Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Raw-input writer retrieval for historical documents by privileged-information distillation

Hussein Adnan Mohammed

hussein.adnan.mohammed@uni-hamburg.de

Universität Hamburg

Research Article

Keywords: writer retrieval, writer identification, historical documents, knowledge distillation, self-supervised learning, learning using privileged information

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11465298/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Raw-input writer retrieval for historical documents by privileged-information distillation

Hussein Mohammed[1*]

[1*]Centre for the Study of Manuscript Cultures (CSMC), The Visual Manuscript Analysis (VMA) lab, Universität Hamburg, Hamburg, Germany.

Corresponding author(s). E-mail(s): hussein.adnan.mohammed@uni-hamburg.de;

Abstract

State-of-the-art page-level writer retrieval for historical documents binarises every page before feature extraction. Binarisation removes the texture and colour of the writing support, stains and uneven lighting, but it also removes information that can carry the handwriting style, such as stroke intensity and ink gradients, and on much historical material it is unreliable. This work removes the binarised-input requirement of the feature-extraction network while improving retrieval accuracy. A page-level supervised-contrastive loss with spatially disjoint positives first improves the leading binarised-input model on HisIR19. This model then serves as the *teacher* in a privileged-information distillation: a larger student, initialised from a modern self-supervised backbone, receives the raw image and learns the teacher's tokens together with their patch-similarity structure, while keeping the teacher's dimensionality, so the retrieval pipeline stays unchanged. The network needs binarised pages only during training. On HisIR19, the raw-input student exceeds its own teacher and the checkpoint provided by the authors of the state-of-the-art system. Measured against that checkpoint under one shared evaluation pipeline, with exemplar-SVM re-encoding applied to both, it improves mean average precision. The mAP of our approach is the best reported on the HisIR19, Historical-WI and BullingerDB datasets. Without retraining, it transfers to modern handwriting (CVL) and, descriptively, to papyrus photographs (GRK-Papyri). Code, configurations and per-query results are released.

Keywords: writer retrieval, writer identification, historical documents, knowledge distillation, self-supervised learning, learning using privileged information

1 Introduction

Writer retrieval is the task of taking one handwritten page as a query and ranking all other pages of a corpus according to the likelihood that they were written by the same writer. On the page-level benchmarks for historical documents, the leading systems extract their features from binarised page images. The current state of the art [47] trains self-supervised vision transformers (ViTs) on binarised pages and aggregates foreground patch descriptors with a vector of locally aggregated descriptors (VLAD) [19]. We call this system the *reference system* and the checkpoint its authors provided to us the *reference checkpoint* (Sec. 3); our measured comparisons use them as the reference, and the system's training recipe is the starting point of our first stage (Fig. 1). Other recent systems also extract their features from binarised input [39, 40]. In all these pipelines,

every stage from pretraining to evaluation assumes a reliable separation of ink from background.

Binarisation gives these systems a clean view, with the texture and colour of the writing support, stains and lighting removed, but it also removes information. This work addresses two consequences. First, even where binarisation works well, it discards components that can carry the handwriting style: stroke intensity, ink-flow and pressure gradients, and fine edge shapes are absent from the binarised view. Whether this discarded signal can be exploited is an empirical question, and the controlled comparisons of Sec. 5 bear on it at the level of the full system, without isolating the contribution of the input alone (Sec. 6). Second, in many cases (ink on papyrus, faded or low-contrast inks, bleed-through, damaged pages) classical binarisation severely damages the script itself (Fig. 2) [34]. Learned binarisation reduces this damage but does not remove it, and its reported downstream gains require binarisers trained or adapted to the target material [1, 8]. The preprocessing then becomes one more corpus-dependent component that must be refitted for each new collection, while the extracted features still describe a degraded version of the script. A model that does not require a binarised input is therefore more generally applicable, and its value can already be measured on the standard benchmarks, where it must compete with the binarised pipeline on the material for which that pipeline was designed. Our network needs binarised input only during training, which can use an in-domain binarisable subset of the target dataset.

Our approach inverts the usual compression setting of knowledge distillation [17] (Fig. 1). A *larger* student network whose input is *richer* (the raw page, from which the teacher's view is computed) is distilled from a *smaller* teacher whose advantage is privileged information at training time [28, 55]: a binarised view that is *reliable* on the training material.

The main contributions of this paper are:

1. a page-level supervised-contrastive auxiliary loss [23] with *spatially disjoint* positives which, added to the training recipe of the reference system, improves this binarised-input model in mean average precision (mAP) on HisIR19, an improvement that transfers to the distilled student (Secs. 4.2, 5.2 and 5.3);
2. a privileged-information distillation setting in which a student larger than its teacher reads the raw image, with no binarisation, and learns through a *relational* term [27, 36, 54] that transfers the teacher's patch-similarity structure; the student exceeds both the reference checkpoint and its own teacher (Sec. 5.3);
3. new state-of-the-art mAP on HisIR19 and the Historical-WI test set in the full configuration of our evaluation pipeline (Sec. 4.4), trained without writer labels and with nothing fitted on the test collections, with a like-for-like margin over the reference checkpoint on HisIR19 under one shared evaluation pipeline; on BullingerDB, trained on its training split with the same two-stage method, the student is level with the strongest published baseline on that corpus in the base configuration and above it in the full configuration (81.9 versus 78.3 mAP, as reported; Sec. 5.4);
4. a zero-shot generalisation study in which the student, trained only on Historical-WI, is at or above the reference checkpoint on modern handwriting (CVL, where the student's 98.6 mAP in the base configuration equals the published result of Raven et al. [47], which uses a codebook fitted on CVL) and, descriptively (n=50), on papyrus photographs (GRK-Papyri), and is measured on BullingerDB against systems trained on that corpus (Sec. 5.5).

The paper is organised as follows. Sec. 2 reviews related work. Sec. 3 establishes the measured baseline and examines the evaluation itself. Sec. 4 presents the two training stages and the shared evaluation pipeline. Sec. 5 reports the controlled comparisons, the comparison with the state of the art and the zero-shot transfer study, and Sec. 6 discusses the findings and concludes.

2 Related work

This paper draws on writer retrieval for historical documents and on knowledge distillation under privileged information.

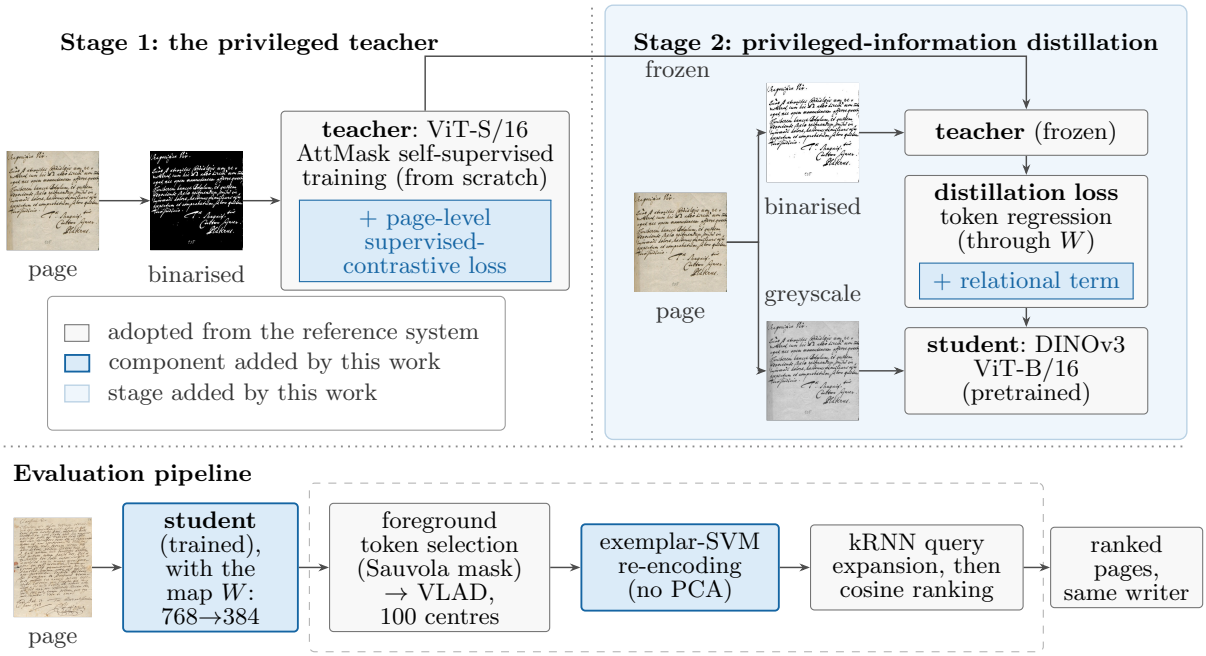

Fig. 1 System overview. Stage 1 trains the teacher, a ViT-S/16, on the binarised Historical-WI training pages provided with the dataset, with the self-supervised recipe of the reference system of Raven et al. [47] plus the page-level supervised-contrastive loss (Sec. 4.2). Stage 2 distils the frozen teacher into a larger ViT-B/16 student initialised from DINOv3, which receives the raw page (as greyscale during distillation); its tokens are regressed onto the teacher's through the trained map W, and the relational term transfers the teacher's patch-similarity structure (Sec. 4.3). The bottom band shows the full configuration of the evaluation pipeline (Sec. 4.4); the base configuration uses PCA whitening in place of exemplar-SVM re-encoding. At inference the student receives the raw colour page. Highlighted components and the highlighted stage are added by this work; outside that stage, the other components are adopted from the reference system, except the VLAD normalisation and codebook fit. Thumbnails: a Historical-WI training page (raw, binarised, greyscale) and a HisIR19 test page; both binarised thumbnails show the Sauvola rendering, for illustration.

