SAGHOG: Self-Supervised Autoencoder for Generating HOG Features for Writer Retrieval

Marco Peer, Florian Kleber, and Robert Sablatnig

Computer Vision Lab, TU Wien
{mpeer, kleber, sab}@cvl.tuwien.ac.at
Code: http://github.com/marco-peer/icdar24

Abstract. This paper introduces SAGHOG, a self-supervised pretraining strategy for writer retrieval using HOG features of the binarized input image. Our preprocessing involves the application of the Segment Anything technique to extract handwriting from various datasets, ending up with about 24k documents, followed by training a vision transformer on reconstructing masked patches of the handwriting. SAGHOG is then finetuned by appending NetRVLAD as an encoding layer to the pretrained encoder. Evaluation of our approach on three historical datasets, Historical-WI, HisFrag20, and GRK-Papyri, demonstrates the effectiveness of SAGHOG for writer retrieval. Additionally, we provide ablation studies on our architecture and evaluate un- and supervised finetuning. Notably, on HisFrag20, SAGHOG outperforms related work with a mAP of 57.2 % - a margin of 11.6 % to the current state of the art, showcasing its robustness on challenging data, and is competitive on even small datasets, e.g. GRK-Papyri, where we achieve a Top-1 accuracy of 58.0 %.

Keywords: Writer Retrieval · Self-Supervised Learning · Masked Autoencoder · Document Analysis.

1 Introduction

Writer Retrieval (WR) is the task of locating documents written by the same author within a dataset, achieved by identifying similarities in handwriting [18]. This task is particularly valuable for historians and paleographers, allowing them to track individuals or social groups across various historical periods [9]. Additionally, WR is used for recognizing documents with unknown authors and uncovering similarities within such documents [6].

WR methods usually consist of multiple steps, such as sampling of handwriting by applying interest point detection, a neural network for feature extraction, and feature encoding such as NetVLAD, followed by aggregation of the encoded features [29]. Although those methods work for large datasets such as Historical-WI [14], WR still lacks performance for datasets that contain less handwriting or noise such as degradation, best exemplified by HisFrag20 [32]. While approaches trained on full fragments currently work best [26], experiments show that those methods do infer features from the background, not necessarily related on the actual handwriting [30].

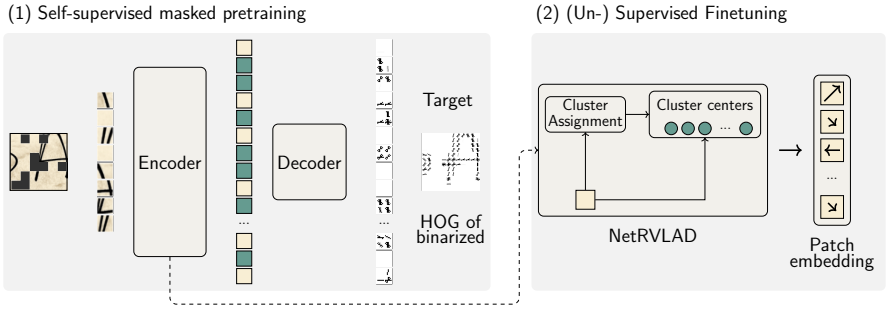


Fig. 1: Overview of SAGHOG. The decoder reconstructs the HOG features of the masked tokens (■) by only using the non-masked tokens (□). In the second stage, the pretrained encoder with NetRVLAD is finetuned, either by training on writer or pseudo labels.

With the recent gain of interest in self-supervised algorithms that make use of large amounts of unlabeled data, this paper introduces **S**elf-Supervised **A**utoencoder for **G**enerating **HOG** features (SAGHOG), a self-supervised pretraining strategy based on Histogram of Oriented Gradients (HOG) features. SAGHOG is two-staged and relies on the Vision Transformer (ViT) architecture, as shown in Fig. 1: Firstly, we aim to close that gap of performance for complex datasets by self-supervised pretraining on reconstructing only HOG features of the handwriting. Since related work relies on clustering SIFT descriptors [6,29], predicting HOG features is an intuitive choice. Secondly, we add NetRVLAD [29] to the pretrained encoder and finetune the model on the respective dataset. We apply and evaluate two finetuning strategies: supervised - using the writer label - or unsupervised finetuning - using the cluster of the corresponding SIFT descriptor of the input patch as a surrogate class (Cl-S) [6]. Additionally, as a preprocessing step, we propose to apply Segment Anything Model (SAM) [20], a foundation model for segmentation, to extract the handwriting of the data, accelerate the training step, and improve the generalization ability of SAGHOG.

We evaluate our results on three datasets: Historical-WI [14], HisFrag20 [32] and GRK-Papyri [25], each corresponding to a different domain. We provide ablation studies on our preprocessing stage, different target features of SAGHOG and hyperparameters of the pretraining stage and the network architecture. Our experiments show that HOG reconstruction during pretraining does improve WR when finetuning with NetRVLAD, in particular on HisFrag20, where we outperform the state of the art by a large margin (11.6 %). Furthermore, we find that only optimizing NetRVLAD is able to outperform finetuning the network, showing that SAGHOG is beneficial for WR. On Historical-WI, the effect of pretraining is limited, but using SAGHOG does not harm the performance and is competitive with other approaches. In the end, we show that even on small, out-of-domain datasets such as GRK-Papyri, SAGHOG shines, in particular in terms

of Top-1 accuracy, making ViTs valuable for further research in the domain of WR.

