CLEAR-WSI: Foundation Model Empowered Whole Slide Image Retrieval

Youssef Wally[1,2] YOUSSEF.M.WALLY@UIT.NO
Jingsong Liu[1,3,4] JINGSONG.LIU@TUM.DE
Elisabeth Wetzer[2] ELISABETH.WETZER@UIT.NO
Peter Schüffler[1,3,4,5] PETER.SCHUEFFLER@TUM.DE

[1] *Institute of Pathology, Technical University of Munich, Munich, Germany*
[2] *Department of Physics and Technology, UiT The Arctic University of Norway, Tromsø, Norway*
[3] *Munich Center for Machine Learning (MCML), Munich, Germany*
[4] *Munich Data Science Institute (MDSI), Munich, Germany*
[5] *German Cancer Consortium (DKTK), partner site Technical University of Munich, Germany*

Editors: Under Review for MIDL 2026

Abstract

The rapid growth of digital pathology has produced vast repositories of hematoxylin and eosin stained whole slide images, yet most of them remain unindexed or unlabelled, limiting their utility for computational analysis. Reverse image search provides a scalable way to organize and access these archives by retrieving visually similar images. While currently deployed retrieval systems exist, they rely on manual configuration, highly affecting their performance. Thus, we propose CLEAR-WSI, Constant Length Embedding & Automatic Retrieval, a fully automated pathology reverse image search engine that leverages Vision Transformer foundation models for histopathology together with attention-based multiple instance learning (AttentionMIL). The AttentionMIL framework jointly identifies diagnostically relevant whole slide images and predicts slide-level diagnoses. To further improve performance, we introduce a self-reviewing classifier filtering mechanism: retrieved candidates are filtered according to their predicted labels, mostly outperforming class-informed filters. Across two public datasets, CAMELYON16 (lymph node metastases) and BRACS (breast cancer subtypes), our method establishes new state-of-the-art results, improving Acc_{MV}@5 from 77.49% to 89.92% on CAMELYON16, from 54.12% to 75.86% on BRACS level-1, and from 36.47% to 51.72% on BRACS level-2. Our general-purpose, annotation-free, dataset-agnostic, search engine that scales across diverse data sources is openly available: https://github.com/youssefwally/CLEAR-WSI

Keywords: Computational Pathology, Image Retrieval, Reverse Image Search, Foundation Models, Vision Transformers, Whole Slide Images, Deep Learning.

1. Introduction

Millions of Whole Slide Images (WSI) remain unlabelled and unindexed, limiting their utility for clinical decision support and large-scale research. Retrieval systems that enable content-based image retrieval (CBIR), identifying diagnostically similar cases without requiring prior diagnostic knowledge, offer a practical way to leverage this untapped data, particularly in the case of rare diseases and situations where diagnostic confidence is low (Böttcher et al., 2025).

However, most CBIR systems rely on expert annotations and carefully curated datasets, which is difficult to scale, especially for infrequent conditions. These challenges have motivated the adoption of transfer learning and Foundation Models (FM), which have demonstrated strong performance in mutation inference (Kather et al., 2020; Fu et al., 2020), cancer grading (Bulten et al., 2022), and survival forecasting (Chen et al., 2022b; Foersch et al., 2023). However, the potential of FM for slide-level retrieval remains underexplored (Tizhoosh and Pantanowitz, 2024; Lahr et al., 2024).

Yottixel represents the first dedicated WSI-level search engine (Kalra et al., 2020) and nonetheless the current state-of-the-art (SOTA) (Lahr et al., 2024). Its Bunch of Barcodes (BoB) representation enables efficient large-scale similarity search, but depends on manual configuration of cluster numbers and patch-sampling rates. Its embeddings scale with patch count, producing inconsistent representation sizes, which hinder generalizability and automation in real-world deployments.

We address these limitations and further propose a self-reviewing (SR) filter that stabilizes retrieval and consistently outperforms class-informed (CI) filtering strategies by introducing CLEAR-WSI, Constant-Length Embedding & Automatic Retrieval, shown in figure 1. Evaluation across two datasets shows that our method establishes a new SOTA for WSI retrieval. We employ Normalized Discounted Cumulative Gain (NDCG), a standard metric in information retrieval, to assess ranking quality in this domain, following (Shi et al., 2018).

