HIDRA: Hierarchical Ink-aware Dual-granularity Retrieval Architecture for Historical Fragments

Jihad Al Akl[1,2⋆], Chady Abou Jaoude[1], Zahi Al Chami[1], Marianne Abi Kanaan[1], and Abdallah Makhoul[3]

[1] Ticket Lab, Antonine University, Hadat–Baabda, Lebanon
{jihad.alakl, chady.aboujaoude, zahi.chami, marianne.abikanaan}@ua.edu.lb
[2] Université Marie et Louis Pasteur, FEMTO-ST Institute, CNRS, Belfort, France
[3] Université Marie et Louis Pasteur, FEMTO-ST Institute, CNRS, Montbéliard, France
abdallah.makhoul@univ-fcomte.fr

Abstract. Historical handwritten fragments often contain limited amounts of text and substantial background noise, which may lead to an unreliable writer retrieval. In this work, we introduce HIDRA, a transformer-based retrieval model that suppresses background noise without explicit ink segmentation by deriving a foreground weighting directly from a self-supervised ViT backbone. HIDRA then aggregates the foreground weighted patch features at two complementary scales: a fine scale to capture stroke details, and a coarse scale to remain stable under severe fragmentation. The resulting global–local descriptor is fused into a compact embedding for retrieval. On HisFrag20, HIDRA achieves 68.4% mAP for writer retrieval, and 45.72% mAP for page retrieval - a margin of 11.2% and 10.4% to the current state of the art, respectively. We also observe strong cross-domain transfer to ScriptNet and GRK-Papyri.

Keywords: Writer Identification · Historical Document Retrieval · Vision Transformer · DINOv2 · Local Feature Aggregation · Metric Learning · HisFrag20

1 Introduction

Writer retrieval and writer identification aim to match handwriting across documents, enabling scribal analysis, and authorship attribution. The problem becomes significantly harder while processing historical fragments: torn or cropped patches with little text, partial characters, and heavy degradation such as stains, bleed-through, and paper texture. In this context, discriminative cues are sparse and local, while background patterns can be misleading, causing models to rely on page texture instead of handwriting.

Many traditional pipelines for writer identification aggregate handcrafted local descriptors into global vectors [20], while modern deep models learn embeddings using end-to-end approaches [5]. However, fragment retrieval exposes two

⋆ Corresponding author.

frequent weaknesses. First, approaches that rely on binarization or explicit ink segmentation tend to achieve reduced performance on degraded manuscripts [33].

Second, approaches that use deep embeddings implicitly capture background cues within a page alongside handwritten features [28,27]. Such feature interference increases similarity scores between fragments with similar background textures, which negatively impacts the objective of isolating writer identities specific characteristics.

These observations indicate three requirements for fragment-level retrieval: (i) suppress background without relying on hard segmentation, (ii) detect sparse stroke level cues while limiting sensitivity to noise, and (iii) avoid post-processing steps that amplify texture-based false neighbors.

HIDRA, our proposed approach, addresses these requirements through three complementary components: (i) an *ink-aware foreground prior* extracted from a strong self-supervised ViT backbone (DINOv2 [24]) to suppress background without external segmentation; (ii) *dual-granularity local aggregation* (DGA) that is robust to fragmentation; and (iii) a *fragment-aware post-processing pipeline* that replaces generic retrieval heuristics with confidence-gated query expansion and PCA whitening [16] adapted to the noisy local structure of document fragments.

Our novel approach combines feature extraction using DINOv2, dual granularity aggregation, and fusion & loss optimization. Therefore, the contributions of this work are summarized as follows:

- **Foreground suppression:** Compute an ink-focused foreground weighting from CLS patch similarity in DINOv2 [24], eliminating external segmentation.
- **Local representation:** Balance detail (strokes) and robustness (fragment noise) by using Pool patch tokens at dual granularities with attention-gated VLAD [23].
- **Training signal:** Introduce a hierarchical contrastive objective that enforces stronger similarity for fragments from the same page than for fragments from different pages of the same writer.
- **Inference:** Use confidence-gated query expansion and unsupervised whitening adapted to document fragments.
- **Results:** Achieve state-of-the-art retrieval on HisFrag20 [30] and strong cross-domain transfer to ScriptNet-HistoricalWI-2017 [14] and GRK-Papyri [22].

The rest of the paper presents the related work (Sec. 2), details the architecture (Sec. 3), and highlights the conducted set of experiments to validate each component through ablations and cross-dataset evaluation (Sec. 4).

2 Related Work

2.1 Writer Identification and Retrieval

Early WR/WI approaches used handcrafted features (e.g., SIFT/pathlets) with bag-of-words or VLAD-style aggregation [20]. VLAD encodings have also been

explored with handcrafted moment descriptors for WI [8]. Deep metric learning and CNN-based embeddings improved robustness, including triplet-based models for writer retrieval [18]. For fragment-level benchmarks such as HisFrag20 [30], recent methods emphasize feature aggregation and page-level context. SAGHOG [27] combines self-supervised HOG pretraining with NetRVLAD encoding, achieving 57.2% mAP on HisFrag20 without re-ranking. Feature Mixer [28] employs random feature mixing to decorrelate background information, demonstrating the importance of background-aware feature extraction.