Concluding, our contributions are summarized to:

- We collect multiple datasets and propose a data curation scheme to enable pretraining of SAGHOG.
- SAGHOG is the first self-supervised masking-based approach for handwriting of 32×32 patches without the need for any carefully designed data augmentation pipeline. We investigate two different finetuning schemes - supervised and unsupervised.
- We thorougly evaluate our approach on Historical-WI [14], HisFrag20 [32] and GRK-Papyri [25] and outperform state-of-the-art methods on WR as a downstream task.

Our code is publicly available. The remainder of our paper is structured as follows. Section 2 gives details on related work in the field of self-supervised learning and WR. Afterwards, we describe SAGHOG in Section 3, including pretraining and finetuning, followed by our evaluation protocol in Section 4 and our results in Section 5. We summarize our findings in Section 6.

2 Related Work

In this section, we briefly discuss related work. We start with the developments in self-supervised learning with a focus on WR, followed by the current state-of-the-art WR methods.

Self-Supervised Learning In recent years, self-supervised learning has gained popularity to learn meaningful representations from unlabeled data. While metric learning approaches like SimCLR [4] or MoCo [16] involve training two separate networks to generate similar embeddings for a sample, they are outperformed by self-distillation methods, e.g. DINO [1], that use a carefully designed augmentation pipeline to generate different views of an image. In the domain of WR, we introduced a self-supervised algorithm based on DINO [1] using morphological operations to generate augmented views of the binarized Historical-WI dataset [28]. Lastilla et al. [22] apply an online bag-of-words pretraining to documents of the Vatican Apostolic Library for WR. Both approaches improve the Mean Average Precision (mAP) on the datasets evaluated, but they use large crops which slows down training. Additionally, they do not use any encoding or finetuning at all. However, those augmentation pipelines suffer when training on small patches only containing a few strokes of handwriting. Therefore, we focus on another approach of self-supervision called masked image modeling or Masked Autoencoder (MAE), introduced by He et al. [15], where a network learns to reconstruct masked parts of an image. An intuitive adaption of MAE is MaskFeat [36], proposed by Wei et al., where the target features are handcrafted descriptors rather than pixel values, e.g. HOG. SAGHOG also applies this strategy by

adapting MaskFeat to the binarized version of the input image. This is an intuitive choice, since dominating approaches for WR rely on SIFT descriptors [6,29] for training, and those descriptors are formed by aggregating HOG features.

Writer Retrieval WR methods fall into two categories: codebook-based and codebook-free approaches. Codebooks serve as models to compute handwriting statistics, with Vector of Locally Aggregated Descriptors (VLAD) standing out as a prominent method for WR [5,6,7,8]. Handwriting characteristics are extracted either through handcrafted features [5,21] or CNNs [6,7,8]. For historical datasets in particular, Christlein et al. [6] demonstrate that training on 32×32 patches with pseudolabels generated by clustering SIFT descriptors outperforms supervised methods [35]. Additionally, Exemplar SVM (ESVM) refines each descriptor's encoding by training multiple SVMs with only one positive sample and the remaining ones as negatives. The HisIR19 competition winners [9] and the current state-of-the-art on Historical-WI dataset utilize handcrafted features (SIFT and pathlet) for retrieval, encoding them via bagged VLAD (bVLAD) [21]. In our previous work [29] we employ NetRVLAD, a NetVLAD-based layer with reduced complexity, to train the neural network in and end-to-end way. Additionally, we introduced Similarity Graph Reranking (SGR), a reranking strategy that shows superior performance compared with state of the art. In contrast to training with small patches, on more complex datasets including fragments, such as HisFrag20 [32], leading approach train on full images: Ngo et al. [26] propose an attention-based VLAD called A-VLAD, while our feature mixer [30] uses an codebook-free encoding stage based on convolutional and fully connected layers. In this work, finetuning of SAGHOG is closely related to our training protocol of NetRVLAD [29], and we show that our approach is outperforming methods on image-level for datasets such as HisFrag20. SAGHOG is also based on the ViT architecture instead of relying on CNNs.

3 Methodology

In this section, we cover each part of our proposed approach. We start with the curation of the dataset for pretraining, followed by describing SAGHOG. In the end, NetRVLAD finetuning and feature aggregation are explained.

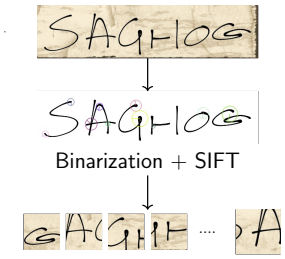
3.1 Data Curation

Self-supervised learning relies on a large amount of (unlabelled) data. Hence, we employ multiple datasets proposed in previous competitions in the domain of historical document analysis for our pretraining. For the datasets used, presented in Table 1, we are mainly inspired by the choice of Zenk et al. [37]. Our preprocessing scheme, also shown in Fig. 2, is two-fold: *handwriting extraction* and *patch sampling*.