In summary, our main contributions are:

- We introduce CLEAR-WSI, a fully automated patch-count independent whole-slide-image context-based image search engine.
- Consistent, substantial performance improvements over two differing datasets compared to the current SOTA Yottixel.
- Release of the full implementation for reproducibility and community use.

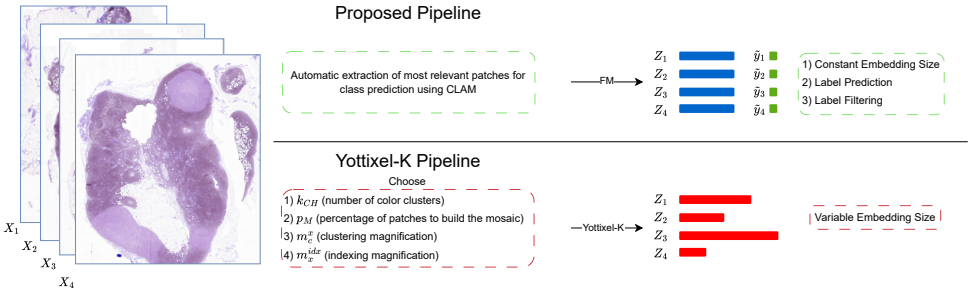

Figure 1: CLEAR-WSI vs. Yottixel (SOTA) (Kalra et al., 2020). Details in Section 3.

2. Background

Traditionally, large-scale medical image archives relied on textual annotations for search and retrieval, which often fails to capture visual similarities effectively. To address this, one of the earliest online CBIR by (Zheng et al., 2003) allowed users to upload a query image with search parameters which then performs similarity comparisons based on features such as colour histograms, texture, Fourier coefficients, and wavelet descriptors.

WSI in digital pathology exceed conventional image-processing pipelines in both scale and diagnostic context (Barker et al., 2016; Gutman et al., 2013). CBIR for histopathology,

addresses the need to find visually and diagnostically similar regions or slides within large archives without relying solely on textual metadata. Early and recent CBIR systems share the core pipeline of (1) decomposing WSI into manageable regions or patches, (2) extracting discriminative descriptors from those regions, (3) aggregating region descriptors into slide-level indices, and (4) performing nearest-neighbour retrieval in the chosen embedding/index space (Tizhoosh and Pantanowitz, 2024; Lahr et al., 2024).

Two practical strategies for handling WSI dominate the literature. The *sub-setting* approach selects a smaller region of interest from the full slide to reduce computation and focus search *(region-centric)*. The *tiling* approach divides the slide into many patches and either indexes patches directly or aggregates their representations into slide-level descriptors *(patch-centric)* (Mehta et al., 2009; Lowe, 1999; Galaro et al., 2011; Sharma et al., 2012; Vanegas et al., 2014). Sub-setting is efficient but requires domain knowledge to select representative regions; tiling is more automatable but raises design choices about sampling, aggregation, and index size (Barker et al., 2016; Gutman et al., 2013). Systems demonstrate these design tradeoffs in practice (Tizhoosh and Pantanowitz, 2024; Lahr et al., 2024).

Recent CBIR systems in pathology have focused on patch-level or region-level similarity search (Barker et al., 2016; Gutman et al., 2013), using handcrafted features (Zheng et al., 2003), CNN-based embeddings (Tommasino et al., 2023), or attention-based aggregation (Li et al., 2023). Early systems leveraged texture and colour descriptors (Zheng et al., 2003), while more recent approaches employ deep convolutional embeddings (Tommasino et al., 2023), deep hashing (Li et al., 2023), attention-driven multiple-instance learning (Lu et al., 2021), and graph-based architectures to capture tissue structure and contextual relationships (Tizhoosh and Pantanowitz, 2024; Lahr et al., 2024). Multimodal and large-scale frameworks have also emerged, ranging from early multimodal efforts to recent multimodal foundation systems such as TITAN (Ding et al., 2025). Several works include deep hashing and deep-learning-based retrieval studies (Chen et al., 2022a; Li et al., 2023; Hegde et al., 2019; Tommasino et al., 2023; Shi et al., 2018) and large-scale WSI search pipelines (Wang et al., 2023). Despite this progress, CBIR methods remain limited in scalability because they depend on patch-level embeddings, require careful parameter tuning, and struggle to produce slide-level representations that generalize across datasets (Tizhoosh and Pantanowitz, 2024; Lahr et al., 2024).