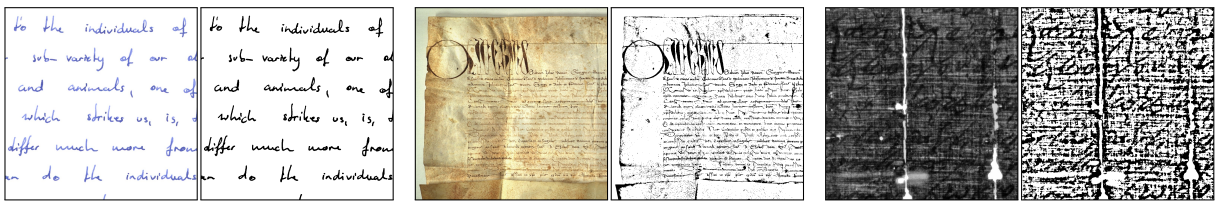

Fig. 2 Raw page crops, each followed by its binarisation with Sauvola (window 51), the preprocessing of the reference system. Left to right: modern handwriting (CVL), where the script survives binarisation; a historical page (HisIR19), where it survives with visible damage; a Greek papyrus (GRK-Papyri), where it is largely destroyed. Each crop is a 1,024-pixel square taken at the page's original resolution with no rescaling, so the visible extent differs between corpora.

2.1 Writer retrieval and identification

Classical pipelines aggregate local descriptors with codebook methods such as VLAD [19]. Deep variants learn per-patch features with surrogate classes [6] or metric learning [22] and keep the codebook aggregation stage, while other CNN approaches combine fragment-level predictions directly [16]. Exemplar support vector machines (SVMs) [31] were introduced to this task by Christlein et al., as a document-specific similarity [5] and as a feature encoding [6]. Self-supervised vision transformers with morphological augmentations [38], vision-transformer backbones [25] and learned feature aggregators [32] have been evaluated for writer retrieval. One recent system learns the aggregation itself, combining a residual network, a NetRVLAD encoding layer and graph-based re-ranking [39]; another pretrains vision

transformers with self-supervised objectives tailored to handwriting [40]. The current state of the art on the recent historical benchmarks [47] trains a vision transformer on binarised handwriting windows with self-supervised objectives of the DINO/iBOT/AttMask family [4, 21, 59], aggregates foreground patch tokens with VLAD, applies principal component analysis (PCA) whitening, and re-ranks with k-reciprocal nearest-neighbour (kRNN) query expansion [20, 46]. Beyond these benchmarks, recent work makes the aggregation interpretable through character-level codebooks [48] and scales writer retrieval to archive-size supervised settings with learned end-to-end aggregation on colour input [49]; the same approach leads historical fragment retrieval on HisFrag20 [52]. Retrieval quality is also known to depend strongly on the amount of handwriting available per sample [41]. Further systems operate on unbinarised input [16, 33, 58], but on the page-level historical benchmarks the leading systems binarise. Writer identification [34, 35, 37] and fragment retrieval [44] have been studied directly on papyri, a material that resists binarisation. Binarisation itself is an active research area with dedicated competitions and learned methods [15, 45]; our argument is not that binarisation cannot be improved, but that its role should be confined to where it is dependable: in our method, binarisation is a network input only for the training pages.

2.2 Distillation and privileged information

Knowledge distillation classically compresses a large teacher into a small student on the same input [17], and intermediate representations can be matched through learned regressors that are discarded after training [50] (our corresponding map is kept at inference). A separate line, learning under privileged information, gives the teacher training-time access to information the student will not have at test time [55]. Generalised distillation unifies the two views [28], and our training setting belongs to this family. Several asymmetric variants are related to it: distillation across modalities [13, 18], self-training with an equal-or-larger noised student [57], and born-again networks, where students exceed their teachers at equal capacity [11].

Our setting differs from these works in one respect. The privileged view $b(x)$, the binarised page, is a deterministic function of the student's input x, so it could be computed at test time and carries no writer information beyond x; its privilege is its reliability. On the binarisable training corpus $b(x)$ is a trustworthy, clean view, from which the teacher learns a target that is invariant to everything the binarisation removes. On papyri (Sec. 5.5) it is not reliable, and confining it to training time keeps it out of the network input at inference. The closest precedent we know of is in another modality: in speech recognition, Watanabe et al. [56] train a teacher on enhanced multi-microphone features and a student on the original noisy input; there the enhanced view uses microphones the student does not receive, so the student's input is not richer than the teacher's. To the best of our knowledge, the composition used here (a larger student, a richer student input from which the teacher's view is computed, and a teacher whose privilege is a view that is reliable only at training time) has not been explored before, in document analysis or elsewhere. For the transfer signal we build on relational distillation, which transfers the structure among the samples of a batch [36, 54] or among the positions of one input [14, 27] instead of individual activations. Our relational term, Eq. (4), has the form of the Gram anchoring of DINOv3 [53], which regularises one network against an earlier checkpoint of itself; here it is applied across two networks and two input views.

3 Establishing the baseline

The baseline of this paper is the system of Raven et al. [47]: it holds the state of the art on HisIR19 and the Historical-WI test set, its code is released, and its authors answered our protocol questions and provided the training checkpoint behind their experiments, which makes exact, like-for-like comparisons possible.

All measured comparisons in this paper are produced under one evaluation pipeline, applied identically to every checkpoint, so that no measured result depends on implementation differences between codebases (published numbers are quoted as reported and never entered into paired comparisons). The pipeline (Sec. 4.4) follows the protocol of Raven et al. [47], with protocol details

confirmed in correspondence with the authors (Appendix A). Pages are processed at their native resolution, and every window enters the aggregation, as in the released code of the reference system (Appendix A). As in the original protocol, the online network of each binarised-input model's self-supervised pair (online network and momentum encoder, Sec. 4.2) is evaluated. All mAP and Top-1 values (defined in Sec. 5.1) are given as percentages. Measured values are printed to two decimals in the controlled comparisons (Tables 1–3 and the text quoting them) and to one decimal in Tables 4–7 and their text, where measured values stand beside published ones, the convention of the compared literature. Paired differences (Δ) are given in percentage points to two decimals, with their 95% confidence intervals where space allows. The reference checkpoint is the like-for-like external comparator in Tables 1–6; it is not publicly distributed by the authors, but our per-query results for it are released, so every paired comparison can be re-tested without it.

We also retrained the published recipe from scratch on the Historical-WI training pages [10]; we call the result *our reproduction*. Training follows the authors' released code exactly (commit 1763137 of https://github.com/Traven16/SSL_ViT_WR), with the peak learning rate confirmed by the authors (Appendix A). Table 1 summarises the outcome on the HisIR19 [7] and Historical-WI test sets (Sec. 5.1).

Because neither the reference checkpoint nor our reproduction reaches the published mAP (Table 1), the evaluation itself is examined first, on the reference checkpoint alone (Table 1, upper block). Starting from the reference system's published evaluation, one element is changed per row, in an order fixed in advance, until the configuration used for every comparison in this paper is reached, which we call the *base configuration* (Sec. 4.4): all windows kept, as in the released code; our VLAD normalisation; our codebook fit. Adjacent rows score the same queries, and their differences are given in mAP points. Of these four rows, the base configuration gives the reference checkpoint its best result, 1.01 points above its published evaluation. The path between the endpoints is not monotone. Keeping all windows, as the released code does, first costs 2.51 points under the publication's VLAD normalisation (power and global ℓ_2 normalisation only), so the window rule as implemented here helps under that normalisation, while under the base configuration it costs 0.22 points (Appendix A). In the last two rows, our VLAD normalisation recovers 3.26 points over the released-code row, and our codebook fit adds a further 0.29, measured against the previous row computed with the randomised PCA solver of the base configuration, so that only the codebook fit changes. None of the four configurations reaches the published 95.0; the gap is smallest (1.46 points) in the base configuration and largest (4.98 points) under the released code's behaviour. The base configuration was fixed before the three rows above it were measured; these rows diagnose the evaluation and were not used to select the configuration.

Comparisons between models are decided by a paired per-query bootstrap on the same queries; each such comparison, here and throughout the paper, reports the mean per-query difference in average precision with its 95% confidence interval (CI) and two-sided p-value (Appendix B). In the base configuration, our reproduction is not significantly different from the reference checkpoint ($+0.01$ points, 95% CI $[-0.12, +0.14]$, $p = 0.86$; 1,142 queries favour one model and 1,141 the other). On the second benchmark, Historical-WI, the binarised-input rows use the dataset's provided binarised test images, the convention of the published numbers, and the difference is again not significant ($+0.22$ points, 95% CI $[-0.02, +0.46]$, $p = 0.067$; Table 1). As an indication of training noise, two training seeds of this recipe differ by 0.06 mAP points at the final checkpoint; this two-seed range was measured on the 1,200 validation pages, not on the test set. The cause of the gap to the published mAP is not identified; Appendix A documents the checks made. Even at its measured values in the base configuration (93.54 on HisIR19, 82.41 on Historical-WI), the reference system remains ahead of every other published system on both benchmarks (next-best mAP: 93.2 and 80.6, Table 4); the shortfall is only against its own published numbers. Our improvements are assessed against the measured 93.54 rather than the published 95.0, and our final results nevertheless exceed the published numbers (Sec. 5.4).