Vision Transformers (ViTs) [13] have become competitive feature extractors, particularly when pretrained with self-supervised objectives. DINO [3] and DINOv2 [24] produce strong, semantically meaningful token representations with emergent localization properties, which we exploit to derive a foreground prior without training a segmentation model.

Several approaches exist for local feature aggregation for retrieval. Global pooling can underrepresent local discriminative cues, motivating aggregation schemes that encode sets of local descriptors. VLAD [17] and its learnable variants (e.g., NetVLAD [1]) are widely used in retrieval pipelines. In historical WR/WI, attention-based VLAD variants such as A-VLAD [23] were proposed to emphasize informative regions. HIDRA follows this line but uses a lightweight, backbone-derived attention gate and explicitly combines *two* granularities of aggregation.

Deep retrieval is commonly trained with margin-based classification heads (e.g., ArcFace [11]) and/or contrastive learning [7,19]. We adopt a hybrid approach, using ArcFace for primary writer supervision and introducing a Hierarchical Contrastive Loss (HCL) to properly model the intra-page and inter-page relationships inherent to fragmented document datasets.

2.2 Post-Processing for Retrieval

Standard post-processing methods such as query expansion (QE) [10] and k-reciprocal re-ranking [34] can improve retrieval when the nearest-neighbor set is reliable. However, as we show empirically, these methods can degrade performance on historical fragments where the initial ranking is contaminated by texture similarity rather than writer similarity. PCA whitening [16] is an unsupervised technique that removes correlated dimensions from the embedding space, and is found to be particularly effective for fragment retrieval where background texture introduces strong correlations.

2.3 Summary

Prior work [28,15] shows that background leakage is a major failure mode in fragment retrieval, and that local aggregation is important for sparse handwriting cues. However, existing approaches [20,9] either require explicit segmentation or rely on retrieval heuristics that assume clean neighborhoods. HIDRA addresses these issues by deriving an ink-aware foreground weighting from ViT tokens and combining dual-scale aggregation with fragment-aware inference.

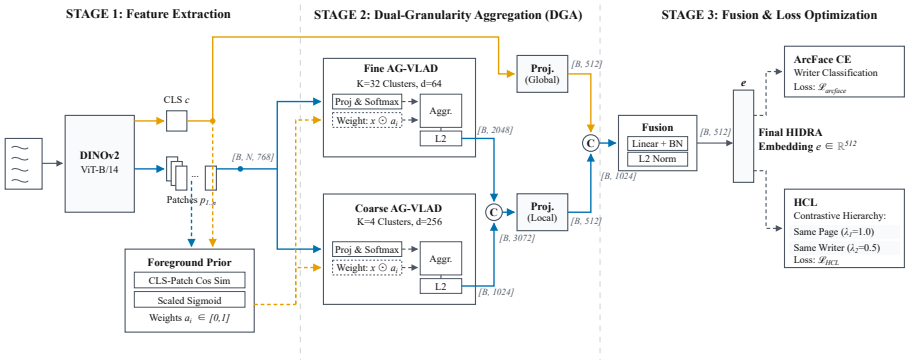

Fig. 1. HIDRA overview. **Stage 1:** DINOv2 outputs a CLS token and patch tokens; their cosine similarity yields foreground weights a_i. **Stage 2:** Dual AG-VLAD aggregates a_i-weighted patches at fine (32 clusters) and coarse (4 clusters) granularities. **Stage 3:** Global and local descriptors are fused into a 512-D retrieval embedding.

3 Proposed solution

3.1 Overview

HIDRA converts a fragment image into a compact 512-D embedding for retrieval in three stages (Fig. 1):

1. **Feature Extraction (Sec. 3.2).** A pretrained DINOv2 backbone tokenizes the image into a global CLS token and N local patch tokens. A lightweight foreground prior estimates which patches contain ink vs. background.
2. **Dual-Granularity Aggregation (Sec. 3.3).** Two parallel AG-VLAD branches pool the foreground-weighted patch tokens into local descriptors at different levels: fine (32 clusters, stroke details) and coarse (4 clusters, overall style).
3. **Global-Local Fusion and Loss Optimization (Sec. 3.4).** The global CLS embedding and the local DGA descriptor are fused through an MLP into a single 512-D embedding, trained with ArcFace and our Hierarchical Contrastive Loss (Sec. 3.5).

At test time, writer retrieval is performed by ranking cosine similarities between fragment embeddings.

3.2 Stage 1: Feature Extraction

Backbone tokenization. We feed the fragment image into a pretrained DINOv2 ViT-B/14 [24], which divides it into a grid of 14×14-pixel patches and outputs one token per patch:

$$T = [c,\ p_1, \ldots, p_N], \quad c, p_i \in \mathbb{R}^D. \tag{1}$$

Here c is the CLS token, which is a learned global summary of the entire image, and $p_1, \ldots, p_N$ are local patch tokens. For a 448×448 input, N=1024. The backbone is initialized from public DINOv2 weights and finetuned after a brief warm-up.