Yottixel (SOTA) (Kalra et al., 2020), a slide-level indexing strategy inspired by bag-of-words concepts, segments a slide into distinct regions (via clustering), samples patches to form a mosaic, extracts features with a pretrained network, and converts them into compact binary barcodes to build a BoB index. Yottixel showed strong retrieval speed and scalability but depends on manual configuration (cluster counts, sampling ratios), and its index size varies with patch sampling, producing inconsistent representation sizes across slides.

Self-supervised methods (Wang et al., 2023; Srinidhi et al., 2022; Chen et al., 2022a; Ciga et al., 2022; Li et al., 2023) explored scalable search by using self-supervised representation learning to produce compact, transferable descriptors suitable for slide-level retrieval and rare-case retrieval. These foundation-style visual models improve generalization across tasks but may still require careful preprocessing (colour normalization, tissue detection) and large-scale compute for pretraining.

Other Approaches (Ding et al., 2025) extend retrieval capability by aligning image and text modalities or pretraining on large WSI corpora. These models enable cross-modal

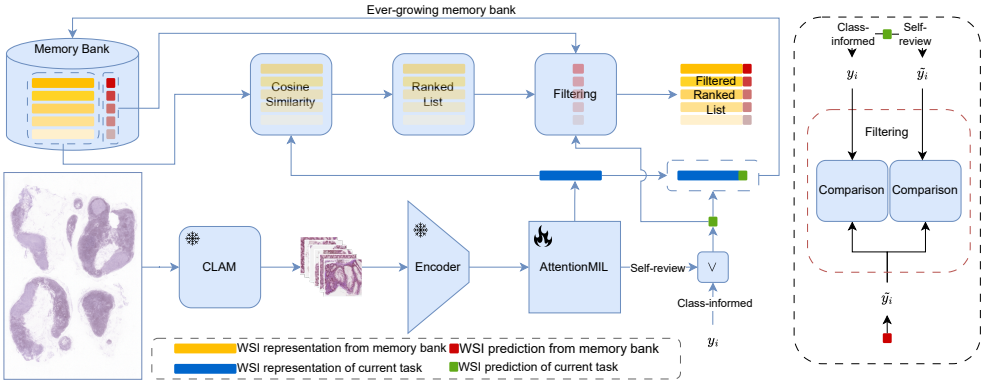

Figure 2: Our pipeline, CLEAR-WSI, using CLAM (Lu et al., 2021). Details in Section 3.

retrieval and zero/few-shot transfer, and they improve robustness to distribution shifts; however, they introduce new dependencies such as large pretraining datasets, compute, and careful evaluation of clinical generalization.

Existing CBIR approaches for WSI expose recurring limitations relevant to deployment and automation (Tizhoosh and Pantanowitz, 2024; Lahr et al., 2024):

- **Manual hyperparameters and sampling bias:** Methods that rely on clustering + sampling (mosaics) require manual choices (number of clusters, sampling fraction) that affect coverage and reproducibility.
- **Inconsistent representation sizes:** Index sizes that scale with the number of selected patches or mosaics produce variable memory and compute requirements.
- **Limited context:** Patch-level retrieval systems may miss slide-level context important for diagnostics due to simple aggregation; losing spatial or structural information.
- **Scalability vs. accuracy:** Compact/hashed indices and binarization increase speed but can degrade fine-grained retrieval necessary for rare or subtle morphologies.