Handwriting extraction Since datasets such as HisIR19 [9] contain additional elements like color palettes, book covers, and background noise, we use SAM

(a) Handwriting extraction via SAM

(b) Examples of input-output pairs



(c) Patch sampling

Fig. 2: Preprocessing scheme. (a) We obtain the handwriting of a page by processing the segmentation of SAM. (b) Examples of the produced images. (c) 32×32 patches are extracted by applying SIFT on the binarized image.

[20], a foundation model for segmentation. We use the default pretrained model and process the data as follows:

1. We apply a Canny edge detector on the image.
2. Since SAM relies on a grid of points and we only want to obtain the handwriting part that is usually a significant part of the page, we use a small 8×8 grid for segmentation.
3. To discard wrong results, we filter masks with less than 10 % edge pixels and a confidence level lower than 0.8.

If the preceding steps generate multiple masks, we merge them, restricting the maximum number of masks per page to two. Additionally, to address pages with minimal handwriting or none at all, we discard pages with fewer than 1000 SIFT keypoints detected. This procedure allows us to take smaller random crops of the pages to extract keypoints, which significantly accelerates the processing time during training and removes non-handwriting areas. As depicted in Fig. 2b, we occasionally oversegment and remove some handwriting. Note that we do not apply SAM on the trainset of Historical-WI [14] since it contains pure handwriting. To show the effectiveness of the preprocessing, we annotate handwriting areas of 200 pages of the HisIR19 [9] test set and evaluate the bounding boxes. We achieve an IoU@0.5 of 74 %. This is mostly due to SAM splitting the handwriting areas in multiple zones (see Fig. 2b). However, all of our detected masks

Table 1: Datasets used for self-supervised pretraining after preprocessing.

Dataset	Subset	Images
CLaMM 2016 [11]	Task 1 / Task 2	2607
Historical-WI [14]	Train (color)	1182
CLaMM 2017 [10]	Task 1 / Task 3	1572
cBAD 2019 [12]	Validation / Test	531
HisIR19 [9]	Validation / Test	1152 / 17641
Total		24685

are inside the annotated handwriting areas, indicating that we can successfully segment the handwriting - although we remove parts of it.

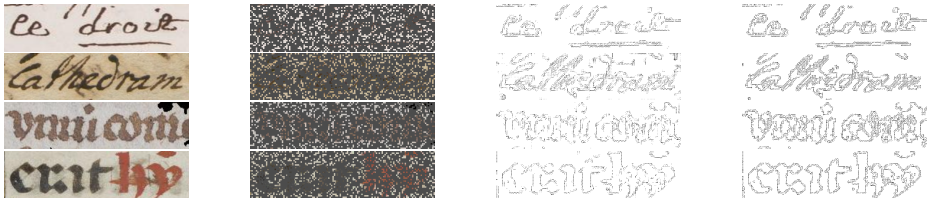
Patch sampling During training, we take a random crop of size 256 and sample small parts of the handwriting by extracting 32×32 patches located around SIFT keypoints. The keypoints are detected on the binarized version of the page via SIFT, and we discard patches with less than 1 % black pixels, similar to previous work [6,29]. We investigate the influence of the binarization in our evaluation in Section 5.

The composition of our final dataset used for pretraining is shown in Table 1. Although the majority of images are included in the HisIR19 [9] dataset, it contains various document types such as manuscript books, letters, or charters.

3.2 Self-Supervised Pretraining

SAGHOG is based on the MAE architecture [15] consisting of an encoder-decoder structure. We use the vanilla ViT [13] structure with a patch size of four. Firstly, random tokens are masked out - by default, we mask 75 % of the image. The encoder projects each unmasked token of the 32×32 input color image to an embedding of size 512. Those embeddings are then projected by the decoder to the output feature space: While MAE [15] and MaskFeat [36] use RGB pixel values or HOG features of the colorized input, we propose to reconstruct the HOG features of the binarized image due to two reasons. By relying on binarization, the network automatically learns to focus on handwriting, and on the other hand, we are able to suppress the influence of the background. HOG features are calculated by applying a convolutional layer with fixed weights to obtain the gradients. Afterward, we build a 9-bin histogram on cells of size 4×4, resulting in a feature vector of dimension 9 for each token. The final HOG features are l_2 normalized.

We choose, similar to related work [36], a larger encoder consisting of 8 layers (depth) and a smaller decoder (one layer) since this shows improved performance on the downstream task. We denote the encoder as ViT-4/8. Qualitative examples of the training process are depicted in Fig. 3. In general, the network is able to reconstruct the missing tokens (and therefore the style of handwriting as well



(a) (b) (c) (d)

Fig. 3: Examples of reconstructed HOG features. For better visualization, we stack multiple patches to create images of (128, 512). The images are not seen during pretraining. (a) Raw input image. (b) Masked input with a ratio of 0.75. (c) Prediction of the missing HOG features of the binarized image. (d) Actual HOG features.

as parts of letters), even if only small parts of the handwriting are visible. We evaluate the choice of target features as well as the hyperparameters of SAGHOG in our evaluation.