Table 1 Baseline measurements (% mAP / % Top-1), binarised input. Upper block: the reference checkpoint on HisIR19, from the reference system's published evaluation, with the window rule as implemented here (Appendix A), changing one element per row until the base configuration is reached. Rows above the base configuration fit PCA by exact singular value decomposition (SVD), the base configuration by randomised SVD (immaterial, Appendix A). Historical-WI rows use the dataset's provided binarised test and training pages. The published row is quoted as reported by Raven et al. [47] (their own evaluation).

	HisIR19		Historical-WI	
Model (ViT-S/16) / evaluation	mAP	Top-1	mAP	Top-1
Reference checkpoint, published evaluation [47]	92.53	96.70	—	—
+ all windows kept (as in the released code)	90.02	95.41	—	—
+ our VLAD normalisation	93.28	96.76	—	—
+ our codebook fit (base configuration)	93.54	97.13	82.41	90.83
Our reproduction (base configuration)	93.55	97.03	82.63	91.39
Published, as reported [47]	95.0	97.6	83.1	90.9

4 Method

Our method has two training stages and one shared evaluation pipeline (Fig. 1), and it involves two models that must not be confused with the two networks inside self-supervised training. In Stage 1 we train *our teacher*: a small vision transformer that sees only the binarised view of each page, trained with the self-supervised recipe of the reference system (which internally pairs an online network with a momentum encoder) plus a page-level supervised-contrastive loss that improves it. In Stage 2 we train *our student*: a larger vision transformer that sees the unbinarised page image and learns, by distillation, to produce the patch tokens that the frozen teacher produces from the binarised view; a small trained map keeps the student's output at the teacher's dimensionality, so the retrieval pipeline downstream is identical for every model. At inference the student network receives only the raw image, with no binarisation or other preprocessing of its input. As a network input, binarisation is therefore needed only for the training pages of both stages, not for the collection that is searched; in principle, a binarisable subset of the target material suffices for training. The evaluation pipeline keeps the components of the reference system, refines its VLAD normalisation and codebook fit, adds exemplar-SVM re-encoding, and is applied identically to all models.

4.1 Setting and notation

Let x denote a page image and $b(x)$ a binarised, white-ink view of it. The teacher f_T is a ViT-S/16 (384-dimensional tokens) and, as in the reference system, is trained on the binarised Historical-WI training pages provided with the dataset. Wherever this paper computes a binarisation itself, $b(x)$ is the Sauvola rendering [51] with window 51, k=0.2 and R=128, the preprocessing of Raven et al. [47]; the binarised Historical-WI rows use the provided binarised test pages. The student f_S is a ViT-B/16 (768-dimensional tokens), initialised from the publicly released DINOv3 ViT-B/16 checkpoint [53] (distilled without labels from the largest self-supervised DINOv3 model; its register tokens are not used in our token extraction), and trained on the raw view with targets computed from $b(x)$. During distillation the raw view is presented as greyscale (luminance replicated to three channels), so that the regression to targets computed from the coloess binarised view does not additionally have to bridge colour variation. At inference the student network receives the unmodified colour image and is evaluated in the same retrieval pipeline as every other model (Sec. 4.4). During distillation and evaluation both networks consume 224-pixel windows, so their 14×14 token grids align one-to-one (the teacher's own pretraining takes random-resized 224-pixel crops from 256-pixel windows, Sec. 4.2).

4.2 Stage 1: the privileged teacher

The teacher training starts from the unmodified recipe of Raven et al. [47]: AttMask [21] (masked-image modelling of the DINO/iBOT family with attention-guided masking) on the 1,182 binarised Historical-WI training pages provided with the dataset, windowed at 256 pixels with stride 32 (each window yields two random-resized 224-pixel global crops and eight smaller local crops), 20 epochs, peak learning rate 5×10^{-4}, batch 128. This recipe minimises $\mathcal{L}_{\text{AttMask}}$, the sum of masked-token and class-token objectives between the online network and its momentum encoder. Our modification is an auxiliary loss together with the page-grouped batch sampler it requires, both described next; everything else in the recipe stays as published.

Page-level supervised contrast with spatially disjoint positives

Page identity is available without annotation and can serve as a label, since the windows of one page share a writer. (Pages with more than one hand are not the common case in the datasets used here, and the label serves only as a training signal.) Training without writer annotation has precedent: clusters of local descriptors serve as surrogate classes in writer identification [6], and fragment identity serves as the class in papyrus fragment retrieval [44]. Our additions are page identity as a supervised-contrastive label inside the AttMask recipe and the spatial-disjointness constraint below. We add a supervised-contrastive (SupCon) term [23] over the online network's class tokens, with positives defined by page identity, leaving $\mathcal{L}_{\text{AttMask}}$ untouched:

$$\mathcal{L}_{\text{T}} = \mathcal{L}_{\text{AttMask}} + \lambda_{\text{sc}}\,\mathcal{L}_{\text{sc}}, \qquad \lambda_{\text{sc}} = 0.5, \quad (1)$$

where λ_{sc} weights the auxiliary term $\mathcal{L}_{\text{sc}}$, defined as

$$\mathcal{L}_{\text{sc}} = -\frac{1}{|B|}\sum_{i\in B}\frac{1}{|P(i)|} \times \sum_{p\in P(i)} \log \frac{\exp(z_i\cdot z_p/\tau)}{\sum_{a\in B,\,a\neq i}\exp(z_i\cdot z_a/\tau)}, \quad (2)$$

with temperature $\tau = 0.1$. Here B indexes the two global crops of every window in the batch, and z_i is the ℓ_2-normalised backbone class token (no projection head) of crop i, taken from the online network's forward pass that $\mathcal{L}_{\text{AttMask}}$ already computes, in which attention-guided masking is applied to each global crop with probability 0.5. $P(i) \subset B$ contains the crops from the same page as crop i, excluding i itself; it always contains the other global crop of the same window. Batches are page-grouped: 16 pages $\times$ 8 windows give the recipe's batch of 128 windows, contributing $|B| = 256$ crops; pages are visited in shuffled round-robin order, so that every page contributes equally often, and an epoch keeps the recipe's number of iterations. The two crops of one window form conventional augmented-view positives; the constraint below governs the *cross-window* positives that carry the page-level signal. At training stride 32, adjacent windows overlap by up to 87.5% of their area, so unconstrained same-page positives would reward shared pixels rather than a shared handwriting style. Windows within a page are therefore selected with pairwise zero overlap (intersection over union 0): candidates from the stride-32 grid are drawn in random order and accepted if disjoint from all previously accepted windows. Pages that cannot supply eight disjoint windows are topped up with the rejected candidates in draw order, so the batch width is always maintained.

We apply this loss at the teacher stage only, for the following reason. On the binarised view, page identity is a nearly pure writer label: binarisation strips paper tone, lighting and degradation, so disjoint windows of one page share little besides the writer's handwriting style. On raw input, windows of one page also share paper colour, texture and lighting, and we expect a page-level contrastive loss there to reward encoding page *appearance*, which a writer representation must discard. The distillation targets transfer the effect of the loss to the student (Sec. 5.3).

All teachers, including a sampler-only control (page-grouped batches, $\lambda_{\text{sc}} = 0$) that separates the effect of the grouping from that of the loss (Sec. 5.2), are trained for the published 20 epochs, and the final checkpoint of the online network is used (no checkpoint selection).

4.3 Stage 2: privileged-information distillation

The student learns to produce, from the unbinarised view, the patch tokens that the frozen teacher produces from the binarised view. For a 224-pixel training window u (Sec. 4.1), let $s_j = f_S(u)_j \in \mathbb{R}^{768}$ denote the student's patch token and $t_j = f_T(b(u))_j \in \mathbb{R}^{384}$ the teacher's patch token at grid position j, where j runs over the $N = 196$ positions of the window; s_{cls} and t_{cls} are the two networks' class tokens. Every position enters the loss with the same weight; there is no foreground weighting. The base objective regresses each student token onto the teacher token at the same position:

$$\mathcal{L}_{\mathrm{S}}^{0} = \tfrac{1}{N}\textstyle\sum_{j=1}^{N}\big(1-\cos(W(s_j),\,t_j)\big) + \lambda_{\mathrm{cls}}\big(1-\cos(W(s_{\mathrm{cls}}),\,t_{\mathrm{cls}})\big), \quad (3)$$

where $\cos(\cdot,\cdot)$ is the cosine similarity and $\lambda_{\mathrm{cls}} = 0.5$ weights the class-token term. The map $W(s) = As + c$, with $A \in \mathbb{R}^{384\times768}$ and $c \in \mathbb{R}^{384}$, is a single trained affine layer that bridges the two token widths. It is trained jointly with the student, as are the hint regressors of FitNets [50]; unlike those regressors, W is kept at inference: the student's tokens are passed through it, so the student keeps the teacher's dimensionality and the retrieval pipeline is dimensionally identical for every model compared.

