficult to exploit supervision effectively.

To address this gap, we introduce FormWR, a benchmark for writer retrieval at scale, see Fig. 1 FormWR is constructed from over 100k distinct *Applications for IRO Assistance* from the CM/1 corpus of the Arolsen Archives[1]. We decompose each application into its constituent pages and define retrieval over pages: given a query page, the task is to retrieve other pages belonging to the same application. Since these forms are primarily filled out by the applicant, application membership serves as a practical proxy for writer identity, while naturally capturing real-world uncertainty (e.g., later annotations by third parties). After semi-automatic curation, the resulting dataset contains around 387k pages, split into roughly 95k pages for training/validation and 290k pages for testing, making it an order of magnitude larger than prior benchmark test sets.

[1] `https://collections.arolsen-archives.org/en/archive/3-2-1`

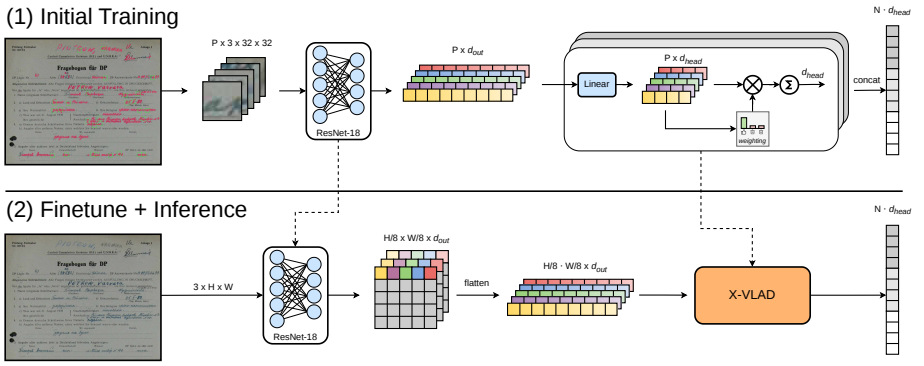


Fig. 2: Overview of the proposed method. For the initial training stage, handwriting masks are used to crop P handwriting-centric patches, which are then separately embedded. X-VLAD employs N parallel heads to obtain a fixed-length document representation. Each head projects local features into a low-dimensional subspace and applies a lightweight selection module that softly assigns each feature to either keep or one of g bin (discard) modes. The head output is the weighted pooled feature of the keep mode. In the second stage, full images are directly embedded. The final feature map is extracted, flattened and encoded with X-VLAD.

Beyond scale, FormWR introduces challenges that are underrepresented in existing benchmarks: (1) multi-hand contamination, where parts of a page may be written by other individuals during the application or archival process; (2) template dominance and other distractors (machine-printed backgrounds, tables, stamps) that can overwhelm handwriting cues; and (3) highly variable handwriting density, ranging from sparse field entries (e.g., name, date of birth) to longer free-text responses.

To tackle these conditions, we propose X-VLAD, a supervised end-to-end aggregation module designed for template-dominated documents with sparse handwriting, see Fig. 2. X-VLAD employs N parallel heads, each operating in its own learned subspace. This decouples selection from multi-centroid competition and allows heads to specialize on complementary aspects of the local feature distribution. We train the model on sets of 32×32 handwriting-centered patches and optionally fine-tune on full pages. In contrast to prior patch-based pipelines for writer retrieval that operate on individual patches [5, 29, 30, 35], we explicitly learn set-level aggregation during initial training by jointly aggregating P patches per sample.

Our contributions are:

1. We present FormWR, a new large-scale benchmark for writer retrieval that increases collection size by roughly an order of magnitude over prior benchmarks and introduces realistic challenges such as multi-hand contamination and strong template dominance.

2. We introduce a supervised end-to-end methodology for writer retrieval built around a learned feature aggregator, X-VLAD, tailored to the conditions posed by FormWR.
3. We transfer our method to HisFragIR20 [38], a challenging benchmark for writer retrieval on historical document fragments, and substantially outperform prior work (+12.7% Top-1, +21.6% mAP).

We introduce FormWR in Section 3 and describe its design choices and curation pipeline. Section 4 presents our end-to-end retrieval model and X-VLAD aggregation. Section 5 reports results and ablations, including a scaling study to the full benchmark. In Section 6 we transfer our method to HisFragIR20. Finally, Section 7 summarizes findings and outlines directions for future work.

2 Related Work

2.1 Datasets and evaluation regimes

Early writer retrieval and identification benchmarks are largely based on controlled modern corpora, such as IAM [24] and CVL [20], and standardized evaluations like the ICDAR 2013 writer identification competition [23]. More challenging regimes emerged from historical material, including the ICDAR2017 Historical-WI competition [11], the ICDAR 2019 historical document retrieval benchmark [7], and fragment-level retrieval in HisFragIR20 [38]. Recent analyses show that retrieval performance depends strongly on available handwriting quantity (page vs. line/word) [32]. FormWR targets a complementary regime: modern form archives where writer signal can be sparse relative to strong printed structure and other distractors while also scaling up gallery size.