3.3 Finetuning

After pretraining of SAGHOG, we drop the decoder and only use the encoder for further finetuning. We use the class token as output of the encoder, a learnable embedding that describes the global information of the token sequence. We also studied pooling of all patch embeddings instead of using the class token, but this did not yield significant differences. Finetuning is then performed by appending the NetRVLAD layer [30] with 100 cluster centers and a final l_2-normalization of the encoded embedding. In our evaluation, we investigate two finetuning schemes commonly used in the domain of WR:

1. *Supervised*: The target label is the writer of the document, this refers to the default supervised finetuning strategy.
2. *Cl-S*: The target label is the cluster membership of the clustered SIFT descriptors extracted at the corresponding keypoints. We follow the protocol used in [29]: We cluster the trainset by using k-Means with 5000 centers and only consider patches that pass the ratio test between the distance to the nearest and the second nearest cluster (< 0.9). This has shown to perform better than supervised training for WR [6,29].

3.4 Writer Retrieval

We aggregate all patch embeddings $\mathbf{x}_i$ with $i = 0, \ldots, n_p - 1$ of a page using l_2 normalization followed by sum pooling $\mathbf{X} = \sum_{i=0}^{n_p-1} \mathbf{x}_i$ to obtain the global page descriptor $\mathbf{X}$. Furthermore, to reduce visual burstiness [17], we apply power-normalization $f(x) = \text{sign}(x)|x|^\alpha$ with $\alpha = 0.4$, followed by l_2-normalization. Finally, a dimensionality reduction with whitening via PCA is performed.

WR is then evaluated by a *leave-one-out strategy*: Each image of the test set is once used as a query q, and the remaining documents are ranked according to the cosine similarity of the global page descriptors.

4 Evaluation

We provide details on the datasets used for evaluation and our experimental setup in this section.

4.1 Datasets

(a) Historical-WI (b) HisFrag20 (c) GRK-Papyri

Fig. 4: Examples of the datasets used. (a) Binarized example of Historical-WI testset. (b) HisFrag20: Sample of train- and testset. (c) GRK-Papyri (Dioscorus-5) color and binarized version [8] - we only use the binarized set.

We evaluate our approach on three datasets, as shown in Fig. 4, each representing a different scenario:

Historical-WI Introduced by Fiel et al. [14], the dataset consists of 3600 images and 760 writers, each contributing five pages. We use the binarized version for evaluation, following related work. The training set includes 1192 pages. We consider Historical-WI as a dataset of moderate size and since it is binarized without containing any additional noise, handcrafted features [21] as well as small networks such as ResNet20 [6,29] work reasonably well.

HisFrag20 The main dataset used in our work is the HisFrag20 [32], introduced via the *ICFHR 2020 Competition on Image Retrieval for Historical Handwritten Fragments*. The trainset consists of about 100k fragments; the respective test set includes 20019 fragments from 1152 writers and 2732 documents. The generation algorithm for the fragments differs between the train- and test set, making the dataset challenging, which results in a lack of performance compared to Historical-WI.

GRK-Papyri Additionally, we provide baseline results for comparison on GRK-Papyri [25], a small dataset consisting of only 50 documents of 10 different writers. Note that this dataset represents a heavy domain shift compared to the datasets used for pretraining - from Latin to ancient Greek, as well as high degradation, e.g., holes, stains, and low contrast. To compare to related work, we use the U-Net binarized version provided by Christlein et al. [8]. Even though binarization tackles most of the degradation, binarized images still suffer from noise and missing strokes.

4.2 Experimental Setup

In the following, the setups for both training stages as well as for the aggregation step are given.

Pretraining We use the vanilla ViT architecture with a patch size of 4 and a depth of 8, denoted as ViT-4/8. The cell size for calculating HOG features is 4×4, and the decoder depth is 1, as related work [15,36] shows negligible influence of those parameters. During training, we take a random crop of size 256×256 and apply grayscale and binarization (both with a probability of $p = 0.2$). One batch consists of 64 pages, each contributing 32 patches, yielding a batch size of 2048. Further hyperparameters are given in Table 2a.

Table 2: Hyperparameters of our training, if not stated otherwise.

(a) Pretraining

Config	ViT-4/8
Optimizer	AdamW [24]
Weight decay	0.05
Learning rate	$8 \cdot 10^{-4}$
Scheduler	Cosine decay [23]
Batch size	2048
Gradient clipping	0.02
Warmup epochs	5
Epochs	200

(b) Finetuning

Config	ViT-4/8	ResNet
Optimizer	AdamW [24]	Adam [19]
Weight decay	0.01	0
Learning rate	10^{-3}	
Scheduler	Cosine decay [23]	
Batch size	1024	
Gradient clipping	1.0	
Warmup epochs	5	
Epochs	50	

Finetuning We finetune all networks by appending NetRVLAD [29] with 100 cluster centers to SAGHOG. The training is done for 50 epochs on triplets with Multi-Similarity loss [34] and a batch size of 1024 with 16 samples per class. For data augmentation, we apply erosion and dilation with a randomly sampled 3×3 kernel and a probability of 0.1. Additionally, token dropout is applied ($p = 0.1$). We use an early stopping of five epochs if the mAP on the validation set does not increase. All of our networks reach convergence with these settings (10 % of the writers are used as validation). Further hyperparameters are given in Table 2b.