The second term transfers the teacher's *patch-similarity structure*. Write $\hat{s}_j$ for the student's ℓ_2-normalised patch token (its own 768-dimensional token, without W) and $\hat{t}_j$ for the teacher's. For each window, each network's Gram matrix collects the pairwise similarities among its own tokens: $G^S_{ij} = \hat{s}_i \cdot \hat{s}_j$ and $G^T_{ij} = \hat{t}_i \cdot \hat{t}_j$, both $N \times N$. The relational loss matches the two matrices by mean squared error:

$$\mathcal{L}_{\mathrm{rel}} = \tfrac{1}{N^2}\big\|G^S - G^T\big\|_F^2, \qquad \mathcal{L}_{\mathrm{S}} = \mathcal{L}_{\mathrm{S}}^{0} + \beta\,\mathcal{L}_{\mathrm{rel}}, \quad (4)$$

where $\|\cdot\|_F$ is the Frobenius norm and $\beta = 5$ weights the term; Eqs. (3) and (4) are averaged over the windows of a batch. The term has the form of the pairwise similarity loss of Liu et al. [27], applied here to the patch tokens of one window. Two properties suit it to the task. First, VLAD assigns tokens to codebook centres by similarity and aggregates deviations, so the similarity structure among the student's tokens shapes, through W, what retrieval consumes. Second, the Gram target constrains only the pairwise similarities among the student's own tokens before W, fixing them up to an orthogonal transformation that may differ from window to window, whereas Eq. (3) requires the single map W to reproduce the teacher's coordinates in every window. The relational term thus transfers which tokens resemble which, while constraining the student's coordinates less than Eq. (3) does.

The student is trained on the same 1,182 Historical-WI training pages (colour scans) at their native resolution, cut into 224-pixel tiles at 50% overlap. A tile is discarded when the bounding box of its dark pixels covers at most 40% of the tile or when fewer than 4% of the pixels in that box are Canny edge pixels, leaving about 130,000 training tiles per epoch (batch 96). The teacher's input is the Sauvola view $b(u)$ of Sec. 4.1, computed on the fly from each tile; during distillation the teacher therefore receives a Sauvola rendering rather than the provided binarisation on which it was trained. No stochastic augmentation is applied, since the loss is a deterministic regression. All students are trained for the same fixed budget of five epochs (linear warm-up from zero over the first epoch, then per-step cosine annealing to zero [29]; AdamW [30] with learning rate 10^{-4}, layer-wise learning-rate decay [3] of 0.75 and weight decay 0.04), and the final checkpoint is evaluated (no checkpoint selection). The budget is fixed in advance and identical for every variant, so every comparison is matched in training cost. The two baseline students differ only in their teacher, our reproduction (Sec. 3) or our improved teacher (Sec. 4.2); this pair measures how teacher gains transfer to the student. Appendix C (Table C1) lists the hyper-parameters.

4.4 Evaluation pipeline

Windows of 224 pixels are taken at stride 56 over the page at its native resolution, and every window is kept, as in the released code of the reference system (Appendix A). Within each window, a Sauvola mask (window 51, k=0.2) selects the patch tokens that are aggregated: a token is kept when at least 4% of the pixels of its 16×16 patch are ink on the mask, the analogue of the foreground threshold t_{fg} of Raven et al. [47]; we call

this selection the *foreground gate*. It is inherited from the reference system and computed identically for every model; the mask is never a network input. The kept tokens of all windows form one set per page.

This token set is aggregated with VLAD [19], using 100 centres. The codebook is fitted on the fitting corpus (Sec. 5.1) by mini-batch k-means on a random sample of at most 256 foreground tokens per page, taken from non-overlapping windows; the three rows above the base configuration in Table 1 instead fit full-batch k-means, with one initialisation, on all foreground tokens of these windows. Tokens are ℓ_2-normalised and assigned to their nearest centre. Residuals are ℓ_2-normalised before summation [9], and the descriptor receives signed square-root, per-centre [2] and global ℓ_2 normalisation. The reference publication uses power and global ℓ_2 normalisation only (Sec. 3). The descriptor is then PCA-whitened to 384 dimensions (randomised SVD, fitted on the same corpus), followed by kRNN query expansion with neighbourhood size k=2 [20, 46] (each ℓ_2-normalised descriptor is replaced by the ℓ_2-normalised sum of itself and its two mutual nearest neighbours) and cosine ranking. On HisIR19, queries are the test pages whose writer contributes at least one further page; singleton pages are gallery-only, and each query is ranked against the remaining pages. These settings define the base configuration (Sec. 3). Every controlled comparison of training recipes uses it, so that measured differences are attributable to training changes alone.

The *full configuration*, the second contributed part of our pipeline after the VLAD normalisation and codebook fit, differs from the base configuration in one stage: exemplar-SVM re-encoding [31], the feature encoding of Christlein et al. [6], replaces PCA whitening. Every page descriptor, queries and gallery alike, is ℓ_2-normalised and replaced by the ℓ_2-normalised weight vector of a linear exemplar SVM (squared hinge loss, ℓ_2 regularisation, C=1, class-balanced weights, bias discarded) trained with that page as the single positive against the ℓ_2-normalised descriptors of the fitting corpus as negatives (for HisIR19, the 1,200 validation pages). The descriptors enter without PCA whitening, because each exemplar SVM already learns a page-specific linear discriminant against the shared negatives. Ranking then proceeds exactly as in the base configuration (the same kRNN query expansion, then cosine ranking).

Comparisons of training recipes and comparisons of evaluation stages are never mixed, and models are always compared within one configuration; the effect of the configuration itself is measured within one model. Same-configuration comparisons are decided by a *paired per-query bootstrap*: both systems score the same queries, so every query contributes one difference in average precision (AP), and the mean of these differences is the mAP gap between the systems. Drawing the queries with replacement 20,000 times (fixed seed) yields a percentile confidence interval and a two-sided p-value for that gap. The bootstrap conditions on the trained models and on one seeded fit of the codebook and PCA: every arm is a single training run, so the intervals reflect query sampling, not run-to-run training variance. Statistical inference is on mAP, and Top-1 is reported descriptively. Code, configurations and the per-query results of every evaluated model are released to facilitate later comparisons by other researchers[https://github.com/vma-lab/WR_by_PID].

5 Experimental results

The teacher and the student are evaluated in the base configuration, the full configuration is compared with the state of the art, and three further datasets test generalisation.

5.1 Experimental setup

Five datasets are used. *HisIR19* [7] (test: 20,000 pages; 10,068 writers) and the *Historical-WI test set* [10] (3,600 pages, 720 writers) are the two standard historical benchmarks used by the reference system. Unless stated otherwise, training happens on the Historical-WI training pages, and the HisIR19 validation set is the fitting corpus of the evaluation pipeline (codebook, PCA and exemplar-SVM negatives). Historical-WI rows fit these on the Historical-WI training pages in each model's input domain (the provided binarised pages for the binarised-input models, the colour pages for the student), and the BullingerDB rows of Table 4 train and fit on that corpus's own split (Sec. 5.4). *GRK-Papyri* [34] (50 photographs, 10

writers, leave-one-out protocol: each photograph in turn queries the remaining 49) represents material where binarisation fails. *CVL* [24] (cropped handwriting fields; 283 test writers with 5 pages each, 27 training writers with 7) represents modern handwriting. *BullingerDB* [43] (16th-century correspondence; writer-retrieval split with 4,396 training and 16,502 test pages by disjoint writers) is a recent large historical benchmark with temporal metadata.

Retrieval is evaluated by mean average precision (mAP), the mean over queries of the average precision of the ranking of all other pages, and by Top-1 accuracy, the share of queries whose highest-ranked page is by the same writer; BullingerDB adds a temporal normalised discounted cumulative gain (nDCG), described in Sec. 5.4. Unless stated otherwise, every result uses the base configuration.

5.2 Teacher

The page-level supervised-contrastive modification improves the teacher's mAP on HisIR19 by +0.20 points over the reference checkpoint (95% CI $[+0.06, +0.34]$, $p = 0.006$) and by +0.18 over our reproduction (95% CI $[+0.07, +0.30]$, $p = 0.002$; Table 2). The sampler-only control of Sec. 4.2, measured on HisIR19 in the base configuration, separates the two ingredients of the modification: the batch composition alone degrades the teacher (−0.29 points against our reproduction, 95% CI $[-0.40, -0.17]$, $p < 10^{-4}$), and adding the loss more than recovers it (+0.47 points over the control, 95% CI $[+0.34, +0.60]$, $p < 10^{-4}$). The improvement is attributable to the loss, not to the grouping it requires.

The HisIR19 margin is modest in absolute terms (the benchmark sits above 93%): of the queries whose AP changes, 1,245 improve and 1,036 degrade, and the paired 95% CI excludes zero. On Historical-WI the margin does not reach significance. The improved teacher is +0.22 points above the reference checkpoint (95% CI $[-0.02, +0.46]$, $p = 0.074$; 3,600 queries), the same margin as our reproduction, so on this benchmark no difference between the three binarised-input models is significant.

5.3 Student

The two direct baselines measure how teacher quality transfers through distillation. The student distilled from the improved teacher is +0.47 mAP points above the student distilled from our reproduction (paired per-query bootstrap: 95% CI $[+0.33, +0.60]$, $p < 10^{-4}$), so the teacher-side gain of +0.18 points grows to +0.47 in the student. The student distilled from our reproduction is not significantly different from the reference checkpoint (+0.08, 95% CI $[-0.08, +0.24]$, $p = 0.32$) or from its own teacher (+0.07, 95% CI $[-0.09, +0.23]$, $p = 0.39$): direct distillation moves the binarised-input teacher onto raw input with no measurable loss. The student distilled from the improved teacher, before the relational term is added, is above the reference checkpoint on raw input (+0.55, 95% CI $[+0.38, +0.72]$, $p < 10^{-4}$). The final student adds the relational term, the one remaining training difference (Sec. 4.3), which contributes +0.46 points over the student distilled from the improved teacher (95% CI $[+0.33, +0.59]$, $p < 10^{-4}$).