Foreground prior: which patches contain ink? Historical fragments are mostly background: blank parchment, stains, tears, and/or ink bleed. We need the model to focus on the few patches that contain actual writing. DINOv2's CLS token naturally encodes a "content summary" of the image [3], and ink-bearing patches tend to align with this summary while blank patches do not. We exploit this by computing a per-patch foreground weight:

$$s_i = \frac{c^\top p_i}{\|c\| \, \|p_i\|}, \qquad a_i = \sigma(\tau \cdot s_i) \in (0, 1), \tag{2}$$

where τ is a single learnable temperature and σ is the sigmoid. A patch of dense handwriting will receive $a_i \approx 1$; a blank parchment corner will receive $a_i \approx 0$. This costs only one extra parameter and requires no external segmentation. We use these weights to downweight background patches before completing local aggregation.

3.3 Stage 2: Dual-Granularity Aggregation

Using only the CLS token (global pooling) discards the spatial detail of individual strokes. This is problematic for fragments where the discriminative cue may be a single distinctive letter or connection stroke. We therefore aggregate patch tokens into *local descriptors* using a VLAD-style mechanism [17,1], adapted for fragment noise.

AG-VLAD: Each AG-VLAD branch performs four steps (see Fig. 1, Stage 2):

1. **Project.** Each patch token is mapped to a lower-dimensional space: $z_i = \mathrm{LN}(W p_i) \in \mathbb{R}^d$.
2. **Assign.** Each patch is soft-assigned to K visual word clusters via $\alpha_{ik} = \mathrm{softmax}_k(w_k^\top z_i)$. This determines *which* stroke pattern each patch represents.
3. **Weight by a_i.** The foreground prior scales each patch's contribution: ink patches ($a_i \approx 1$) contribute fully, background patches ($a_i \approx 0$) are suppressed.
4. **Aggregate.** We sum the weighted tokens per cluster:

$$v_k = \sum_{i=1}^{N} a_i \cdot \alpha_{ik} \cdot z_i, \tag{3}$$

followed by intra-normalization of each v_k and global L_2 normalization [17].

Why two granularities? A single AG-VLAD branch faces a trade-off: many clusters capture fine details but are noise-sensitive; few clusters are robust but lose detail. To balance detail and robustness, we run *two branches in parallel*:

- **Fine branch** (K_f=32, d_f=64): 32 visual words that encode micro-stroke textures, serifs, and ligature shapes.
- **Coarse branch** (K_c=4, d_c=256): 4 broad clusters that summarize overall writing style (slant, stroke width, spacing), and remain stable even on heavily damaged fragments.

We concatenate the fine and coarse descriptors, $d = [\text{AGVLAD}_f;\ \text{AGVLAD}_c]$, and then fuse them with the global token to form the final embedding.

3.4 Stage 3: Global–Local Fusion

We now have two complementary views of the fragment: the CLS token c (a holistic summary) and the DGA descriptor d (detailed stroke-level information). We project each to 512 dimensions and fuse:

$$g = W_g\, c, \quad \ell = W_\ell\, d, \quad e = \frac{f([g;\ell])}{\|f([g;\ell])\|_2} \in \mathbb{R}^{512}, \tag{4}$$

where:

- g is a global or content feature vector that is generated by multiplying a weight matrix W_g by an input context/control vector c.
- l is a local or positional feature vector that is generated by multiplying a weight matrix W_ℓ by an input input d.
- $[g;\ell]$ is the concatenation of vectors g and ℓ.
- f is a two-layer MLP with GELU activation.

The final embedding e is L_2-normalized, making cosine similarity equivalent to a dot product.

3.5 Training Objectives

We train HIDRA using a joint objective that balances robust global writer classification with local structural coherence.

ArcFace for Writer Supervision. We apply an ArcFace head [11] over embeddings e to classify writers with an additive angular margin, optimized with cross entropy. This provides the primary discriminative signal across the dataset.

Hierarchical Contrastive Loss (HCL). Historical fragment datasets have a natural three-level hierarchy: a *writer* produces multiple *pages*, and each page is broken into multiple *fragments*. ArcFace alone does not model this structure; it treats all fragments from the same writer equally. But intuitively, two fragments from the *same page* should be closer in embedding space than two fragments from *different pages* of the same writer, because same-page fragments share identical ink, paper texture, and writing context.

HCL encodes this intuition as a weighted contrastive loss. For each pair of embeddings (e_i, e_j) in a batch, we assign a positive weight:

$$w_{ij} = \begin{cases} \lambda_p & \text{same page (strongest pull),} \\ \lambda_w & \text{same writer, different page (weaker pull),} \\ 0 & \text{different writers (negative pair).} \end{cases} \tag{5}$$

With temperature t, we optimize:

$$\mathcal{L}_{\text{HCL}} = -\frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \frac{\sum_{j \neq i} w_{ij} \, \log \frac{\exp(e_i^\top e_j / t)}{\sum_{k \neq i} \exp(e_i^\top e_k / t)}}{\sum_{j \neq i} w_{ij} + \epsilon}, \tag{6}$$

where $\mathcal{V}$ indexes anchors with at least one positive.