Thus, we propose CLEAR-WSI (Constant-Length Embedding & Automatic Retrieval), a fully automated WSI CBIR that (1) removes dependence on manual mosaic construction and patch-count–dependent indices, (2) leverages Vision Transformer (ViT)–based foundation encoders for robust patch descriptors, and (3) aggregates patch descriptors into dimensionally-consistent, compact slide-level representations suitable for scalable retrieval without manual configuration. The method aims to preserve diagnostic context while addressing the limitations above through automated preprocessing, unified embedding dimensionality, and retrieval-optimized indexing.

3. Methodology

CLEAR-WSI is designed to identify and rank WSI or their patches based on similarity to a given image query as illustrated in Figure 2. The process consists of (1) patch selection using CLAM (Lu et al., 2021), (2) feature extraction with a frozen FM encoder, (3) aggregation of patch embeddings through a trained AttentionMIL model, (4) similarity-based ranking, (5) whole slide label prediction with the same AttentionMIL model, (6) label consistency filtering, and (7) storage of the resulting outputs in a growing memory bank for future use.

This modular design enables the framework to operate at both the slide and patch levels, depending on whether patch selection and aggregation are enabled.

3.1. Patch Selection using CLAM

CLAM obtains patch-level feature embeddings by passing each WSI patch through a pre-trained ResNet-50 encoder, then computes an attention score for each patch using a class-specific attention network; these embeddings and scores are optimized while training on slide-level diagnostic classification tasks so that the features capture morphology relevant to the slide label, and the highest-attention patches correspond to regions most informative for predicting that diagnosis.

CLAM's feature vectors are optimized for weakly supervised slide-level classification, not for producing stable, semantically aligned patch embeddings. The representation is shaped by attention-based multiple instance learning (MIL) and instance-level clustering, so it becomes highly class specific, emphasizing features useful only for the diagnostic labels on which CLAM was trained. For retrieval, this is undesirable: class conditioned embeddings distort cross-class similarity structure and do not generalize further. Returning to the raw patch and re-embedding it with an FM encoder such as UNI (Chen et al., 2024) avoids these constraints because FM are trained to produce uniform, semantically consistent, pathology-domain embeddings that preserve morphological similarity independent of any specific downstream label.

Direct end-to-end processing of a WSI is computationally infeasible. Thus, we apply CLAM to select a subset of M patches, such that each WSI is represented by a set of patches $\tilde{X}_l$:

$$X_l = \{p_1, p_2, \ldots, p_M\}, \quad p_j \in \mathbb{R}^{h \times w \times c}.$$

CLAM assigns attention scores α_j to each patch and selects the top-ranked patches:

$$\alpha_j = \frac{\exp(w^\top z_j^{clam})}{\sum_{k=1}^{M} \exp(w^\top z_k^{clam})}, \quad \tilde{X}_l = \{p_j : \alpha_j \geq \tau\}.$$

Here, z_j^{clam} denotes the feature embedding of patch p_j, and τ is a threshold. This produces a compact yet informative subset of patches for downstream retrieval.

3.2. Feature Extraction and Aggregation

For each WSI X_l, selected patches $p_j \in \tilde{X}_l$ are embedded by a ViT–based FM encoder f_θ:

$$z_j^{FM} = f_\theta(p_j), \quad z_j \in \mathbb{R}^d.$$

where f_θ is a deep feature extractor parameterized by θ that maps each image patch into a latent representation. To capture the varying relevance of patch embeddings, a attention-based multi-instance learning (AttentionMIL) model is trained to weigh patches according to their contribution to the WSI label. Afterwards, the attention scores β_j from the AttentionMIL are utilised to form a fixed-length global representation Z_l for each WSI X_l such that

$$Z_l = \sum_{j=1}^{|\tilde{X}_l|} \beta_j f_\theta(p_j), \quad p_j \in \tilde{X}_l$$