2.2 Set aggregation for retrieval

Scalable retrieval requires mapping a variable-size set of local descriptors to a fixed-length embedding. Bag-of-Words [39] is efficient but discards residual information; Fisher Vectors [33] and VLAD [17] improve discrimination by aggregating residual statistics. NetVLAD makes VLAD differentiable via soft assignments for end-to-end learning [1], while GhostVLAD introduces discard clusters for low-quality descriptors [45]. GeM provides a compact learnable pooling operator without explicit descriptor filtering [34]. Recent work refines assignment/selection (e.g., optimal-transport aggregation with a dustbin) [16], and SuperVLAD shows that aggressively reducing clustering complexity can improve robustness and compactness [10]. These methods can be viewed as permutation-invariant set functions [43, 44]; our X-VLAD follows this line with lightweight selection tailored to sparse-handwriting pages.

2.3 Writer retrieval methods

Classical pipelines combine local descriptors with explicit encodings and ranking. Examples include Fisher Vector aggregation [12], GMM supervectors [4], and

VLAD-style encodings of handcrafted features [3, 21], with further gains from re-ranking [19]. Deep approaches learn descriptors directly from image data [13] and can still benefit from classical encoders applied to learned activations [6]. For historical documents, unsupervised surrogate-class feature learning has been explored [5]. More recently, self-supervised and Transformer-based pipelines for writer retrieval have been proposed, often coupled with VLAD-type aggregation [27, 28, 30, 36], and methods often include an additional reranking step [29, 35].

3 FormWR Dataset

Increasing scale and complexity is the central goal of FormWR: we aim to move writer retrieval beyond the small, comparatively clean benchmarks where progress has largely saturated, and into a regime where both the number of pages and the nature of the documents meaningfully challenge current methods. FormWR is built from hand-filled *Applications for IRO Assistance*. Beyond scale, it targets three difficulty factors that are underrepresented in existing benchmarks: (1) multi-hand contamination, where parts of a page may be written by other individuals during the application or archival process (e.g., officers or archivists adding notes, corrections, stamps, or signatures); (2) strong distractors from the machine-printed template and tabular structure, which can dominate sparse handwriting and mislead naive appearance-based matching; and (3) highly variable handwriting density, ranging from only a few field entries (e.g., name, date of birth) to longer free-text responses. In addition, the template itself is not fixed: FormWR contains closely related template variants (languages and revisions) that shift field locations and printed content, making the background signal strong but non-stationary.

3.1 Source corpus and page normalization

FormWR is derived from hand-filled *Applications for IRO Assistance* in the CM/1 corpus of the Arolsen Archives. We build on the subset and weak template-cluster annotations reported by Wolf et al. [42], which provide the first page for a large number of applications. We crawl the remaining pages belonging to each application to construct a multi-page retrieval corpus. The raw scans include both single-page and double-page captures; we split double-page scans vertically into two single pages. The resulting raw pool contains ≈460k pages before curation.

3.2 Curation protocol

We apply a semi-automatic pipeline to remove pages that are ill-posed for writer retrieval (e.g., attachments or pages without applicant handwriting), while preserving variation in handwriting quantity and contamination.

(1) Size-based filtering. We discard pages with atypical dimensions. These are often attachments (cards, photos, appendices) with little relevant handwriting. Concretely, we retain pages near the dominant modes of the size distribution (single- and double-page scans; Fig. 3) and remove outliers.

Table 1: Dataset statistics for FormWR. Process identifiers serve as a proxy for writer identity.

Statistic	**Train**	**Val**	**Test**	**Total**
Pages	88,291	6,928	291,891	387,110
Processes (writer proxies)	24,538	1,932	70,601	97,071
Mean pages / process	3.6	3.59	4.13	3.99
Template clusters (distinct)	9	2	26	37

(2) Handwriting coverage estimation. To identify pages with insufficient handwriting, we train a handwriting-segmentation model on synthetic form pages following [18]. We sample representative template variants, construct handwriting-free backgrounds, and paste handwriting snippets into form fields to generate training data. A U-Net-style model estimates handwriting coverage per page; we manually inspect low-coverage candidates and remove pages without at least three handwritten words.

(3) Template-cluster inspection. Based on the weak template clusters from Wolf et al. [42], we inspect clusters for systematic issues (e.g., variants whose final page contains only case-worker signatures) and remove such pages while retaining the remaining pages of the corresponding processes.

3.3 Proxy labels, noise model, and evaluation protocol

Obtaining verified writer identities at archive scale is infeasible. We therefore use the process identifier (application identifier) as a proxy label: pages from the same process are treated as belonging to the same writer, motivated by the fact that the forms are primarily filled by the applicant. This choice intentionally admits label noise. First, pages may include handwriting from additional contributors (officers/archivists), and second, non-handwritten marks (stamps, marginalia) can create spurious similarity. Consequently, FormWR evaluates robustness under realistic uncertainty: methods must match applicant handwriting while discounting strong but non-writer cues.