3.6 Fragment-Aware Post-Processing

Retrieval systems [29] commonly improve results with post-processing steps, such as query expansion and re-ranking. However, these techniques assume that the top-ranked results are already mostly correct. This assumption is not effective when retrieving noisy fragments since two fragments may rank highly for sharing paper texture rather than writer identity. We therefore design two fragment-aware alternatives.

Confidence-Gated Query Expansion (CG-QE). Standard query expansion [10] refines each query by averaging it with its top-k neighbors. On fragments, this can corrupt the query because some top-k neighbors are texture matches, not writer matches. We fix this by only including neighbors whose similarity exceeds a confidence threshold θ:

$$e_i' = \alpha \, e_i + (1-\alpha) \frac{\sum_{j \in \mathcal{N}_k(i)} \mathbf{1}[s_{ij} > \theta] \, s_{ij} \, e_j}{\sum_{j \in \mathcal{N}_k(i)} \mathbf{1}[s_{ij} > \theta] \, s_{ij} + \epsilon}. \tag{7}$$

where:

- α **(Residual Weight):** A hyperparameter controlling the balance between the original query e_i and the aggregated neighborhood information.
- s_{ij} **(Cosine similarity Score):** Measures the relationship between the i-th query and element j in its top-k neighbors.

- $\mathbf{1}(\cdot)$ **(Indicator Function):** Acts as a hard threshold. It equals 1 if $s_{ij} > \theta$, and 0 otherwise, effectively pruning weak or noisy edges.
- **Weighted Aggregation:** The term

$$\sum_j \mathbf{1}(s_{ij} > \theta)\, s_{ij}\, e_j$$

  ensures that more similar neighbors contribute more heavily to the updated embedding e'_i.
- ϵ **(Smoothing Constant):** A very small value added to the denominator to prevent division by zero when a query has no neighbors exceeding the threshold θ.

If no neighbor passes the threshold, the original query is kept unchanged.

PCA Whitening. The raw embedding space contains dimensions that encode domain-specific signals, such as paper texture and scan artifacts, rather than writer identity. PCA whitening [16] removes these correlated dimensions by projecting onto the principal components and normalizing their variance:

$$e'_i = \frac{W(e_i - \mu)}{\|W(e_i - \mu)\|_2}, \quad W = U_d \,\mathrm{diag}(S_d^{-1/2}), \tag{8}$$

where U_d and S_d are the top-d eigenvectors and eigenvalues of the gallery covariance matrix, and μ is the gallery mean. This removes dominant texture-correlated directions that are informative for page identity but harmful for writer discrimination.

4 Experiments

4.1 Dataset and Protocol

We evaluate primarily on HisFrag20 [30,31], introduced for the ICFHR 2020 competition on historical handwritten fragment retrieval. The test split contains 20,019 fragments from 8,717 writers. Following the standard evaluation [27], we report mean Average Precision (mAP) and Top-1 accuracy (R@1) for writer retrieval and page retrieval, based on cosine similarity between embeddings.

For cross-dataset evaluation, we deliberately use ScriptNet-HistoricalWI-2017-color [14] and GRK-Papyri [22] as a domain-shift, by testing on a different language and material.

Since these datasets provide full-page images and our model operates on fragments, we apply synthetic fragmentation [27,28]. Specifically, we extract random 448×448 crops with up to 50 random placement attempts per crop before any model sees the data. For ScriptNet (3,600 pages, 720 writers), we extract 6 crops per page, resulting in ~21,600 fragments. For GRK-Papyri (50 pages, 10 writers), we extract 8 crops per page, resulting in ~400 fragments. Crops with less than 5% ink coverage (estimated via Otsu binarization [25]) are

discarded and replaced with a new random crop. We note that this threshold is intentionally conservative: in practice, fewer than 3% of crops are discarded, and the filter primarily removes pure-margin crops that would be uninformative for any retrieval method. Sensitivity to this threshold is minimal; raising it to 10% changes writer mAP by less than ±0.3 points on both datasets.

4.2 Implementation Details

We use a DINOv2 ViT-B/14 backbone [24] and input size 448 × 448 (grayscale replicated to 3 channels). We train for 60 epochs with AdamW [21] and OneCycle learning rate scheduling [32]. Each batch samples 16 writers with 4 fragments per writer ($B = 64$). We apply lightweight augmentations using Albumentations [2], including small rotations (±5°), Gaussian blur, contrast jitter, and coarse dropout [12]. The backbone is frozen for the first 5 epochs then finetuned end-to-end with mixed precision training on an NVIDIA A100 80 GB GPU. We use Exponential Moving Average (EMA) with decay 0.999 for the inference checkpoint.