3.3. Ranking by Similarity

Given a query embedding Z_q and a database $\mathcal{D} = \{Z_1, \ldots, Z_N\}$, similarity is measured using cosine similarity. The ranking is defined as:

$$\pi(q) = \text{argsort}_{i=1..N} \, \text{CosineSim}(Z_q, Z_i).$$

3.4. Label Consistency Filtering

To improve retrieval robustness, we apply post-retrieval filtering strategies that enforce label consistency. Specifically, we evaluate two schemes:
1.) SR Classifier Filtering: A retrieved item i is retained if and only if its predicted label $\tilde{y}_i$ matches the query label y_q:

$$\pi_{\text{SR}}(q) = \{i \in \pi(q) \mid \tilde{y}_i = \tilde{y}_q\}.$$

2.) CI Filtering: Ground-truth labels are used to filter the ranked list:

$$\pi_{\text{CI}}(q) = \{i \in \pi(q) \mid \tilde{y}_i = y_q\}.$$

Each embedding and prediction learned or processed for a query image are also added back into the memory bank for future comparisons.

Lastly, the top-k ranked items from $\tilde{\pi}(q)$ form the retrieval candidate set $\tilde{\pi}_k(q)$, where $\tilde{\pi}(q)$ is either $\pi(q)$, $\pi_{SR}(q)$, $\pi_{CI}(q)$ depending on the chosen filtering method.

3.5. Evaluation Metrics

We assess retrieval performance by evaluating how effectively the pipeline ranks relevant WSIs. A WSI is considered relevant if its label matches the label of the query image.

3.5.1. Majority-Vote Accuracy

We report majority-vote accuracy ($Acc_{MV}@k$) at Top-1, Top-3, and Top-5 such that:

$$Acc_{MV}@k = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left\{ y_q = \text{mode}(\{y_j : j \in \tilde{\pi}_k(q)\}) \right\}.$$

3.5.2. Normalized Discounted Cumulative Gain (NDCG)

In addition to $Acc_{MV}@k$, we use NDCG (Järvelin and Kekäläinen, 2002) as an evaluation metric in the context of WSI CBIR. NDCG is widely adapted in recommender system evaluation (Mao et al., 2021; Bellogín et al., 2011; Xue et al., 2017; He et al., 2020; Rendle et al., 2009; He et al., 2017) however, it is rarely adapted in WSI CBIR evaluation. Unlike accuracy-based measures, NDCG explicitly accounts for the ranking position of relevant

results, rewarding models that place correct matches earlier in the retrieval list. Thus, we follow (Shi et al., 2018) and report NDCG evaluation metrics.

Formally, NDCG at cutoff k is defined as:

$$NDCG@k = \frac{1}{\lambda} \sum_{i=1}^{k} \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)}, \quad rel_i = \begin{cases} 1 & \text{if } \tilde{y}_i = y_i, \\ 0 & \text{otherwise.} \end{cases}$$

where rel_i is the binary relevance score of the i-th ranked item, and λ is a normalization constant ensuring $NDCG@k \in [0, 1]$. The logarithmic discounting penalizes relevant results that appear lower in the ranking, thereby providing a more nuanced measure of retrieval quality compared to accuracy-based metrics alone.

4. Datasets

This study leverages two publicly available Hematoxylin and Eosin (H&E)-stained datasets widely used in contemporary computational pathology. The datasets differ in their clinical diagnosis, scale, and labelling schemes, thereby providing complementary challenges for evaluating our retrieval framework.

4.1. CAMELYON16

The CAMELYON16 dataset (Bejnordi et al., 2017) is among the earliest benchmarks for automated metastasis detection in lymph node tissue. It provides large WSI from sentinel lymph node biopsies and supports binary classification. It contains 399 WSI, comprising 160 tumours and 239 normal cases with varying resolutions.

4.2. BRACS

The BReAst Carcinoma Subtyping (BRACS) dataset (Brancati et al., 2022) contains H&E WSI of breast carcinoma subtyping. Unlike CAMELYON16, BRACS is designed for multi-class classification with a hierarchical label structure. At level-1, slides are categorized into three groups; at level-2, level-1 classes are refined further into subclasses; with a total of seven subclasses in level-2. The BRACS dataset comprises 547 whole-slide images categorized into 265 benign tumors (44 normal, 147 pathological benign, and 74 usual ductal hyperplasia), 89 atypical tumors (41 flat epithelial atypia and 48 atypical ductal hyperplasia), and 193 malignant tumors (61 ductal carcinoma in situ and 132 invasive carcinoma). All slides are scanned at 40× magnification and include two-level hierarchical labels.