Evaluation uses a leave-one-out cross validation, where each sample with at least one relevant match is used once as query. Considered metrics are Top-1 accuracy and mean average precision (mAP).

3.4 Splits and scale

Let $\mathcal{D}$ denote the curated page set and $p(x)$ the process identifier of page x. We provide a three-way partition

$$\mathcal{D} \;=\; \mathcal{D}_{\text{train}} \,\dot\cup\, \mathcal{D}_{\text{val}} \,\dot\cup\, \mathcal{D}_{\text{test}},$$

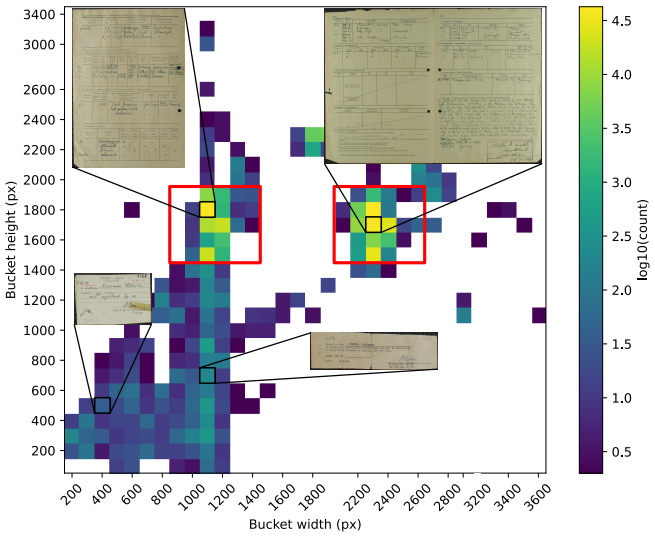


Fig. 3: Size statistics used for curation. We visualize page dimensions as a 2D histogram (log-counts). Two dominant modes correspond to single- and double-page scans. We retain pages near these modes (red markers) and remove outliers such as small attachments (e.g., signature cards).

constructed to prevent leakage and to separate model selection from final reporting. All splits are process-disjoint and template-cluster disjoint, reducing the risk of exploiting the printed layouts. We allocate a compact validation benchmark of ≈7k pages for hyperparameter search, ablations, and reference-method comparisons, and reserve the large-scale test benchmark of ≈290k pages for final reporting. Table 1 summarizes the resulting statistics.

4 X-VLAD

We learn compact document embeddings end-to-end for large-scale writer retrieval. In form archives, writer-discriminative signal (handwriting) is sparse and highly variable, while non-writer content (printed structure, stamps, layout drift) can dominate dense feature maps. X-VLAD addresses this by learning to *select* informative local descriptors and pool them into a fixed-length embedding suitable for nearest-neighbor search.

4.1 Preliminaries: VLAD-style set aggregation

Let $\mathbf{F} = \{f_i\}_{i=1}^{M}$ denote local descriptors extracted from an image, $f_i \in \mathbb{R}^D$, with variable cardinality M (e.g., $M{=}H{\cdot}W$ for a dense feature map). NetVLAD

aggregates descriptors by soft-assigning each f_i to one of K clusters and summing residuals [1]:

$$v_k^{\text{NetVLAD}} = \sum_{i=1}^{M} a_{ik}\,(f_i - c_k), \qquad k \in \{1, \ldots, K\}, \tag{1}$$

with learnable centroids $c_k \in \mathbb{R}^D$ and soft assignments a_{ik} (softmax over cluster logits). GhostVLAD introduces additional "ghost" bins that absorb low-quality descriptors and are discarded in the final embedding [45]. SuperVLAD removes residuals and pools the assigned descriptors directly [10]:

$$v_k^{\text{SuperVLAD}} = \sum_{i=1}^{M} a_{ik}\, f_i, \qquad k \in \{1, \ldots, K\}. \tag{2}$$

4.2 X-VLAD aggregation

X-VLAD builds on the SuperVLAD principle, but shifts aggregation into N parallel *single-cluster* heads operating in learnable d-dimensional subspaces. Each head learns its own projection and head-specific reject modes, enabling multiple complementary low-dimensional views without multi-centroid competition and yielding a compact embedding of total dimension Nd.