For CG-QE, we set k=5, α=0.3, θ=0.65. These values were selected by a coarse grid search on a 10% held-out subset of the *training* set (not the test set); Sec. 4.5 shows that performance is stable across a wide range. For PCA whitening, we retain the full d=512 dimensions. Multi-scale inference uses scales $\{0.85, 1.0, 1.15\}$ with Generalized Mean pooling (p=3).

We report efficiency metrics in Table 1. We present model size, inference speed, and memory. HIDRA adds only 11.75 M parameters over the frozen DINOv2-B backbone (86 M). Single-image latency is 13.1 ms on an A100; batch processing at B=64 achieves 589 img/s. The 512-D embedding requires 2 KB per fragment; the full HisFrag20 gallery (20K fragments) fits in ∼39 MB.

Table 1. Efficiency measured on a single NVIDIA A100 (80 GB) with mixed precision. Embedding storage assumes float32.

	Params	Head params	Emb dim	Latency(1 img)	Throughput (B=64)
HIDRA	98.1 M	11.75 M	512	13.1 ms	589 img/s

4.3 Comparison to State of the Art

Table 2 compares HIDRA to prior work on HisFrag20. We report both model-only performance (center crop, no post-processing) and our best post-processing configuration. Without any post-processing, HIDRA already achieves 68.40% writer mAP, exceeding the strongest model-only baseline (SAGHOG + NetRVLAD512, 57.2%) by +11.2 points. With transductive PCA whitening (computed on the evaluation gallery, no labels), mAP reaches **79.50%**.

Table 2. HisFrag20 writer retrieval results (mAP / Top-1). Baselines reproduced from [27]. "–" indicates not reported. SGR uses extra graph-based re-ranking. *Transductive: PCA whitening computed on the evaluation gallery (unsupervised, no labels used).

Method	Writer retrieval		Page retrieval	
	mAP	Top-1	mAP	Top-1
Cl-S + mVLAD3200 + ESVM [4]	33.7	68.9	–	–
FragNet512 [30]	33.5	77.1	22.6	36.4
ResNet50 + NetVLAD6400 + ESVM [6]	38.1	77.9	–	–
Cl-S + ResNet56 + NetRVLAD512 [26]	41.2	68.5	23.5	33.3
Feature Mixer256 [28]	44.0	81.9	29.3	45.0
A-VLAD500 [23]	46.6	85.2	–	–
SAGHOG + NetRVLAD512 [27]	57.2	81.7	35.2	47.6
SAGHOG + SGR (re-ranking) [27,26]	68.7	88.7	**54.5**	**70.6**
HIDRA (model-only)	68.40	94.52	45.66	70.6
HIDRA + CG-QE	76.26	94.16	48.99	69.35
HIDRA + Whiten*	**79.50**	**99.24**	–	–

Table 3. Component ablations on HisFrag20. "DGA w/o fg prior" sets a_i=1 for all patches at inference (same trained checkpoint).

Variant	Writer mAP	Writer R@1	Page mAP	Page R@1
Global CLS only	62.25	92.32	40.89	66.09
Fine AG-VLAD only	64.39	93.07	42.99	67.42
Coarse AG-VLAD only	37.59	81.29	27.72	52.10
DGA w/o fg prior (a_i=1)	68.40	94.51	45.64	70.58
Full HIDRA (DGA + fg prior)	**68.50**	**94.58**	**45.68**	**70.63**

4.4 Ablations and Analysis

We report component ablations on the HisFrag20 test split (N = 20,019 fragments). Unless stated otherwise, ablations use the same checkpoint and protocol as the main model.

The dual-granularity configuration provides the strongest performance (+6.16 writer mAP over global-only and +30.82 over coarse-only). Disabling the foreground prior at inference (setting a_i=1 for all patches) costs 0.10 writer mAP, confirming that the prior provides a measurable benefit even after training has converged. Qualitative attention maps (Fig. 2) further confirm that the learned prior correctly highlights strokes over texture.

Impact of Hierarchical Contrastive Learning As shown in Table 4, adding HCL to ArcFace provides a small but consistent improvement in writer mAP (68.27 → 68.43, +0.16). While the mAP gain is modest, we include HCL as an architectural contribution for two reasons. First, it acts as a *structural regularizer*:

Table 4. Loss-function ablations on HisFrag20 (paired checkpoints at epoch 60).

Objective	Writer mAP	Writer R@1	Page mAP	Page R@1
ArcFace only	68.27	94.78	46.07	71.47
ArcFace + HCL	**68.43**	**94.87**	45.82	71.24

Table 5. Post-processing ablations on HisFrag20 writer retrieval (full test set, $N = 20{,}019$). *Transductive: PCA computed on evaluation gallery without labels.