Retrieval If not stated otherwise, we report the mAP on the task of WR for the respective datasets when SAGHOG is jointly finetuned with NetRVLAD in an unsupervised manner (Cl-S). The Top-1 metric indicates if the first document of the ranked list is a hit. The retrieval is performed on PCA-whitened global descriptors of dimension 512.

5 Results

In this section, we evaluate different parts of our pipeline and study the gain of performance by (un-)supervised finetuning of SAGHOG.

5.1 Ablation studies

Preprocessing Firstly, in Table 3, we study the influence of preprocessing via SAM. For both datasets, SAGHOG performs better when we preprocess data for pretraining with SAM, in particular on the HisFrag20 dataset, where we report a gain of 4.1 % mAP. Note that pretraining with no preprocessing includes 27k images instead of 24k due to the filtering applied after using SAM.

Table 3: SAM preprocessing.

	Historical-WI	HisFrag20
From scratch	71.1	47.4
Full image	72.8	50.5
With SAM	**73.3**	**54.6**

Target features Next, we evaluate the choice of the target features of the decoder. We train four networks with different features, as shown in Table 4: *Pixel* - which refers to the default MAE proposed by He et al. [15], and three different settings which refer to the modality used for calculating HOG features: *RGB*, *Gray*, *Binarized* (Bin.). While for both datasets, the largest gain of performance is reported for switching from pixel values to HOG features, SAGHOG works best when using HOG features of the binarized image as target on HisFrag20. All of our experiments outperform training from scratch, showing the effectiveness of pretraining.

We provide results for the critical parts of our pipeline in Table 5 when finetuning on Historical-WI and HisFrag20: the ratio of the masked patches, indicating the difficulty of the reconstruction task, the dimension of the encoder and the binarization algorithm used for patch sampling and calculating HOG features. We summarize our findings in the following:

Masking ratio SAGHOG performs best when using a *masking ratio* of 75 %. This is higher than values reported in related work ([15]: 60 %, [36]: 40 %). We

Table 4: Decoder target features.

Target feature	Historical-WI	HisFrag20
From scratch	71.1	47.4
Pixel (RGB)	71.9	50.6
HOG (RGB)	**73.3**	53.1
HOG (Gray)	**73.3**	54.1
HOG (Bin.)	**73.3**	**54.6**

Table 5: Ablation studies of SAGHOG on Historical-WI (HW) and HisFrag20 (HF) with finetuned NetRVLAD. Default settings are marked in gray.

(a) Masking ratio

	mAP	
Ratio	HW	HF
45	72.2	54.4
60	72.1	53.6
75	**73.3**	**54.6**
90	72.1	44.6

(b) Encoder dimension

	mAP	
Dim.	HW	HF
128	71.3	47.6
256	72.0	51.7
384	72.5	49.1
512	**73.3**	**54.6**

(c) Binarization algorithm

	mAP	
Method	HW	HF
Otsu [27]	73.0	54.1
Sauvola [31]	**73.3**	**54.6**
Su [33]	72.5	53.6

argue that this is mostly related to the fact that the background is not considered for the reconstruction task since we are only focusing on the HOG features of the handwriting. We observe a drop in performance when setting the masking ratio too high (90 %).

Embedding dimension The performance increases with higher *encoder dimensions*. The highest performance is obtained for 512 with a mAP of 73.3 % and 54.6 %. Note that the encoder dimension increases the network size, we assume that higher dimensions would further increase the performance, but this introduces an additional computational burden.

Binarization Finally, the choice of *binarization algorithm* is the least influencing factor. Even global binarization algorithms such as Otsu do perform well and close to our best one, Sauvola. We think this is because the data used does not contain heavy degradation, decreasing the difficulty of the task.

5.2 Finetuning

In this section, we assess two finetuning strategies: the conventional supervised approach (Sup.) and the unsupervised method (Cl-S [6]). While supervised finetuning is the commonplace default in various computer vision tasks, this shifts for WR, where the benchmark is unsupervised training on pseudolabels, generated by the clustering of the corresponding SIFT descriptors. In Table 6, the results for both finetuning strategies are shown. For a fair comparison, we also

train ResNet56 and the ViT architecture ViT-4/8 from scratch for both settings to show the influence of SAGHOG.

Table 6: Finetuning NetRVLAD with different backbones and target labels: Supervised (*Sup.*) and Cl-S [6]. We report mAP in %.

	Historical-WI		HisFrag20	
	Sup.	Cl-S	Sup.	Cl-S
ResNet56	**67.2**	**73.3**	42.5	41.2
ViT-4/8 (from scratch)	60.0	71.1	43.4	47.4
SAGHOG (ours)				
frozen backbone	64.7	64.0	**56.3**	**57.2**
finetuned	61.4	**73.3**	50.7	54.6
Δ to best	-2.5	0.0	+12.9	+9.8

Historical-WI In contrast to [6], we have a surprisingly good baseline when training ResNet56 with NetRVLAD on writer labels (67.2 % mAP). This setting even outperforms SAGHOG with supervised finetuning, indicating that the pretraining task is contrary to the finetuning - freezing the backbone also improves the mAP. For unsupervised training, SAGHOG is on par with ResNet56 and outperforming training ViT from scratch. We think this is mainly due to the dataset only containing pure handwriting, where a simple CNN is enough to achieve good performance. The approach seems to be limited by the finetuning task (the pseudolabels generated by Cl-S).