On raw network input, the final student exceeds the binarised-input reference checkpoint by +1.01 mAP points (95% CI $[+0.85, +1.17]$, $p < 10^{-4}$; 94.55 against 93.54; 1,469 queries improve, 816 degrade) in the base configuration, which also gives the reference checkpoint its best result among the upper-block rows of Table 1; the margin was not measured under the publication's own VLAD normalisation. The final student also exceeds its own improved teacher across the input-domain boundary (94.55 raw against 93.73 binarised; paired +0.81, 95% CI $[+0.66, +0.97]$, $p < 10^{-4}$), a comparison we return to in Sec. 6. The undistilled control shows that the pretrained initialisation alone is not sufficient: the same backbone, evaluated identically but without distillation, reaches 87.26 (Table 3); compared at the unprojected representation, the only one the undistilled backbone has, the final student is +6.95 points above it (95% CI $[+6.61, +7.30]$, $p < 10^{-4}$).

On Historical-WI, where the student receives the raw colour scans, the final student is above the reference checkpoint at the significance boundary (+0.26, 95% CI $[+0.00, +0.52]$, $p = 0.049$; 82.66 against 82.41; 653 queries improve, 672 degrade) and not significantly different from its own teacher (+0.04, 95% CI $[-0.19, +0.27]$, $p = 0.76$). With

Table 2 Teacher results (% mAP / % Top-1), base configuration, binarised input. Δ in points on HisIR19, paired per-query bootstrap against the reference checkpoint (confidence intervals in the text). Bold marks the improved teacher's mAP on HisIR19, the one margin whose interval excludes zero; on Historical-WI no difference between the three models is significant.

	HisIR19		Historical-WI		
Model (ViT-S/16)	mAP	Top-1	mAP	Top-1	paired Δ (HisIR19)
Reference checkpoint [47]	93.54	97.13	82.41	90.83	—
Our reproduction (published recipe)	93.55	97.03	82.63	91.39	+0.01 ($p = 0.86$)
Improved teacher (+ page-level SupCon)	**93.73**	97.01	82.63	90.89	+0.20 ($p = 0.006$)

Table 3 Distillation results (% mAP / % Top-1), base configuration. Teacher rows use binarised input; student rows use the raw image as network input. The undistilled row measures the student's pretrained initialisation without distillation. Distilled student rows use the projected tokens (384-dimensional, through W, Sec. 4.3); the undistilled and ablation rows use 768-dimensional tokens. Historical-WI is measured for the binarised-input rows and the final student only. Bold marks the final student's mAP.

		HisIR19		Historical-WI	
Row	Input	mAP	Top-1	mAP	Top-1
Reference checkpoint [47]	binarised	93.54	97.13	82.41	90.83
Improved teacher (Sec. 4.2)	binarised	93.73	97.01	82.63	90.89
DINOv3 ViT-B/16, no distillation (control)	raw	87.26	93.36	—	—
Student distilled from our reproduction	raw	93.62	96.99	—	—
Student distilled from the improved teacher	raw	94.08	97.30	—	—
+ relational term (final student)	raw	**94.55**	97.34	**82.66**	91.39
Final student, unprojected tokens (ablation)	raw	94.21	97.43	—	—

3,600 queries the confidence intervals are about 1.5 to 1.8 times as wide as on HisIR19.

A representation ablation evaluates the same final checkpoint with its unprojected 768-dimensional tokens instead of the tokens projected through W. It yields 94.21 / 97.43, 0.34 mAP points below the projected tokens (95% CI $[-0.45, -0.23]$, $p < 10^{-4}$; the same holds in the full configuration, -0.36, $p < 10^{-4}$), although its tokens, and its VLAD descriptor before PCA whitening, are twice as wide. All student rows therefore use the projected tokens.

5.4 Full configuration and comparison with the state of the art

Table 4 places the reference checkpoint and the final student, measured in both configurations of our pipeline (Sec. 4.4), next to the published state of the art on both benchmarks, and adds BullingerDB as a third corpus, for which the two-stage method is trained on that corpus's own split (below). On HisIR19 the full configuration adds 0.69 points over the base configuration for the reference checkpoint and 0.75 for the final student.

In the base configuration (VLAD with kRNN query expansion, as in the protocol of Raven et al. [47]), the student's 94.5 on HisIR19 lies between the published 94.4 (no re-ranking) and 95.0 (with re-ranking) of that system, while our like-for-like comparison with its reference checkpoint gives +1.01 points (Sec. 5.3). In the full configuration the student reaches 95.3 / 98.0 on HisIR19 and 83.4 / 91.9 on Historical-WI. On both benchmarks this mAP is above every published result, including the published mAP of Raven et al., which neither the reference checkpoint nor our reproduction reaches in the base configuration (Sec. 3). Measured like-for-like in the full configuration, the student's margin over the reference checkpoint on HisIR19 is +1.07 points (95% CI $[+0.93, +1.20]$, $p < 10^{-4}$); on Historical-WI the

Table 4 Comparison with the state of the art (% mAP / % Top-1). Published rows are quoted as reported, each under its own evaluation; the reference checkpoint and our student are measured in our evaluation pipeline (Sec. 4.4). All listed systems except our student binarise the page before feature extraction; our student (a ViT-B/16 from DINOv3, tokens mapped to 384 dimensions) reads the raw image, and Raven et al. use a ViT-S/16. kRNN: k-reciprocal nearest-neighbour query expansion; ESVM: exemplar-SVM re-encoding; SGR: graph-based re-ranking; nDCG: temporal normalised discounted cumulative gain; mVLAD, bVLAD, NetRVLAD: VLAD variants of the cited works. Bold marks the best mAP per benchmark. The BullingerDB block gives that corpus's three metrics under its own cosine ranking protocol; both of our rows there are trained on its training split with the two-stage method (Sec. 5.4), and the reference checkpoint was not evaluated on this corpus.

		HisIR19		Historical-WI	
Method	Post-processing	mAP	Top-1	mAP	Top-1
Christlein et al. [6], mVLAD	ESVM	—	—	76.2	88.7
Lai et al. [26], bVLAD	—	92.5	97.4	77.1	90.1
Peer et al. [39], NetRVLAD	—	91.6	96.1	73.4	88.5
Peer et al. [39], NetRVLAD	SGR	93.2	96.7	80.6	91.1
Raven et al. [47] (published), VLAD	—	94.4	97.8	82.6	91.9
Raven et al. [47] (published), VLAD	kRNN	95.0	97.6	83.1	90.9
Reference checkpoint [47], VLAD	kRNN	93.5	97.1	82.4	90.8
Reference checkpoint [47], VLAD	ESVM + kRNN	94.2	97.8	83.3	91.9
Ours (raw-input student), VLAD	kRNN	94.5	97.3	82.7	91.4
Ours (raw-input student), VLAD	ESVM + kRNN	**95.3**	98.0	**83.4**	91.9

		BullingerDB		
		mAP	nDCG	Top-1
Peer et al. [43], NetRVLAD	—	65.1	84.8	93.4
Peer et al. [43], AttMask+VLAD	—	78.3	90.3	95.6
Ours (raw-input student), VLAD	—	78.6	90.3	95.7
Ours (raw-input student), VLAD	ESVM	**81.9**	92.1	97.1

same margin is not significant (+0.13, 95% CI $[-0.10, +0.35]$, $p = 0.27$), and there the reference checkpoint in the full configuration is also above the published 83.1. The improved teacher reaches 94.6 / 97.8 on HisIR19 in the full configuration, and the raw-input student is above it (+0.68, 95% CI $[+0.55, +0.80]$, $p < 10^{-4}$).

BullingerDB [43] extends this comparison to a third corpus. Its authors evaluate the 16,502 test pages of 632 writers by mAP, Top-1 and a temporal nDCG, in which the relevance of a same-writer match decays with the distance in years between query and match. The two-stage method is applied to this corpus with three corpus-specific settings. The binarisation provided with the dataset, produced with the method of Gatos et al. [12], is the binarised view in both stages, as the teacher's training input and as the student's privileged view. All pages (raw and binarised, training and test) are downscaled by a fixed factor of 0.5 (a hardware constraint), the colour pages re-saved as JPEG at quality 92. Stage 1 uses bfloat16 instead of float16 mixed precision, for numerical stability on this corpus. Otherwise the teacher is trained as in Sec. 4.2 and the student distilled as in Sec. 4.3. The codebook is fitted on a random subset of 2,000 training pages, PCA and exemplar-SVM negatives on all training pages. Under the corpus's protocol, which ranks the final descriptors by plain cosine without query expansion, the student is level with the strongest published baseline on this corpus, AttMask+VLAD trained on the same split, with the PCA-whitened VLAD descriptors of the base configuration, and above it with the exemplar-SVM descriptors of the full configuration (Table 4; published values as reported, not paired). Under our own ranking protocol the two rows give 80.2 / 95.5 and 82.8 / 96.9 (mAP / Top-1).