For head $n \in \{1, \ldots, N\}$ we learn a linear projection

$$z_i^{(n)} = P^{(n)} f_i, \qquad P^{(n)} \in \mathbb{R}^{d \times D}. \tag{3}$$

A lightweight selection module predicts logits over one *keep* mode (m=0) and g *bin* modes (m=1, …, g):

$$s_{i,m}^{(n)} = {w_m^{(n)}}^{\top} z_i^{(n)} + b_m^{(n)}, \qquad m \in \{0, \ldots, g\}, \quad w_m^{(n)} \in \mathbb{R}^d, \tag{4}$$

$$a_{i,m}^{(n)} = \frac{\exp\left(s_{i,m}^{(n)}\right)}{\sum_{m'=0}^{g} \exp\left(s_{i,m'}^{(n)}\right)}. \tag{5}$$

Only the keep-assignment $a_{i,0}^{(n)}$ contributes to the head representation:

$$v^{(n)} = \sum_{i=1}^{M} a_{i,0}^{(n)}\, z_i^{(n)} \in \mathbb{R}^d. \tag{6}$$

We apply per-head ℓ_2 intra-normalization, concatenate, and apply a final ℓ_2 normalization.

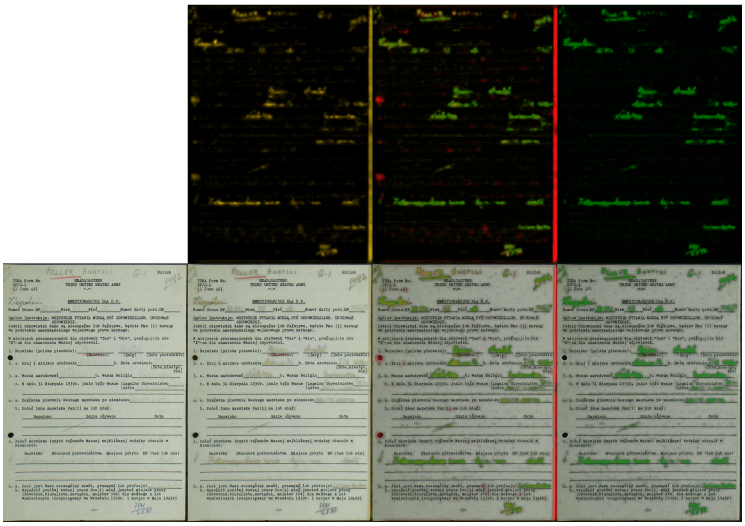


Fig. 4: Selectiveness of X-VLAD. From left to right: input page; local feature norms after the backbone; per-head keep vs. reject scores aggregated across heads; final keep score aggregated across heads.

4.3 End-to-End Representation Learning

We train the backbone encoder and X-VLAD jointly end-to-end. Training follows a two-stage protocol: (i) patch pretraining on handwriting-centered 32×32 patches sampled using the handwriting detector described in Section 3.2 to increase the signal-to-noise ratio, and (ii) full-page fine-tuning on complete pages to improve robustness to distractors and to reduce error propagation from imperfect handwriting detection. Optimization combines an ArcFace-style angular-margin classification loss [9] with a batch-hard triplet loss using a soft margin [37]. At inference time we use the normalized X-VLAD embedding directly for retrieval, without additional post-processing.

5 Results on FormWR

5.1 Implementation details

As a backbone, we use an ImageNet [8]-pretrained ResNet-18 [14] and remove the first two downsampling stages to retain high resolution activation maps (total downsampling factor is 1/8). We first pretrain on handwriting-centered 32×32 patches for 100 epochs using AdamW [22] (weight decay $2 \cdot 10^{-2}$) with a cosine schedule (peak LR $5 \cdot 10^{-4}$, 25k warmup steps). Patch pretraining uses batch size 128, where each batch element contains 64 patches. We apply Dropout (p=0.3) after the feature extractor. Augmentations include ColorJitter(brightness=contrast=saturation=0.3, hue=0.1) and, with probability 0.1 each,

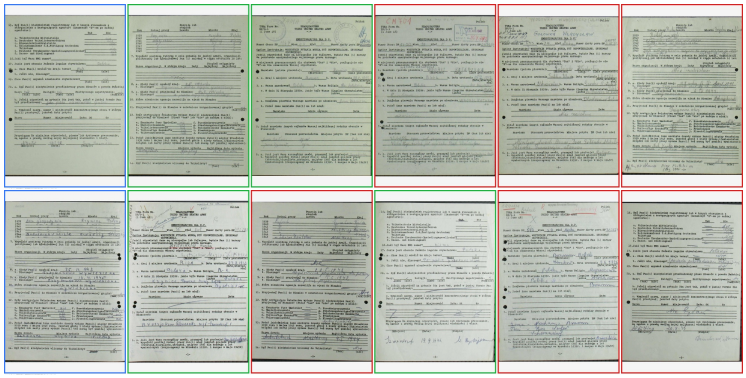


Fig. 5: Qualitative retrieval results. For each query (left), we show the top-5 retrieved pages and mark true matches. For the displayed queries, two relevant matches exist. Zoom in for details.

horizontal/vertical flips and grayscale conversion. Patches are extracted offline; augmentations are applied consistently per page (shared parameters across its patches). We then fine-tune on full pages (resized to 1568×1120 pixels) for 4 epochs (batch size 8, LR $5 \cdot 10^{-5}$). Model selection and ablations are performed on the FormWR validation split; final reporting and scaling are performed on the FormWR test split.