HisFrag20 However, the application of SAGHOG on HisFrag20, a more complex and larger dataset, drastically shifts the results towards SAGHOG. Firstly, training ViT on Cl-S, even from scratch, outperforms the current state of the art. Furthermore, ViTs outperforms the residual baseline in all experiments. Our best result is achieved by freezing the encoder of SAGHOG and only training NetRVLAD, achieving 56.3 % for supervised and 57.2 % for unsupervised training, which is a margin of 12.9 %/9.8 % to the best setting when training from scratch. In conclusion, we argue that SAGHOG works best for large, complex datasets, where networks trained from scratch fail to perform due to noisy data - either on image- or label-level - or due to a domain shift between training and test data.

5.3 Comparison to state of the art

We compare the results of SAGHOG to state of the art. In addition to the previous evaluation, we extend our analysis to include a comparison of the GRK-Papyri dataset. All results are shown in Table 7.

Table 7: Comparison to state of the art. (†) denotes our own implementation, subscript $_i$ the dimension of the final page descriptor.

(a) Historical-WI

Method		mAP	Top-1
Cl-S + ResNet56	+ NetRVLAD$_{512}$ [29]	73.4	88.5
Cl-S + ResNet20	+ mVLAD$_{400}$ [6]	73.2	87.6
Cl-S + ResNet20	+ mVLAD$_{32000}$ [6]	74.8	88.6
Pathlet/SIFT	+ bVLAD$_{1000}$ [21]	**77.1**	**90.1**
Cl-S + SAGHOG	+ NetRVLAD$_{512}$	73.3	88.2
with SGR$^{\dagger}_{k=2}$ [29]		**80.3**	**91.1**

(b) GRK-Papyri

Method		mAP	Top-1
Cl-S + ResNet20	+ mVLAD$_{32000}$ [8]	42.2	52.0
R-SIFT	+ mVLAD$_{32000}$ [8]	**42.8**	48.0
Cl-S + ResNet20	+ NetRVLAD$^{\dagger}_{2048}$ [29]	38.8	56.0
Cl-S + SAGHOG	+ NetRVLAD$_{2048}$	38.2	**58.0**
with SGR$^{\dagger}_{k=3}$ [29]		**47.6**	**58.0**

(c) HisFrag20

Method	Writer Retrieval mAP	Writer Retrieval Top-1	Page Retrieval mAP	Page Retrieval Top-1
Cl-S + mVLAD$_{3200}$ + ESVM [2]	33.7	68.9	-	-
FragNet$_{512}$ [32]	33.5	77.1	22.6	36.4
ResNet50 + NetVLAD$_{6400}$ + ESVM [3]	38.1	77.9	-	-
Cl-S + ResNet56 + NetRVLAD$^{\dagger}_{512}$ [29]	41.2	68.5	23.5	33.3
Feature Mixer$_{256}$ [30]	44.0	81.9	29.3	45.0
A-VLAD$_{500}$ [26]	46.6	**85.2**	-	-
Cl-S + SAGHOG + NetRVLAD$_{512}$	**57.2**	81.7	**35.2**	**47.6**
with SGR$^{\dagger}_{k=2}$ [29]	**68.7**	**88.7**	**54.5**	**70.6**

Historical-WI As discussed in the previous section, self-supervised pretraining has a limited effect on the Historical-WI dataset - probably due to the dataset containing only (binarized) handwriting. However, SAGHOG is on par with all deep-learning-based approaches (Table 7a) and does not harm the performance. Interestingly, the best approach in terms of mAP (without considering reranking) is still relying on handcrafted features [21].

GRK-Papyri We also finetune SAGHOG on GRK-Papyri in an unsupervised manner, a small dataset of only 50 documents. The results are shown in Table 7b. The global page descriptors used are averaged on five different models trained, similar to the ensemble method used by Christlein et al. [8]. We also choose to decrease the dimension of SAGHOG to 128 instead of 512 for a fair comparison. While NetVLAD is trailing the traditional VLAD on mAP for both networks, ResNet20 and SAGHOG, the end-to-end learned method achieves better Top-1 accuracy, with SAGHOG outperforming related work with 58.0 %.