Table 5 Zero-shot transfer to GRK-Papyri (% mAP / % Top-1): leave-one-out over the 50 photographs, codebook fitted without labels on the corpus, no PCA, no full configuration (see text). Left, tokens selected by the standard foreground gate (Sec. 4.4); right, a foreground mask taken from the binarised pages published with GRK-120 [42] as the token selector for every model, with network inputs and weights unchanged. With n=50 queries all values are descriptive.

	standard gate		external mask	
Model	mAP	Top-1	mAP	Top-1
Reference checkpoint	27.3	26.0	29.0	26.0
Final student	33.8	36.0	35.0	40.0

5.5 Generalisation

The final study tests transfer beyond the training benchmarks: to papyri, where binarisation fails (GRK-Papyri), to modern handwriting (CVL), and to BullingerDB (trained-on-target in Table 4, zero-shot here). Apart from the trained-on-target rows repeated in Table 7, every row of this study is zero-shot: the feature extractors are the frozen models of the previous subsections, no target corpus contributes training data, and no writer labels of a target corpus are used; per-corpus evaluation settings are stated with each experiment and applied identically to the compared models. The final student is compared with the reference checkpoint on GRK-Papyri and CVL, and evaluated alone on BullingerDB. Paired statistics are used where the query count permits.

GRK-Papyri

Top-1 is the metric of the original publication [34] and mAP that of the recent literature on this dataset. With n=50 queries and four to seven samples per writer, all rows on this corpus are reported descriptively, following the practice of this literature [34, 35, 40]. The VLAD codebook is fitted, without any labels, on the 50 evaluated photographs themselves, as in the corpus's own literature; the fit is therefore transductive, unlike the CVL fit of Raven et al. [47], which uses CVL's writer-disjoint training split. PCA is not applied, since 50 pages cannot support a 384-dimensional fit. The full configuration is not applied, since its exemplar SVMs need in-domain negatives and none exist for this corpus, a constraint also noted by Christlein et al. [8]. Query expansion is kept, since it has no fitted parameters.

Two settings are reported. In the first, tokens are selected by the standard foreground gate, and each model receives its usual input; the student reads the raw photograph. In the second, a foreground mask replaces the gate as the token selector, identically for every model. The mask comes from the binarised pages published with GRK-120, an extension of this corpus [42]; they were produced there by a U-Net binariser trained on the papyri of DIBCO 2019 [45] without writer labels (whether any of its training images is a GRK-Papyri page is not stated by its authors), and we threshold them to two levels. This gives both models the corpus-adapted foreground knowledge that papyri-specialised systems build in, for token selection only; the learned binarisation never becomes a network input. Selecting with a binarised map rather than feeding it to the network has precedent on papyri, in fragment retrieval [44] and in the pretraining of SAGHOG [40]. The second setting thus measures how much of the gap to corpus-adapted systems is foreground detection rather than representation. The reference checkpoint reading a learned binarisation as its network input is not evaluated, so whether a better binariser closes the gap is not measured here.

Descriptively (Table 5), the student is +6.5 mAP / +10.0 Top-1 above the reference checkpoint with the standard gate, and +6.0 / +14.0 under the shared mask. The external mask raises the mAP of both models by a similar amount (reference checkpoint +1.7, student +1.2); the student's Top-1 rises from 36.0 to 40.0, while that of the reference checkpoint stays at 26.0. Training both stages on papyri whose binarisation is reliable, as Sec. 5.4 does for BullingerDB with that corpus's provided binarised training pages, would need binarisation only for those training pages. Published results use the same leave-one-out protocol and mAP definition [40]. The published systems ahead of these values are all adapted to the material in some form: learned binarisation adapted to papyri, unsupervised or self-supervised training on the corpus images, or model ensembling [8, 40]; they remain above both settings measured here. Neither of our training stages uses a papyrus image, and our pipeline uses the gallery only for the codebook fit, query expansion and, in

Table 6 Zero-shot transfer to CVL (% mAP / % Top-1): 283 test writers, 1,415 query pages, foreground gate disabled for both models, fully frozen fitting (codebook, PCA and exemplar negatives fitted on the HisIR19 validation set under the same disabled gate; nothing of CVL is used for training or fitting). Paired comparisons are quoted in the text.

	base config.		full config.	
Model	mAP	Top-1	mAP	Top-1
Reference checkpoint	98.2	99.6	96.9	99.6
Final student	98.6	99.7	98.2	99.8

Table 7 Zero-shot transfer to BullingerDB (% mAP, nDCG, % Top-1; the corpus's cosine ranking protocol, pages downscaled by 0.5): codebook, PCA and exemplar negatives fitted on the HisIR19 validation set. Trained-on-target rows as in Table 4.

Model	mAP	nDCG	Top-1
Trained on BullingerDB			
AttMask+VLAD [43]	78.3	90.3	95.6
Ours, full config.	81.9	92.1	97.1
Zero-shot			
Final student, base config.	73.9	88.6	95.6
Final student, full config.	75.9	89.7	96.7

the second setting, the external mask. The closest published comparator keeps the on-corpus feature learning of Christlein et al. [8] but replaces its learned binarisation with classical methods, reaching 33.0 mAP at its best [1]; both student results, with no binarisation of the network input and no training on the corpus, are above this level (33.8 and 35.0 mAP).

CVL

CVL contains modern handwriting on white paper. Its crops contain handwriting only, and on pages written in light ink the gate can discard every token, so the gate is disabled for both models (Table 6). The published CVL result of Raven et al. [47], 98.6, uses a codebook fitted on CVL's 27 training writers; ours does not. In the base configuration the student is above the reference checkpoint (+0.36, 95% CI $[+0.06, +0.67]$, $p = 0.016$), and its 98.6 equals the published number without using any CVL data. The full configuration lowers the mAP of both models on this corpus (by 1.35 points for the reference checkpoint and 0.38 for the student); its exemplar negatives come from the HisIR19 validation set. On BullingerDB, negatives from the same validation set raise the zero-shot row (below), so the effect of cross-corpus negatives is corpus-dependent. Within the full configuration the student's margin over the reference checkpoint grows to +1.33 (95% CI $[+0.99, +1.68]$, $p < 10^{-4}$).

BullingerDB

The final student, trained only on Historical-WI, is evaluated with fully frozen fitting (Table 7). Against the published trained-on-target baseline, the zero-shot student is 4.4 mAP and 1.7 nDCG points lower in the base configuration and 2.4 mAP and 0.6 nDCG points lower in the full configuration, where its Top-1 is 1.1 points higher. In the full configuration it is also 6.0 mAP points below our own trained-on-target student. Under our own ranking protocol the two zero-shot rows give 75.0 / 95.3 and 75.2 / 96.0 (mAP / Top-1). No binarised-input model was measured zero-shot on this corpus, so these rows are not a relative transfer comparison.

6 Discussion and conclusion

The setting gives the student four assets: an input from which the teacher's input is computed and which keeps what binarisation removes; a relational target that transfers structure without coordinates; a large-scale pretrained initialisation that the teacher (trained from scratch on 1,182 pages) does not have; and more capacity (ViT-B/16 against ViT-S/16). The student ends above its own teacher while receiving no signal beyond that teacher's outputs, so the gap must come from these assets. The pretrained backbone alone, at the student's capacity, does not explain it: undistilled, it is about 7 points below the final student. The relational term adds +0.46 points, and the representation ablation shows that token width does not explain the margin, since the student scores higher with its 384-dimensional projected tokens than with its own 768-dimensional ones (Sec. 5.3). No student was trained on the binarised view or at the teacher's capacity, so the shares of the richer input and of the larger, pretrained network remain unquantified.

On papyrus photographs (GRK-Papyri) the frozen student is 6.5 mAP points above the reference checkpoint with the standard foreground gate, and 6.0 when a corpus-adapted token selector is granted to both models; the strongest papyri-adapted published systems remain above both settings, while a corpus-trained pipeline with classical binarisation (33.0 mAP [1]) is below the student's level. On a large unseen historical corpus (BullingerDB) the frozen student is 2.4 to 4.4 mAP points (with and without exemplar-SVM re-encoding) below the corpus's published trained-on-target baseline; trained on that corpus with the same two-stage method, it is level with that baseline in the base configuration and above it in the full configuration (Table 4). Two limitations remain. First, the leave-one-out papyri benchmark is too small (n=50) for statistical claims, and closing the gap to papyri-specialised systems there would require adaptation to the material, which we deliberately avoid in this study. Second, the distillation transfers the teacher's notion of foreground. Where a corpus differs strongly from the training corpus (for example, in script or with dense marginalia), both stages can be retrained on pages of that domain, and only these training pages need a reliable binarisation. The BullingerDB experiment follows this path with the corpus's provided binarised training pages; for papyri it is untested.

We reproduced the state of the art in historical writer retrieval, with protocol details confirmed by its authors, improved it on binarised input with a page-level supervised-contrastive loss, and exceeded it with a raw-input student trained by privileged-information distillation. As a network input, the binarised view is confined to training time, where it is reliable; at inference the network receives the raw image. Beyond document analysis, we hypothesise that the same composition applies wherever a preprocessing step that is reliable only on well-behaved input, such as segmentation in medical imaging or cloud and shadow masking in remote sensing, limits an otherwise capable model.

Acknowledgements. We thank Tim Raven for correspondence that confirmed protocol details of the reproduction and for providing the training checkpoint.

Statements and Declarations

Funding. The research for this article was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2176 'Understanding Written Artefacts: Material, Interaction and Transmission in Manuscript Cultures', project no. 390893796. The research was conducted at the Centre for the Study of Manuscript Cultures (CSMC) at Universität Hamburg.

Competing Interests. The authors have no competing interests.

Ethics approval. Not applicable.