5.2 Ablation study

We ablate key design choices on the FormWR validation split. Table 2 shows that moving from a single head to N=4 yields the best patch-level mAP, while larger N degrades performance. Varying the number of discard modes g within each head has only a minor effect. Finally, combining ArcFace [9] with batch-hard triplet loss (soft margin) [15] is crucial: each loss alone underperforms substantially, while the joint objective improves mAP by $> 5\%$. We use the best-performing configuration (N=4, g=2, ArcFace+Triplet) in subsequent experiments.

Table 3 compares X-VLAD to other aggregators: a mean-pooling baseline, GhostVLAD [45], SuperVLAD [10] using the respective default configurations. Two key insights emerge: First, while global pooling achieves competitive results on the handwriting-centric patches and actually even outperforms GhostVLAD in this regime, adding full-page fine-tuning flips the results. The learned aggregators improve by around 11 mAP points, whereas mean-pooling totally collapses, consistent with the strong dominance of printed template regions over handwriting. Second, X-VLAD achieves the best performance while using a compact embedding: d_{out}=256 with N=4 heads and outperforms higher-dimensional alternatives, suggesting that head-wise low-dimensional specialization is effective without requiring large descriptor budgets.

Table 2: Patch-level ablations on the FormWR validation split (7k pages), reporting $\text{mAP}_{\text{patch}}$ ↑. (a) varies the number of heads at fixed per-head dimensionality d_{head}=64 (total dim Nd_{head}). (b) varies the number of discard modes. (c) compares loss criteria.

(a) Heads ($d_{\text{head}} = 64$)

N	$\text{mAP}_{\text{patch}}$ ↑
1	63.2
4	**63.7**
8	62.9
32	62.3

(b) Discard modes

g	$\text{mAP}_{\text{patch}}$ ↑
1	63.5
2	**63.7**
3	**63.7**
4	63.6

(c) Loss criterion

Method	$\text{mAP}_{\text{patch}}$ ↑
ArcFace	57.7
Triplet (batch-hard)	56.0
ArcFace + Triplet	**63.7**

Table 3: Aggregator comparison on the FormWR validation split. We report mAP after patch pretraining and after full-page fine-tuning (FT). $\Delta_{\text{patch}\to\text{FT}}$ denotes the absolute change in mAP.

Aggregator	d_{out}	$\text{mAP}_{\text{patch}}$ ↑	mAP_{FT} ↑	$\Delta_{\text{patch}\to\text{FT}}$
AvgPool	512	61.0	3.4	−57.6
$\text{GhostVLAD}_{g=1}^{k=8}$	4096	58.9	70.5	+11.6
$\text{SuperVLAD}_{g=1}^{k=4}$	2048	62.8	74.7	+11.9
X-VLAD$^{n=4}$ (Ours)	256	**63.7**	**75.4**	+11.7

5.3 Baselines and unsupervised reference methods

Table 4 compares the final X-VLAD configuration to two unsupervised reference methods on the FormWR validation split. As a handcrafted baseline we use RootSIFT descriptors [2] with VLAD encoding [17]. As a learned unsupervised baseline we use a self-supervised ViT feature extractor trained with AttMask [36] and apply VLAD to its local features. For both references, we tune the number of VLAD clusters on the training split and apply PCA whitening to $d_{\text{PCA}} = 256$. To reduce confounding from printed structure, we restrict feature extraction to handwriting regions (via our handwriting detector). Both unsupervised references lag far behind the supervised model, reinforcing that FormWR's combination of sparse handwriting and strong distractors makes label supervision materially beneficial at scale. As our method outperforms the reference methods by such a large margin, we only provide results of our method on the test set.

5.4 Visualizations and qualitative results

Fig. 4 illustrates how selectiveness emerges in X-VLAD. Already at the backbone level, the spatial distribution of feature norms correlates well with handwriting regions, indicating that the encoder learns to respond strongly to ink even in the

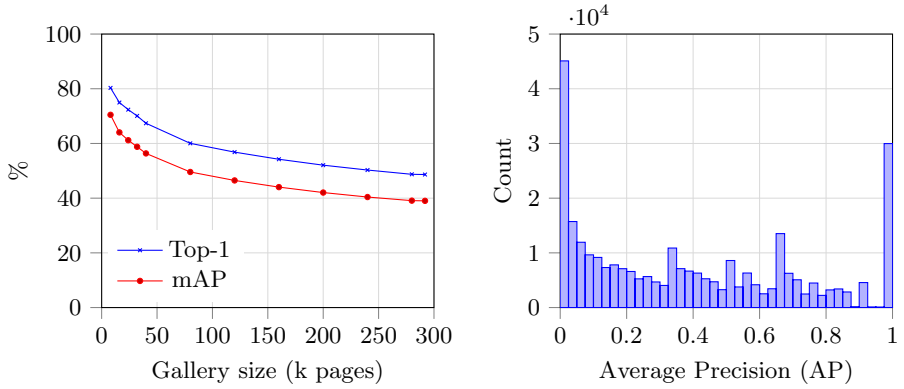
Table 4: Performance of our baseline method and two unsupervised reference methods on FormWR validation, and baseline results of our method on FormWR test.