5. Experimental Setup

CLEAR-WSI consists of four stages: *feature extraction*, *training*, *memory bank construction*, and *evaluation*. Dataset-splits were done according to the respective dataset release.

We utilise 4 different ViT-based FM for feature extraction; DeiT (Touvron et al., 2021), MoCov3 (Chen et al., 2021), Prov-GigaPath (Xu et al., 2024) and UNI (Chen et al., 2024). DeiT and MoCov3 are trained on general images, with DeiT using a distillation-based objective and MoCov3 using a self-supervised architecture on ImageNet-1K, providing efficient general-purpose visual representations but limited transfer to histopathology. Thus,

we finetune both on CAMELYON16 (Bejnordi et al., 2017) and BRACS (Brancati et al., 2022). In contrast, Prov-GigaPath and UNI are pathology FMs. Prov-GigaPath is trained on 1.3 billion tiles from more than 170,000 WSI. UNI is pretrained on more than 100 million tissue patches from over 100,000 diagnostic WSI. Therefore, we do not finetune them. These models differ primarily in domain specificity and scale: DeiT and MoCov3 capture general visual features, whereas Prov-GigaPath and UNI provide domain-specialized representations optimized for computational pathology.

In the *training* stage, the AttentionMIL model is trained to aggregate patch-level embeddings into slide-level representations while simultaneously predicting slide-level labels. Once trained, the pipeline is applied to the validation set in order to *construct a memory bank*: for every slide, we store its embedding, the predicted label obtained from the trained AttentionMIL or the corresponding ground-truth label, according to which filtering mechanism that will be tested. This memory bank serves as the retrieval database against which all subsequent queries are compared. During *evaluation*, test slides are processed through the pipeline. Depending on the chosen filtering strategy, SR or CI, the ranked retrieval results are refined to enforce label consistency. In the case of CI filtering, we assume that the label of the query WSI is known, thus filtering with the ground truth label. Note that all WSI in the memory bank only have predicted labels. While in the case of SR filtering, we assume that the user does not know the label of the query WSI, thus using a predicted label for the query WSI from the AttentionMIL. **The ground truth labels of the test set are only used when evaluating the retrieved WSI.**

Beyond its static role during evaluation, the memory bank is designed as an ever-growing repository. Each query processed through the retrieval pipeline is not only compared against the existing entries but also added back into the memory bank together with its embedding and predicted label. This incremental update mechanism allows the memory bank to continuously expand as more queries are made, effectively transforming it into a dynamic knowledge base. Over time, the system benefits from a richer and more diverse set of stored representations, which can improve retrieval performance and enhance the adaptability of the framework to evolving data distributions.

6. Results

We assess retrieval performance on CAMELYON16 (binary) and BRACS (multi-class with level-1 and level-2 labels) using Majority-Vote Accuracy (AccMV@k) and NDCG@k for $k \in \{1, 3, 5\}$. A fixed memory bank is constructed from the validation split, and all test queries are ranked against this bank. We report CI only for UNI as it is the best performing model. However, other models consistently followed the same pattern as UNI.

Across datasets, CLEAR-WSI using a frozen UNI foundation encoder, AttentionMIL aggregation, cosine similarity ranking, and CLEAR-WSI built on SR+UNI filtering, consistently surpasses Yottixel (Table 1). On CAMELYON16, CLEAR-WSI built onSR+UNI achieves AccMV@5 of 89.92% and NDCG@5 of 89.07%, surpassing Yottixel's 77.49% and 73.69%. The performance gains extend to early ranks as well; with Top-1 accuracy and NDCG@k of 89.92% and 90.02% versus 75.96% and 76.21% for Yottixel. On BRACS level-1, CLEAR-WSI built on SR+UNI improves SOTA at AccMV@5 from 54.12% to 75.86% and

Table 1: WSI retrieval performance: $\text{Acc}_{MV}(\Uparrow)$ and NDCG($\Uparrow$) @$k = 1, 3, 5$; *finetuned.