Data availability. All benchmarks used in this study are publicly available.

Code availability. Code and configurations reproducing all reported results are available at https://github.com/vma-lab/WR_by_PID.

Author contributions. H.M. is the sole author of this work.

d

Appendix A Reproduction details

Protocol details confirmed by the authors. The authors confirmed by correspondence: the training batch is 128 on a single GPU, giving a peak learning rate of 5×10^{-4} (the printed 5×10^{-3} is a typographical error); the HisIR19 validation set is the fitting corpus for both the VLAD codebook and the PCA; the Sauvola window is 51; and the reference checkpoint is the one underlying the published results.

Checkpoint arguments and inference path. The reference checkpoint stores the argument list of its training run; 56 of its 62 keys equal those of our reproduction, and the other six are logging and path settings with no effect on the learned weights. Our checkpoint conversion is tensor-identical over all 150 tensors, and identical inputs give identical class and patch tokens in our wrapper and in the authors' released model code.

Known discrepancies. The released preparation code binarises the HisIR19 evaluation pages with window 25 while the publication states 51;

measured on the reference checkpoint, the difference is −0.02 points (at an earlier page geometry).

The window rule. The publication states that windows with less than 2.5% stroke pixels are discarded; the released code does not apply this rule, and every other row of this paper keeps all windows, as that code does. Our implementation of the stated rule, used only in the first row of Table 1 and in the measurement below, measures the stroke pixels relative to the bounding box of the ink in each window, which keeps more windows than a window-area basis would. Measured on the reference checkpoint, enabling it changes mAP by −0.22 points in the base configuration and by −0.01 in the full configuration.

Solver and geometry. The PCA solver is immaterial: randomised minus exact SVD gives −0.02 points on the first row of Table 1 and −0.04 on the row that adds our VLAD normalisation. The foreground gate of the base configuration (4%) and the published $t_{\mathrm{fg}} = 10/256$ differ only for patches with exactly ten ink pixels, so Table 1 treats them as one element. Every test-set measurement uses the page at its native resolution, without rescaling or padding, except the BullingerDB rows (all pages at half scale, Sec. 5.4) and the window-25 check above (an earlier page geometry). The provided Historical-WI binarised pages are stored inverted (white ink on black); our evaluation code re-binarises them with the same Sauvola step as every binarised-input page, so every binarised view reaches the network as white ink on black.

Appendix B Statistical procedure

Every comparison between two models in one configuration is a paired per-query bootstrap on the identical scored-query set; differences between evaluation configurations of one model, including the upper-block rows of Table 1, are reported as plain differences without a test. The statistic is the mean difference of per-query average precision, in percentage points. Query indices are resampled with replacement 20,000 times with a fixed seed; the 95% confidence interval is the 2.5th and 97.5th percentile of the resampled means, and the two-sided p-value is twice the smaller of the two tail fractions of resampled means at or below and at or above zero. Top-1 differences are reported descriptively, and query wins and losses count the positive and negative per-query differences. The released code contains the exact tool and the per-query files behind every reported test, so each test can be recomputed bit-identically. Queries of the same writer share their relevant pages and are therefore not independent. Resampling writers instead of queries (an option of the released tool) is expected to widen the intervals.

On multiple comparisons: the paper's primary claims rest on effects reported at $p < 10^{-4}$, which survive any standard correction over the number of tests conducted. The teacher improvement is supported by two pairings that share the teacher's per-query results ($p = 0.006$ against the reference checkpoint, $p = 0.002$ against our reproduction) and by its student-side transfer at $p < 10^{-4}$; beyond that, p-values between 0.001 and 0.05 are interpreted cautiously and no single one carries a claim on its own.

Appendix C Hyper-parameters

Table C1 lists the distillation hyper-parameters and, below the rule, the two Stage-1 settings added with the page-level loss; all other Stage-1 settings follow the published recipe [47]. All selected values were fixed from validation-set measurements before any test-set evaluation. The map W is excluded from weight decay and layer-wise decay; clipping covers the student and W, and mixed precision also covers the Gram matrices of Eq. (4). BullingerDB differs in teacher view and page scale (Sec. 5.4).

References

[1] Akt D, Peer M, Kleber F (2025) Assessing the impact of binarization for writer identification in Greek papyrus. arXiv:2506.15852, https://doi.org/10.48550/arXiv.2506.15852, accepted at the Austrian Symposium on AI, Robotics, and Vision (AIRoV 2025)

[2] Arandjelović R, Zisserman A (2013) All about VLAD. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp 1578–1585

Table C1 Distillation hyper-parameters.

Parameter	Value
Window / patch size	224 / 16
Map W	affine, 768→384
Teacher view (per tile)	Sauvola 51, k=0.2
Batch size	96
Optimiser	AdamW [30]
Peak learning rate	10^{-4}
Layer-wise decay [3]	0.75
Weight decay	0.04
Epochs (linear warm-up)	5 (1)
Annealing	cosine to zero [29]
Gradient-norm clipping	3.0
Mixed precision	bfloat16
λ_{cls}, β, Eqs. (3), (4)	0.5, 5
Tile overlap (training)	0.5
Tile keep-filter (Sec. 4.3)	box $>$ 40%, edges $\geq$ 4%
Page geometry (training)	native resolution
λ_{sc}, τ (Stage 1)	0.5, 0.1
Pages $\times$ windows (Stage 1)	16×8

[3] Bao H, Dong L, Piao S, et al (2022) BEiT: BERT pre-training of image transformers. In: Proc. Int. Conf. on Learning Representations (ICLR)

[4] Caron M, Touvron H, Misra I, et al (2021) Emerging properties in self-supervised vision transformers. In: Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), pp 9630–9640, https://doi.org/10.1109/ICCV48922.2021.00951

[5] Christlein V, Bernecker D, Hönig F, et al (2017) Writer identification using GMM supervectors and exemplar-SVMs. Pattern Recognition 63:258–267. https://doi.org/10.1016/j.patcog.2016.10.005

[6] Christlein V, Gropp M, Fiel S, et al (2017) Unsupervised feature learning for writer identification and writer retrieval. In: Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), pp 991–997, https://doi.org/10.1109/ICDAR.2017.165

[7] Christlein V, Nicolaou A, Seuret M, et al (2019) ICDAR 2019 competition on image retrieval for historical handwritten documents. In: Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), pp 1505–1509, https://doi.org/10.1109/ICDAR.2019.00242

[8] Christlein V, Marthot-Santaniello I, Mayr M, et al (2022) Writer retrieval and writer identification in Greek papyri. In: Intertwining Graphonomics with Human Movements – IGS 2022, Lecture Notes in Computer Science, vol 13424. Springer, Cham, pp 76–89, https://doi.org/10.1007/978-3-031-19745-1_6

[9] Delhumeau J, Gosselin PH, Jégou H, et al (2013) Revisiting the VLAD image representation. In: Proc. ACM Int. Conf. on Multimedia (MM), pp 653–656, https://doi.org/10.1145/2502081.2502171

[10] Fiel S, Kleber F, Diem M, et al (2017) ICDAR2017 competition on historical document writer identification (Historical-WI). In: Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), pp 1377–1382, https://doi.org/10.1109/ICDAR.2017.225

[11] Furlanello T, Lipton Z, Tschannen M, et al (2018) Born again neural networks. In: Proc. Int. Conf. on Machine Learning (ICML), Proceedings of Machine Learning Research, vol 80. PMLR, pp 1607–1616

[12] Gatos B, Pratikakis I, Perantonis SJ (2006) Adaptive degraded document image binarization. Pattern Recognition 39(3):317–327. https://doi.org/10.1016/j.patcog.2005.09.010

[13] Gupta S, Hoffman J, Malik J (2016) Cross modal distillation for supervision transfer. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp 2827–2836, https://doi.org/10.1109/CVPR.2016.309

[14] Hao Z, Guo J, Jia D, et al (2022) Learning efficient vision transformers via fine-grained manifold distillation. In: Advances in Neural Information Processing Systems (NeurIPS), vol 35. Curran Associates, Inc., pp 9164–9175

[15] He S, Schomaker L (2019) DeepOtsu: Document enhancement and binarization using iterative deep learning. Pattern Recognition 91:379–390. https://doi.org/10.1016/j.patcog.2019.01.025

[16] He S, Schomaker L (2020) FragNet: Writer identification using deep fragment networks. IEEE Transactions on Information Forensics and Security 15:3013–3022. https://doi.org/10.1109/TIFS.2020.2981236

[17] Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. NIPS Deep Learning and Representation Learning Workshop. arXiv:1503.02531, https://doi.org/10.48550/arXiv.1503.02531

[18] Hoffman J, Gupta S, Darrell T (2016) Learning with side information through modality hallucination. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp 826–834, https://doi.org/10.1109/CVPR.2016.96

[19] Jégou H, Douze M, Schmid C, et al (2010) Aggregating local descriptors into a compact image representation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp 3304–3311, https://doi.org/10.1109/CVPR.2010.5540039

[20] Jordan S, Seuret M, Král P, et al (2020) Re-ranking for writer identification and writer retrieval. In: Document Analysis Systems – DAS 2020, Lecture Notes in Computer Science, vol 12116. Springer, Cham, pp 572–586, https://doi.org/10.1007/978-3-030-57058-3_40

[21] Kakogeorgiou I, Gidaris S, Psomas B, et al (2022) What to hide from your students: Attention-guided masked image modeling. In: Computer Vision – ECCV 2022, Lecture Notes in Computer Science, vol 13690. Springer, Cham, pp 300–318, https://doi.org/10.1007/978-3-031-20056-4_18