Method	Validation		Test	
	Top-1 ↑	mAP ↑	Top-1 ↑	mAP ↑
RootSIFT + VLAD [2, 17]	7.11	6.7	–	–
AttMask ViT + VLAD [36]	35.1	30.2	–	–
X-VLAD$^{n=4}$ (Ours)	**84.7**	**75.4**	**48.6**	**39.0**

presence of a dominant printed template. The selection mechanism then sharpens this separation: across heads, keep scores consistently suppress background and machine print while focusing on handwriting, yielding a clean aggregated keep map that highlights writer-discriminative evidence. Fig. 5 shows exemplary retrieval outputs.

5.5 Scaling study

We evaluate the performance on the test split as a function of the gallery size. For this, we evaluate performance at different subset sizes. We always either include all pages for a writer proxy or none. Fig. 6 (a) shows that retrieval difficulty increases steadily with gallery size, as expected when hard negatives accumulate at scale. The largest changes occur when transitioning from small galleries to the large-scale regime (e.g., mAP decreases from 70.5 at 8k pages to 49.6 at 80k pages). Beyond that point, the curve flattens and performance degrades more gradually (e.g., from 42.1 at 200k pages to 39.0 at 292k pages), suggesting a logarithmic relationship between gallery size and retrieval performance. Fig. 6 (b) highlights that test-query difficulty is highly heterogeneous, with many samples either achieving near-perfect average precision (≈30k) or near-zero (≈45k), indicating a strong variation in difficulty.



(a) Scaling to the full test split.

(b) Distribution of AP over test queries.

Fig. 6: FormWR test split analysis: (left) scaling of Top-1/mAP with gallery size, (right) distribution of per-query retrieval difficulty via AP over test queries.

6 Results on HisFragIR20

We further evaluate on HisFragIR20 [38], a fragment-level benchmark with strong degradation and limited context. The dataset consists of around 100k fragments for training and a writer-disjoint test set of 20k fragments.

6.1 Implementation details

To sample handwriting-centered patches on fragments, we apply a simple edge-based heuristic: Canny edge detection followed by morphological closing and a final erosion step to suppress border responses while retaining ink strokes. For hyperparameter search, we hold out 10% of the training writers as a validation split; final models are trained on the full training set. We train for 70 epochs with a cosine learning-rate schedule (4k warmup steps, peak LR $2 \cdot 10^{-4}$, final LR $1 \cdot 10^{-4}$) and sample four fragments per writer in each batch. Due to the highly variable fragment sizes, we omit full-page fine-tuning and evaluate at patch level using up to 2k patches per fragment.

6.2 Comparison to prior work

Table 5 compares our approach to published results. Without re-ranking, our method achieves 78.8% mAP and 97.9% Top-1 accuracy, improving over prior work by +21.6% mAP (vs. 57.2% [30]) and +12.7% Top-1 (vs. 85.2% [25]). We also report results with SGR reranking [29], which increases mAP by 5.4 points at a small cost in Top-1 accuracy. Qualitative examples of remaining failure cases are given in Fig. 7. In most cases, either the query or the falsely retrieved fragment contains extremely little handwriting.

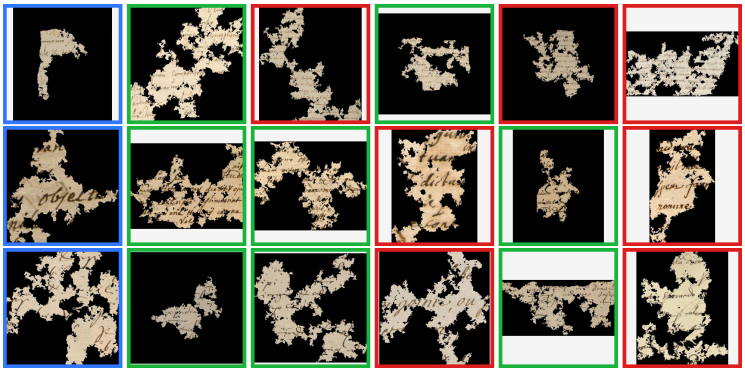


Fig. 7: Failure examples on HisFragIR20. For each query (blue) we show the top-5 retrieved fragments and mark correct (green) vs. incorrect (red) writers. Most errors occur when the query or retrieved fragments contain extremely sparse handwriting.

Table 5: Comparison with prior methods on HisFragIR20. † Reference numbers taken from [30].