Setting	Writer mAP	Writer R@1
Center crop (model-only)	68.40	94.52
Standard techniques:		
+ TTA (5-crop)	68.40	94.52
+ k-Reciprocal Re-ranking [34]	61.01	84.34
+ CG-kR ($k = 20, \theta = 0.65$)	71.87	93.54
Fragment-aware (ours):		
+ Multi-Scale GeM ($p = 3$)	69.06	94.88
+ Confidence-Gated QE ($k = 5, \theta = 0.65$)	76.26	94.16
+ PCA Whitening* [16]	**79.50**	**99.24**

by explicitly enforcing same-page coherence (λ_p) alongside writer consistency (λ_w), it prevents the margin-based ArcFace head from aggressively separating visually dissimilar fragments that originate from the same physical page. Second, HCL requires proper hierarchical batch sampling (multiple pages per writer, multiple fragments per page) to produce a meaningful gradient signal. With the standard writer-only sampler used in prior work, the proportion of same-page pairs per batch is negligible. We therefore view the sampler–loss co-design as the key insight, and expect the HCL benefit to grow with improved sampling strategies in future work.

4.5 Post-Processing Analysis

We evaluate the impact of various post-processing techniques on HisFrag20 writer retrieval. Table 5 demonstrates that standard techniques designed for person re-identification (TTA, k-reciprocal re-ranking) *degrade* performance on document fragments, while our fragment-aware alternatives provide significant improvements. We further evaluate a confidence-gated k-reciprocal variant (CG-kR), where a candidate is admitted to the reciprocal neighborhood only if its mutual similarity exceeds the same threshold used by CG-QE (θ=0.65).

Why standard TTA fails. The 5-crop TTA (center + 4 corners) performs no better than center crop alone. On fragments, corner crops frequently sample regions devoid of handwriting (pure background or margins), producing uninformative embeddings that render the average negligible.

Why k-reciprocal re-ranking fails. K-reciprocal re-ranking [34] assumes an initially pure nearest-neighbor set, but on degraded fragments the top-ranked neighbors can share paper texture, bleed-through patterns, or similar paleographic artifacts rather than writer identity. This causes the reciprocal neighborhood expansion to amplify initial errors, reducing mAP by 7.4 points. Adding the same confidence gate used by CG-QE partially mitigates this failure: CG-kR improves over plain k-reciprocal re-ranking (71.87 vs. 61.01 mAP), but remains below CG-QE because hard reciprocal neighborhoods can still reject valid writer matches that are asymmetric under severe fragmentation.

Why CG-QE and whitening succeed. Confidence-gated QE avoids these pitfalls by only incorporating neighbors exceeding a similarity threshold (θ=0.65), effectively filtering out texture matches without imposing a strict reciprocal-neighborhood constraint. PCA whitening removes the dominant correlated directions that encode domain-specific signals (paper texture, scan artifacts), revealing the underlying writer-discriminative structure. The +11.1 mAP improvement from whitening alone confirms that a significant portion of the embedding variance encodes non-writer information.

Note on transductive evaluation. PCA whitening is computed on the evaluation gallery in an unsupervised manner (no labels are used), following the standard protocol of [16]. This is a transductive post-processing step that adapts the embedding space to the statistics of the test set. We therefore clearly distinguish our *model-only* result (68.40% mAP) from the post-processed result (79.50% mAP) throughout the paper, and emphasize that the model-only performance already surpasses all prior model-only baselines.

CG-QE Hyperparameter Sensitivity Table 6 reports CG-QE performance when varying each hyperparameter while fixing the others at their default values (k=5, α=0.3, θ=0.65). Performance is stable across a wide range of thresholds ($\theta \in [0.50, 0.70]$) and neighborhood sizes ($k \in [3, 15]$), confirming that the results are not sensitive to a specific threshold choice. The residual weight α mainly trades mAP against R@1: lower values place more weight on the confident neighborhood and slightly improve mAP, while higher values preserve more of the original embedding and improve nearest-neighbor accuracy. All hyperparameters were selected via grid search on a 10% held-out subset of the training data.

4.6 Cross-domain Dataset Generalization

To evaluate the domain-agnostic properties of the learned representations, we perform cross-domain dataset evaluation on two external datasets using synthetic fragmentation (Sec. 4.2).

Table 7 reports the results.

ScriptNet-HistoricalWI-2017-color shares similar morphological characteristics and writing surfaces with HisFrag20 (Latin manuscripts on parchment). HIDRA achieves an impressive 80.41% mAP in a purely cross-domain setting, demonstrating that the learned features generalize well across time periods and individual scribal hands when the fundamental material domain remains similar.

Table 6. CG-QE sensitivity analysis on HisFrag20. Each row varies one parameter while fixing the others at defaults (k=5, α=0.3, θ=0.65). Best in **bold**; used values are marked with $*$.