HisFrag20 On the largest and most diverse dataset, HisFrag20, we show that SAGHOG is superior to all previous approaches by a large margin on both tasks - WR (+10.6 %) and page retrieval (+5.9 %). A thorough comparison to related work is given in Table 7c. Although the mAP and therefore the ranked list is better, SAGHOG has a slightly worse Top-1 accuracy. As we showed in our work introducing the feature mixer [30], networks trained on full-color images tend to infer features from the background during training, which is also shown in our comparison: For example, approaches relying on 32×32 patches such as [2,29] suffer from a limited Top-1 accuracy. We think this is due to approaches on image-level retrieving images from the same page, where the focus is not on retrieving documents based on the similarity of the handwriting, but rather on background similarity, which is shared within a page. This phenomenon is also observed by considering the page retrieval performance (e.g., [29] vs. [30]). However, we argue that due to SAGHOG achieving a superior performance in terms of mAP, we can apply a reranking scheme such as SGR which is shown to be beneficial in particular for Top-1 accuracy [29]. By using SGR, we are able to boost both metrics by 11.5 % mAP / 7 % Top-1 for WR and 19.3 % mAP / 23 % Top-1 for page retrieval, which outperforms related work, even on the Top-1 accuracy. We also provide qualitative results in Fig. 5.

6 Conclusion

In this paper, we introduce SAGHOG, the first masking-based self-supervised approach for the downstream task of WR. Our pretraining includes reconstructing HOG features of the binarized image, which leads the network to focus on handwriting and suppress the background. Additionally, we propose a preprocessing via SAM to make use of previous historical documents for our pretraining. We evaluate three different datasets and two different finetuning methods commonly used in the domain of WR. In the following, we summarize our key results and ideas for future work.

Our findings Firstly, we investigate our preprocessing scheme and show that the segmentation of handwriting is beneficial for the pretraining of SAGHOG. Secondly, SAGHOG shines on WR for large, complex datasets containing sources of noise (e.g. HisFrag20 with small, torn fragments). We are able to outperform related work by a significant margin on this difficult dataset. Furthermore, we

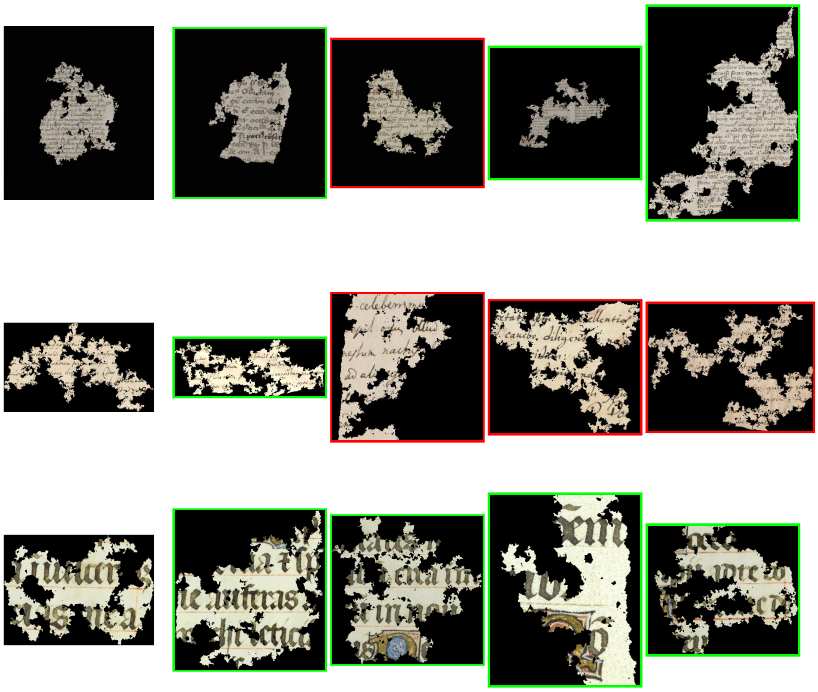


Fig. 5: Qualitative results of the retrieval on HisFrag20. Left: Query. We show the four nearest documents. If the retrieved document is written by the same author as the query, we highlight the image in green, otherwise red.

show that for a binarized dataset such as Historical-WI the effect of pretraining is limited, for which current methods are already working decent. In the end, we also evaluate on GRK-Papyri, a small dataset, and show that the application of SAGHOG improves the Top-1 accuracy upon state of the art. SAGHOG also shows the applicability of the ViT architecture for WR, which heavily depends on CNNs.

Future work We further want to investigate self-supervised learning for. On the one hand, one could improve the diversity of data for pretraining (we mainly use Latin script and documents including a limited amount of degradation) to improve results also on smaller datasets such as GRK-Papyri. On the other hand, tuning the reconstruction task, e.g., using a U-Net for improved binarization or integrating style embeddings from specific domains to refine the self-supervised embeddings, promises to further improve the results.

Acknowledgements We thank Vincent Christlein for providing the binarized images of GRK-Papyri.