[22] Keglevic M, Fiel S, Sablatnig R (2018) Learning features for writer retrieval and identification using triplet CNNs. In: Proc. Int. Conf. on Frontiers in Handwriting Recognition (ICFHR), pp 211–216, https://doi.org/10.1109/ICFHR-2018.2018.00045

[23] Khosla P, Teterwak P, Wang C, et al (2020) Supervised contrastive learning. In: Advances in Neural Information Processing Systems (NeurIPS), vol 33. Curran Associates, Inc., pp 18661–18673

[24] Kleber F, Fiel S, Diem M, et al (2013) CVL-DataBase: An off-line database for writer retrieval, writer identification and word spotting. In: Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), pp 560–564, https://doi.org/10.1109/ICDAR.2013.117

[25] Koepf M, Kleber F, Sablatnig R (2022) Writer identification and writer retrieval using vision transformer for forensic documents. In: Document Analysis Systems – DAS 2022, Lecture Notes in Computer Science, vol 13237. Springer, Cham, pp 352–366, https://doi.org/10.1007/978-3-031-06555-2_24

[26] Lai S, Zhu Y, Jin L (2020) Encoding pathlet and SIFT features with bagged VLAD for historical writer identification. IEEE Transactions on Information Forensics and Security 15:3553–3566. https://doi.org/10.1109/TIFS.2020.2991880

[27] Liu Y, Chen K, Liu C, et al (2019) Structured knowledge distillation for semantic segmentation. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp 2604–2613, https://doi.org/10.1109/CVPR.2019.00271

[28] Lopez-Paz D, Bottou L, Schölkopf B, et al (2016) Unifying distillation and privileged information. In: Proc. Int. Conf. on Learning Representations (ICLR)

[29] Loshchilov I, Hutter F (2017) SGDR: Stochastic gradient descent with warm restarts. In: Proc. Int. Conf. on Learning Representations (ICLR)

[30] Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: Proc. Int.

Conf. on Learning Representations (ICLR)

[31] Malisiewicz T, Gupta A, Efros AA (2011) Ensemble of exemplar-SVMs for object detection and beyond. In: Proc. IEEE Int. Conf. on Computer Vision (ICCV), pp 89–96, https://doi.org/10.1109/ICCV.2011.6126229

[32] Mattick A, Mayr M, Seuret M, et al (2024) Evaluating learned feature aggregators for writer retrieval. International Journal on Document Analysis and Recognition (IJDAR) 27(3):265–274. https://doi.org/10.1007/s10032-024-00482-x

[33] Mohammed H, Märgner V, Konidaris T, et al (2017) Normalised local naïve Bayes nearest-neighbour classifier for offline writer identification. In: Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), pp 1013–1018

[34] Mohammed H, Marthot-Santaniello I, Märgner V (2019) GRK-Papyri: A dataset of Greek handwriting on papyri for the task of writer identification. In: Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), pp 726–731, https://doi.org/10.1109/ICDAR.2019.00121

[35] Nasir S, Siddiqi I, Moetesum M (2021) Writer characterization from handwriting on papyri using multi-step feature learning. In: Document Analysis and Recognition – ICDAR 2021 Workshops, Lecture Notes in Computer Science, vol 12916. Springer, Cham, pp 451–465, https://doi.org/10.1007/978-3-030-86198-8_32

[36] Park W, Kim D, Lu Y, et al (2019) Relational knowledge distillation. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp 3967–3976, https://doi.org/10.1109/CVPR.2019.00409

[37] Peer M, Sablatnig R (2023) Feature mixing for writer retrieval and identification on papyri fragments. In: Proc. ACM Int. Workshop on Historical Document Imaging and Processing (HIP), pp 31–36, https://doi.org/10.1145/3604951.3605515

[38] Peer M, Kleber F, Sablatnig R (2022) Self-supervised vision transformers with data augmentation strategies using morphological operations for writer retrieval. In: Frontiers in Handwriting Recognition – ICFHR 2022, Lecture Notes in Computer Science, vol 13639. Springer, Cham, pp 122–136, https://doi.org/10.1007/978-3-031-21648-0_9

[39] Peer M, Kleber F, Sablatnig R (2023) Towards writer retrieval for historical datasets. In: Document Analysis and Recognition – ICDAR 2023, Lecture Notes in Computer Science, vol 14187. Springer, Cham, pp 411–427, https://doi.org/10.1007/978-3-031-41676-7_24

[40] Peer M, Kleber F, Sablatnig R (2024) SAGHOG: Self-supervised autoencoder for generating HOG features for writer retrieval. In: Document Analysis and Recognition – ICDAR 2024, Lecture Notes in Computer Science, vol 14805. Springer, Cham, pp 121–138, https://doi.org/10.1007/978-3-031-70536-6_8

[41] Peer M, Sablatnig R, Kleber F (2025) Towards the influence of text quantity on writer retrieval. In: Document Analysis and Recognition – ICDAR 2025, Lecture Notes in Computer Science, vol 16024. Springer, Cham, pp 129–145, https://doi.org/10.1007/978-3-032-04617-8_8

[42] Peer M, Sablatnig R, Serbaeva O, et al (2025) KaiRacters: Character-level-based writer retrieval for Greek papyri. In: Pattern Recognition – ICPR 2024, Lecture Notes in Computer Science, vol 15319. Springer, Cham, pp 73–88, https://doi.org/10.1007/978-3-031-78495-8_5

[43] Peer M, Scius-Bertrand A, Scheurer P, et al (2027) BullingerDB: A dataset for handwritten text recognition and writer retrieval. In: Document Analysis and Recognition – ICDAR 2026, Lecture Notes in Computer Science, vol 16973. Springer, Cham, pp 555–571, https://doi.org/10.1007/978-3-032-36033-5_33

[44] Pirrone A, Beurton-Aimar M, Journet N (2021) Self-supervised deep metric learning for ancient papyrus fragments retrieval. International Journal on Document Analysis and Recognition (IJDAR) 24(3):219–234. https://doi.org/10.1007/s10032-021-00369-1

[45] Pratikakis I, Zagoris K, Karagiannis X, et al (2019) ICDAR 2019 competition on document image binarization (DIBCO 2019). In: Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), pp 1547–1556, https://doi.org/10.1109/ICDAR.2019.00249

[46] Qin D, Gammeter S, Bossard L, et al (2011) Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp 777–784, https://doi.org/10.1109/CVPR.2011.5995373

[47] Raven T, Matei A, Fink GA (2024) Self-supervised vision transformers for writer retrieval. In: Document Analysis and Recognition – ICDAR 2024, Lecture Notes in Computer Science, vol 14805. Springer, Cham, pp 380–396, https://doi.org/10.1007/978-3-031-70536-6_23

[48] Raven T, Christlein V, Fink GA (2026) Interpretable writer recognition via vectors of locally aggregated characters. In: Document Analysis and Recognition – ICDAR 2025, Lecture Notes in Computer Science, vol 16026. Springer, Cham, pp 429–445, https://doi.org/10.1007/978-3-032-04627-7_25

[49] Raven T, Hallyburton T, Fink GA (2027) Writer retrieval at scale. In: Document Analysis and Recognition – ICDAR 2026, Lecture Notes in Computer Science, vol 16974. Springer, Cham, pp 37–53, https://doi.org/10.1007/978-3-032-36039-7_3

[50] Romero A, Ballas N, Kahou SE, et al (2015) FitNets: Hints for thin deep nets. In: Proc. Int. Conf. on Learning Representations (ICLR)

[51] Sauvola J, Pietikäinen M (2000) Adaptive document image binarization. Pattern Recognition 33(2):225–236. https://doi.org/10.1016/S0031-3203(99)00055-2

[52] Seuret M, Nicolaou A, Stutzmann D, et al (2020) ICFHR 2020 competition on image retrieval for historical handwritten fragments. In: Proc. Int. Conf. on Frontiers in Handwriting Recognition (ICFHR), pp 216–221, https://doi.org/10.1109/ICFHR2020.2020.00048

[53] Siméoni O, Vo HV, Seitzer M, et al (2025) DINOv3. arXiv:2508.10104, https://doi.org/10.48550/arXiv.2508.10104

[54] Tung F, Mori G (2019) Similarity-preserving knowledge distillation. In: Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), pp 1365–1374, https://doi.org/10.1109/ICCV.2019.00145

[55] Vapnik V, Vashist A (2009) A new learning paradigm: Learning using privileged information. Neural Networks 22(5–6):544–557. https://doi.org/10.1016/j.neunet.2009.06.042

[56] Watanabe S, Hori T, Le Roux J, et al (2017) Student-teacher network learning with enhanced features. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp 5275–5279, https://doi.org/10.1109/ICASSP.2017.7953163

[57] Xie Q, Luong MT, Hovy E, et al (2020) Self-training with noisy student improves ImageNet classification. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp 10687–10698, https://doi.org/10.1109/CVPR42600.2020.01070

[58] Xing L, Qiao Y (2016) DeepWriter: A multi-stream deep CNN for text-independent writer identification. In: Proc. Int. Conf. on Frontiers in Handwriting Recognition (ICFHR), pp 584–589, https://doi.org/10.1109/ICFHR.2016.0112

[59] Zhou J, Wei C, Wang H, et al (2022) iBOT: Image BERT pre-training with online tokenizer. In: Proc. Int. Conf. on Learning Representations (ICLR)