Method	mAP ↑	Top-1 ↑
Cl-S + ResNet56 + NetRVLAD$^{k=100}$ [29]†	41.2	68.5
Feature Mixer [31]	44.0	81.9
ResNet34 + A-VLAD [25]	46.6	85.2
SAGHOG + NetRVLAD$^{k=100}$ [30]	57.2	81.7
X-VLAD$^{n=4}$ (Ours)	**78.8**	**97.9**
SAGHOG + SGR [30]	68.7	88.7
X-VLAD$^{n=4}$ + SGR$^{k=1,L=1}_{\gamma=0.08}$ (Ours)	**84.2**	**97.5**

7 Conclusion

We introduced FormWR, an archival-scale benchmark for writer retrieval, and X-VLAD, a supervised end-to-end aggregator that learns to keep writer-discriminative evidence and discard the printed structure dominating form pages. Two findings stand out. First, scale reshapes the problem: as the gallery grows to the full 290k-page test split, hard negatives accumulate and retrieval difficulty rises steadily—an effect invisible on small, clean datasets, and a concrete explanation for their apparent saturation. Second, supervision becomes decisive where the setting is hard and data is plentiful: unsupervised references that lead on small benchmarks fall well behind, while learned embeddings with selective aggregation hold up under full-page fine-tuning, where naive pooling collapses. These results caution against extrapolating from curated benchmarks and position selection—learning what to ignore—as central to retrieval at scale. Our main cost is reliance on labels; a natural next step is self-supervised pretraining on large unlabeled

collections with light supervised adaptation, bringing writer retrieval closer to the unlabeled reality of real archives.

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5297–5307. IEEE, Las Vegas, NV, USA (Jun 2016)
2. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: CVPR (2012)
3. Christlein, V., Bernecker, D., Angelopoulou, E.: Writer identification using VLAD encoded contour-Zernike moments. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 906–910 (2015)
4. Christlein, V., Bernecker, D., Hönig, F., Angelopoulou, E.: Writer identification and verification using GMM supervectors. In: IEEE Winter Conference on Applications of Computer Vision. pp. 998–1005. IEEE (2014)
5. Christlein, V., Gropp, M., Fiel, S., Maier, A.: Unsupervised Feature Learning for Writer Identification and Writer Retrieval. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 991–997 (Nov 2017)
6. Christlein, V., Maier, A.: Encoding CNN Activations for Writer Recognition. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS). pp. 169–174. IEEE (2018)
7. Christlein, V., Nicolaou, A., Seuret, M., Stutzmann, D., Maier, A.: ICDAR 2019 Competition on Image Retrieval for Historical Handwritten Documents. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1505–1509. IEEE (2019)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
9. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4685–4694. IEEE, Long Beach, CA, USA (Jun 2019)
10. Dong, S., Lan, X., Lu, F., Ye, C., Yuan, C., Zhang, L., Zhang, X.: SuperVLAD: Compact and Robust Image Descriptors for Visual Place Recognition. In: Advances in Neural Information Processing Systems 37. pp. 5789–5816. Neural Information Processing Systems Foundation, Inc. (NeurIPS), Vancouver, BC, Canada (2024)
11. Fiel, S., Kleber, F., Diem, M., Christlein, V., Louloudis, G., Nikos, S., Gatos, B.: ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI). In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 1377–1382. IEEE (2017)
12. Fiel, S., Sablatnig, R.: Writer identification and writer retrieval using the fisher vector on visual vocabularies. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 545–549. IEEE (2013)
13. Fiel, S., Sablatnig, R.: Writer identification and retrieval using a convolutional neural network. In: Computer Analysis of Images and Patterns: 16th International Conference, CAIP 2015, Valletta, Malta, September 2-4, 2015, Proceedings, Part II 16. pp. 26–37. Springer (2015)