References

1. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 9630–9640 (2021)
2. Chammas, M., Makhoul, A., Demerjian, J.: Writer identification for historical handwritten documents using a single feature extraction method. In: 19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020, Miami, FL, USA, December 14-17, 2020. pp. 1–6 (2020)
3. 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 (2022)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607 (2020)
5. Christlein, V., Bernecker, D., Angelopoulou, E.: Writer identification using VLAD encoded contour-zernike moments. In: 13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015. pp. 906–910 (2015)
6. Christlein, V., Gropp, M., Fiel, S., Maier, A.K.: Unsupervised feature learning for writer identification and writer retrieval. In: 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017. pp. 991–997 (2017)
7. Christlein, V., Maier, A.K.: Encoding CNN activations for writer recognition. In: 13th IAPR International Workshop on Document Analysis Systems, DAS 2018, Vienna, Austria, April 24-27, 2018. pp. 169–174 (2018)
8. Christlein, V., Marthot-Santaniello, I., Mayr, M., Nicolaou, A., Seuret, M.: Writer retrieval and writer identification in greek papyri. In: Intertwining Graphonomics with Human Movements - 20th International Conference of the International Graphonomics Society, IGS 2021, Las Palmas de Gran Canaria, Spain, June 7-9, 2022, Proceedings. vol. 13424, pp. 76–89 (2022)
9. Christlein, V., Nicolaou, A., Seuret, M., Stutzmann, D., Maier, A.: ICDAR 2019 competition on image retrieval for historical handwritten documents. In: 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019. pp. 1505–1509 (2019)
10. Cloppet, F., Eglin, V., Helias-Baron, M., Kieu, V.C., Vincent, N., Stutzmann, D.: ICDAR2017 competition on the classification of medieval handwritings in latin script. In: 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017. pp. 1371–1376 (2017)
11. Cloppet, F., Eglin, V., Kieu, V.C., Stutzmann, D., Vincent, N.: ICFHR2016 competition on the classification of medieval handwritings in latin script. In: 15th International Conference on Frontiers in Handwriting Recognition, ICFHR 2016, Shenzhen, China, October 23-26, 2016. pp. 590–595 (2016)
12. Diem, M., Kleber, F., Sablatnig, R., Gatos, B.: cBAD: ICDAR2019 competition on baseline detection. In: 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019. pp. 1494–1498 (2019)

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. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 (2021)
14. Fiel, S., Kleber, F., Diem, M., Christlein, V., Louloudis, G., Nikos, S., Gatos, B.: ICDAR2017 competition on historical document writer identification (historical-wi). In: 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017. pp. 1377–1382 (2017)
15. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 15979–15988. IEEE (2022)
16. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 9726–9735. Computer Vision Foundation / IEEE (2020)
17. Jégou, H., Douze, M., Schmid, C.: On the burstiness of visual elements. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. pp. 1169–1176 (2009)
18. Keglevic, M., Fiel, S., Sablatnig, R.: Learning features for writer retrieval and identification using triplet cnns. In: 16th International Conference on Frontiers in Handwriting Recognition, ICFHR 2018, Niagara Falls, NY, USA, August 5-8, 2018. pp. 211–216 (2018)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
20. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (2023)
21. Lai, S., Zhu, Y., Jin, L.: Encoding pathlet and SIFT features with bagged VLAD for historical writer identification. IEEE Trans. Inf. Forensics Secur. **15**, 3553–3566 (2020)
22. Lastilla, L., Ammirati, S., Firmani, D., Komodakis, N., Merialdo, P., Scardapane, S.: Self-supervised learning for medieval handwriting identification: A case study from the vatican apostolic library. Information Processing & Management **59**(3), 102875 (2022)
23. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019 (2019)
25. Mohammed, H.A., 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 2019, Sydney, Australia, September 20-25, 2019. pp. 726–731 (2019)

26. Ngo, T.T., Nguyen, H.T., Nakagawa, M.: A-VLAD: an end-to-end attention-based neural network for writer identification in historical documents. In: 16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II. vol. 12822, pp. 396–409 (2021)
27. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (Jan 1979)
28. Peer, M., Kleber, F., Sablatnig, R.: Self-supervised vision transformers with data augmentation strategies using morphological operations for writer retrieval. In: Frontiers in Handwriting Recognition - 18th International Conference, ICFHR 2022, Hyderabad, India, December 4-7, 2022, Proceedings. pp. 122–136 (2022)
29. Peer, M., Kleber, F., Sablatnig, R.: Towards writer retrieval for historical datasets. In: Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part I. pp. 411–427 (2023)
30. 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 (2023)
31. Sauvola, J., Pietikäinen, M.: Adaptive document image binarization. Pattern Recognition **33**(2), 225–236 (Feb 2000)
32. Seuret, M., Nicolaou, A., Maier, A., Christlein, V., Stutzmann, D.: ICFHR 2020 competition on image retrieval for historical handwritten fragments. In: 17th International Conference on Frontiers in Handwriting Recognition, ICFHR 2020, Dortmund, Germany, September 8-10, 2020. pp. 216–221 (2020)
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 (Jun 2010)
34. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 5022–5030 (2019)
35. Wang, Z., Maier, A., Christlein, V.: Towards end-to-end deep learning-based writer identification. In: 50. Jahrestagung der Gesellschaft für Informatik, INFORMATIK 2020 - Back to the Future, Karlsruhe, Germany, 28. September - 2. Oktober 2020. vol. P-307, pp. 1345–1354 (2020)
36. Wei, C., Fan, H., Xie, S., Wu, C., Yuille, A.L., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 14648–14658 (2022)
37. Zenk, J., Kordon, F., Mayr, M., Seuret, M., Christlein, V.: Investigations on self-supervised learning for script-, font-type, and location classification on historical documents. In: Proceedings of the 7th International Workshop on Historical Document Imaging and Processing. HIP '23 (Aug 2023)