Param	Value	Writer mAP	Writer R@1
	0.50	76.33	93.96
	0.55	76.36	94.00
	0.60	**76.38**	94.08
θ	0.65$*$	76.23	94.18
	0.70	75.80	94.33
	0.75	74.73	94.31
	0.80	72.54	**94.34**
	1	73.18	93.81
	2	75.20	94.14
	3	75.84	**94.39**
k	5$*$	76.23	94.18
	7	**76.25**	94.01
	10	76.17	93.98
	15	75.99	93.92
	0.1	**76.37**	93.43
	0.3$*$	76.23	94.18
α	0.5	75.42	94.60
	0.7	73.67	**94.87**
	0.9	70.55	94.75

GRK-Papyri presents a more challenging domain shift: Greek script on fibrous papyrus substrates, substantially different from the Latin parchment fragments in HisFrag20. Despite this gap, HIDRA achieves 44.03% mAP with very high R@1 (98.85%), suggesting that the model can reliably identify the closest writer match even when the overall ranking is degraded by the domain shift. The high R@1 but moderate mAP pattern indicates that the model correctly identifies individual writer pairs but struggles with the full ranking across all 10 writers due to the small dataset size (50 images from 10 scribes) and substantial texture differences.

We generated foreground-attention overlays and retrieval montages directly from the trained model. Figures 2 and 3 show representative examples.

5 Conclusion

We introduced HIDRA, a transformer-based retrieval model for writer identification on historical handwritten fragments. HIDRA leverages a foreground prior computed from DINOv2 tokens that highlights ink regions to suppress background noise without explicit segmentation, and encodes local handwriting cues with dual-granularity attention-gated aggregation. Without any post-processing, HIDRA achieves 68.40% writer mAP on HisFrag20, surpassing the strongest

Table 7. Cross-domain dataset writer retrieval. HIDRA is trained only on HisFrag20 and evaluated directly on synthetic fragments from the target datasets. SAGHOG results from [27].

Dataset	Method	Writer mAP	Writer R@1
ScriptNet-HistoricalWI-2017 [14]	SAGHOG + NetRVLAD512 [27]	–	–
	HIDRA (ours)	**80.41**	**99.21**
GRK-Papyri [22]	SAGHOG + NetRVLAD512 [27]	–	58.0
	HIDRA (ours)	**44.03**	**98.85**

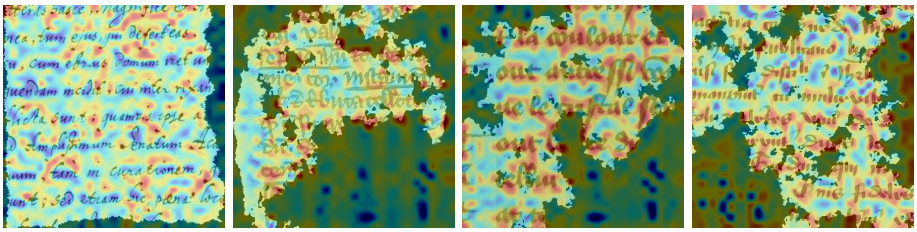

Fig. 2. DINOv2-derived foreground prior overlays on HisFrag20 fragments. The heatmaps concentrate on stroke-bearing regions and suppress large parts of the background texture/noise.

prior model-only baseline (SAGHOG, 57.2%) by +11.2 points. With transductive PCA whitening, mAP increases to 79.50%. Cross-domain dataset evaluation confirms strong generalization to related historical domains (80.41% mAP on ScriptNet-HistoricalWI-2017, 44.03% on GRK-Papyri), while highlighting that severe structural domain shifts (e.g., Greek papyri) remain an open challenge. Future work will explore improved hierarchical sampling strategies to unlock the full potential of the HCL objective, and domain-adaptive inference techniques for extreme cross-domain scenarios.

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). IEEE (2016). https://doi.org/10.1109/CVPR.2016.572
2. Buslaev, A., Parinov, A., Khvedchenya, E., Iglovikov, V.I., Kalinin, A.A.: Albumentations: Fast and flexible image augmentations (2018). https://doi.org/10.48550/arXiv.1809.06839
3. Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers (2021). https://doi.org/10.48550/arXiv.2104.14294
4. Chammas, M., Makhoul, A., Demerjian, J.: Writer identification for historical handwritten documents using a single feature extraction method. In: 2020 19th

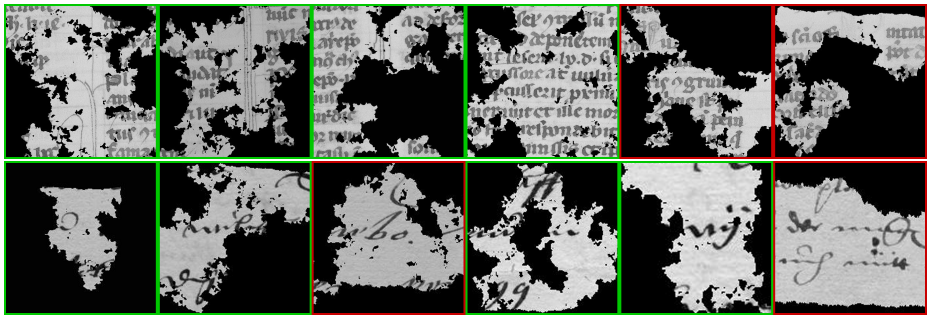

Fig. 3. Top-5 writer retrieval montages on HisFrag20. Top row: query and ranked matches with mostly correct writer hits. Bottom row: a harder case with stronger visual ambiguity and ranking noise in the tail.

IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE (Dec 2020). https://doi.org/10.1109/ICMLA51294.2020.00010
5. Chammas, M., Makhoul, A., Demerjian, J.: An end-to-end deep learning system for writer identification in handwritten arabic manuscripts. Journal of King Saud University - Computer and Information Sciences (2024). https://doi.org/10.1007/s11042-023-17303-8
6. Chammas, M., Makhoul, A., Demerjian, J., Dannaoui, E.: A deep learning based system for writer identification in handwritten arabic historical manuscripts. Multimedia Tools and Applications **81**(21), 30769–30784 (Apr 2022). https://doi.org/10.1007/s11042-022-12673-x
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations (2020). https://doi.org/10.48550/arXiv.2002.05709
8. 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. IEEE (Aug 2015). https://doi.org/10.1109/ICDAR.2015.7333893
9. Christlein, V., Bernecker, D., Maier, A., Angelopoulou, E.: Writer identification using gmm supervectors and exemplar-svms. Pattern Recognition **63**, 258–267 (2017). https://doi.org/10.1016/j.patcog.2016.10.005
10. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8. IEEE (2007). https://doi.org/10.1109/ICCV.2007.4408891
11. 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. 4690–4699. IEEE (Jun 2019). https://doi.org/10.1109/CVPR.2019.00482
12. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout (2017). https://doi.org/10.48550/arXiv.1708.04552
13. 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 (2020). https://doi.org/10.48550/arXiv.2010.11929

14. Fiel, S., Kleber, F., Diem, M., Christlein, V., Louloudis, G., Stamatopoulos, N., Gatos, B.: ICDAR2017 competition on historical document writer identification (Historical-WI). In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE (Nov 2017). https://doi.org/10.1109/ICDAR.2017.225
15. He, S., Schomaker, L.: Fragnet: Writer identification using deep fragment networks. IEEE Transactions on Information Forensics and Security **15**, 3013–3022 (2020). https://doi.org/10.1109/TIFS.2020.2979203
16. Jégou, H., Chum, O.: Negative evidences and co-occurrences in image retrieval: The benefit of PCA and whitening. In: Computer Vision – ECCV 2012. pp. 774–787. Springer Berlin Heidelberg (2012). https://doi.org/10.1007/978-3-642-33709-3_55
17. Jegou, H., Douze, M., Schmid, C., Perez, P.: Aggregating local descriptors into a compact image representation. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3304–3311. IEEE (Jun 2010). https://doi.org/10.1109/CVPR.2010.5540039
18. Keglevic, M., Fiel, S., Sablatnig, R.: Learning features for writer retrieval and identification using triplet CNNs. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 211–216. IEEE (Aug 2018). https://doi.org/10.1109/ICFHR-2018.2018.00045
19. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning (2020). https://doi.org/10.48550/arXiv.2004.11362
20. 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). https://doi.org/10.1109/TIFS.2020.2991880
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019). https://doi.org/10.48550/arXiv.1711.05101
22. Mohammed, H., Marthot-Santaniello, I., Märgner, V.: Grk-papyri: A dataset of greek handwriting on papyri for the task of writer identification. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 726–731. IEEE (2019). https://doi.org/10.1109/ICDAR.2019.00121
23. Ngo, T.T., Nguyen, H.T., Nakagawa, M.: A-VLAD: An End-to-End Attention-Based Neural Network for Writer Identification in Historical Documents, pp. 396–409. Springer International Publishing (2021). https://doi.org/10.1007/978-3-030-86331-9_26
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023). https://doi.org/10.48550/arXiv.2304.07193
25. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (1979). https://doi.org/10.1109/TSMC.1979.4310076
26. Peer, M., Kleber, F., Sablatnig, R.: Towards Writer Retrieval for Historical Datasets, pp. 411–427. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-41676-7_24
27. Peer, M., Kleber, F., Sablatnig, R.: SAGHOG: Self-supervised Autoencoder for Generating HOG Features for Writer Retrieval, pp. 121–138. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-70536-6_8

28. 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 '23). pp. 31–36. ACM (Aug 2023). https://doi.org/10.1145/3604951.3605515
29. 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 (2018). https://doi.org/10.1109/TPAMI.2018.2846566
30. 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 (Sep 2020). https://doi.org/10.1109/ICFHR2020.2020.00048
31. Seuret, M., Nicolaou, A., Stutzmann, D., Maier, A., Christlein, V.: Icfhr 2020 competition on image retrieval for historical handwritten fragments (hisfrag20) dataset (2020). https://doi.org/10.5281/ZENODO.3893807
32. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates (2017). https://doi.org/10.48550/arXiv.1708.07120
33. Su, B., Lu, S., Tan, C.L.: Binarization of historical document images using the local maximum and minimum. In: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems. pp. 159–166. ACM (2010). https://doi.org/10.1145/1815330.1815351
34. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3652–3661. IEEE (Jul 2017). https://doi.org/10.1109/CVPR.2017.389