14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
15. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
16. Izquierdo, S., Civera, J.: Optimal Transport Aggregation for Visual Place Recognition. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17658–17668. IEEE, Seattle, WA, USA (Jun 2024)
17. Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C.: Aggregating Local Image Descriptors into Compact Codes. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(9), 1704–1716 (2011)
18. Jo, J., Koo, H.I., Soh, J.W., Cho, N.I.: Handwritten text segmentation via end-to-end learning of convolutional neural networks. Multimedia Tools and Applications **79**(43), 32137–32150 (2020)
19. Jordan, S., Seuret, M., Král, P., Lenc, L., Martínek, J., Wiermann, B., Schwinger, T., Maier, A., Christlein, V.: Re-ranking for writer identification and writer retrieval. In: Document Analysis Systems: 14th IAPR International Workshop, DAS 2020, Wuhan, China, July 26–29, 2020, Proceedings 14. pp. 572–586. Springer (2020)
20. Kleber, F., Fiel, S., Diem, M., Sablatnig, R.: CVL-DataBase: An Off-Line Database for Writer Retrieval, Writer Identification and Word Spotting. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 560–564 (Aug 2013)
21. Lai, S., Zhu, Y., Jin, L.: Encoding Pathlet and SIFT Features With Bagged VLAD for Historical Writer Identification. IEEE Transactions on Information Forensics and Security **15**, 3553–3566 (2020)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Louloudis, G., Gatos, B., Stamatopoulos, N., Papandreou, A.: ICDAR 2013 Competition on Writer Identification. In: 2013 12th International Conference on Document Analysis and Recognition (ICDAR). pp. 1397–1401. IEEE (2013)
24. Marti, U.V., Bunke, H.: The IAM-database: An English sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition **5**(1), 39–46 (Nov 2002)
25. Ngo, T.T., Nguyen, H.T., Nakagawa, M.: A-VLAD: An end-to-end attention-based neural network for writer identification in historical documents. In: Document Analysis and Recognition – ICDAR 2021 (2021)
26. Okabe, K., Koshinaka, T., Shinoda, K.: Attentive Statistics Pooling for Deep Speaker Embedding. In: Interspeech 2018. pp. 2252–2256. ISCA (Sep 2018)
27. Peer, M., Kleber, F., Sablatnig, R.: Self-supervised Vision Transformers with Data Augmentation Strategies Using Morphological Operations for Writer Retrieval. In: Porwal, U., Fornés, A., Shafait, F. (eds.) Frontiers in Handwriting Recognition, vol. 13639, pp. 122–136. Springer International Publishing, Cham (2022)
28. Peer, M., Kleber, F., Sablatnig, R.: Writer Retrieval using Compact Convolutional Transformers and NetMVLAD. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 1571–1578 (2022)
29. Peer, M., Kleber, F., Sablatnig, R.: Towards Writer Retrieval for Historical Datasets. In: International Conference on Document Analysis and Recognition. pp. 411–427. Springer (2023)
30. Peer, M., Kleber, F., Sablatnig, R.: SAGHOG: Self-supervised Autoencoder for Generating HOG Features for Writer Retrieval. In: Barney Smith, E.H., Liwicki, M.,

Peng, L. (eds.) Document Analysis and Recognition - ICDAR 2024. pp. 121–138. Springer Nature Switzerland, Cham (2024)
31. Peer, M., Sablatnig, R.: Feature mixing for writer retrieval and identification on papyri fragments. In: Proceedings of the 7th International Workshop on Historical Document Imaging and Processing, HIP@ICDAR 2023, San Jose, CA, USA, August 25-26, 2023. pp. 31–36. ACM (2023)
32. Peer, M., Sablatnig, R., Kleber, F.: Towards the Influence of Text Quantity on Writer Retrieval. In: Yin, X.C., Karatzas, D., Lopresti, D. (eds.) Document Analysis and Recognition – ICDAR 2025. pp. 129–145. Springer Nature Switzerland, Cham (2026)
33. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) Computer Vision – ECCV 2010. pp. 143–156. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
34. Radenović, F., Tolias, G., Chum, O.: Fine-tuning CNN image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(7), 1655–1668 (2019)
35. Rasoulzadeh, S., BabaAli, B.: Writer identification and writer retrieval based on NetVLAD with Re-ranking. IET Biometrics **11**(1), 10–22 (Jan 2022)
36. Raven, T., Matei, A., Fink, G.A.: Self-supervised Vision Transformers for Writer Retrieval. In: Barney Smith, E.H., Liwicki, M., Peng, L. (eds.) Document Analysis and Recognition - ICDAR 2024. pp. 380–396. Springer Nature Switzerland, Cham (2024)
37. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823. IEEE, Boston, MA, USA (Jun 2015)
38. Seuret, M., Nicolaou, A., Stutzmann, D., Maier, A., Christlein, V.: ICFHR 2020 competition on image retrieval for historical handwritten fragments. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 216–221. IEEE (2020)
39. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proceedings Ninth IEEE International Conference on Computer Vision. pp. 1470–1477 vol.2 (2003)
40. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-Vectors: Robust DNN Embeddings for Speaker Recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5329–5333. IEEE, Calgary, AB (Apr 2018)
41. Wan, L., Wang, Q., Papir, A., Moreno, I.L.: Generalized end-to-end loss for speaker verification. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4879–4883. IEEE (2018)
42. Wolf, F., Tüselmann, O., Matei, A., Hennies, L., Rass, C., Fink, G.A.: CM1 - a dataset for evaluating few-shot information extraction with large vision language models. In: Yin, X.C., Karatzas, D., Lopresti, D. (eds.) Document Analysis and Recognition – ICDAR 2025. pp. 23–39. Springer Nature Switzerland, Cham (2026)
43. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? In: International Conference on Learning Representations (ICLR) (2019)
44. Zaheer, M., Kottur, S., Ravanbhakhsh, S., Póczos, B., Salakhutdinov, R., Smola, A.: Deep sets. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
45. Zhong, Y., Arandjelović, R., Zisserman, A.: GhostVLAD for set-based face recognition. In: Proceedings of the Asian Conference on Computer Vision (ACCV). pp. 35–50. Springer (2018)