	Pipeline	CAMELYON16 NDCG|Acc_{MV}	BRACS-1 NDCG|Acc_{MV}	BRACS-2 NDCG|Acc_{MV}
@k=1	Yottixel-K (SOTA)	76.21|75.96	49.91|49.41	33.61|32.94
	CLEAR-WSI (SR DeiT*)	69.30|68.99	52.20|51.72	38.93|37.93
	CLEAR-WSI (SR MoCov3*)	67.00|66.67	54.48|54.02	28.30|27.59
	CLEAR-WSI (SR Prov-GigaPath)	73.14|72.87	51.06|50.57	30.58|29.89
	CLEAR-WSI (CI UNI)	**96.93|96.90**	43.87|43.68	10.53|10.34
	CLEAR-WSI (SR UNI)	90.02|89.92	**67.00|66.67**	**44.24|43.68**
@k=3	Yottixel-K (SOTA)	74.59|78.29	49.58|54.12	31.25|36.47
	CLEAR-WSI (SR DeiT*)	65.45|67.44	51.47|52.87	32.56|34.48
	CLEAR-WSI (SR MoCov3*)	67.10|72.09	54.45|50.47	28.81|25.28
	CLEAR-WSI (SR Prov-GigaPath)	67.99|71.32	51.20|49.43	33.18|35.63
	CLEAR-WSI (CI UNI)	**97.69|99.22**	42.02|45.98	11.24|16.09
	CLEAR-WSI (SR UNI)	89.56|89.15	**66.82|72.41**	**44.94|48.28**
@k=5	Yottixel-K (SOTA)	73.69|77.49	48.64|54.12	29.17|36.47
	CLEAR-WSI (SR DeiT*)	65.84|70.54	51.44|57.47	31.48|36.78
	CLEAR-WSI (SR MoCov3*)	67.07|72.87	55.30|62.07	28.99|29.89
	CLEAR-WSI (SR Prov-GigaPath)	65.99|65.89	52.18|54.02	32.42|36.78
	CLEAR-WSI (CI UNI)	**97.23|98.93**	41.39|50.57	11.46|14.94
	CLEAR-WSI (SR UNI)	89.07|89.92	**66.98|75.86**	**43.19|51.72**

NDCG@5 from 48.64% to 66.98%, while on the more challenging BRACS level-2, AccMV@5 rises from 36.47% to 51.72% with NDCG@5 from 29.17% to 43.19%.

These improvements are consistent across different k and demonstrate superior rank quality as captured by NDCG. Variants using DeiT, MoCoV3, and Prov-GigaPath encoders are sometimes competitive, particularly on BRACS, but generally UNI outperfomes them, highlighting the importance of domain-tuned foundation features for slide level retrieval. CI filtering, which assumes oracle access to the query's ground truth label, yields near-ceiling measures on CAMELYON16, validating the quality of the learned representations and aggregation; however, on BRACS it does not match SR due to label prediction noise. Together, the results show that the proposed pipeline is both accurate and rank-sensitive, retrieving correct diagnoses and surfacing them early in the list.

7. Discussion

The improvements stem from three complementary design choices: First, patch count independent slide representations produced via ViT-based foundation encoders and Attention-MIL aggregation; avoiding the dependence on mosaic size, clustering hyperparameters and variable embeddings size that constrains Yottixel. This produces compact, consistent, and morphology aware slide embeddings. Second, CLAM-guided patch selection focuses computation on diagnostically salient regions, reducing noise prior to FM embedding and enabling the aggregator to capture context while remaining efficient. Third, SR filtering introduces a lightweight, automated consistency constraint: by retaining only neighbors whose predicted labels match the query's predicted label, the method stabilizes retrieval, especially

in heterogeneous archives where visually similar but label-inconsistent slides would otherwise degrade the ranked list. SR filtering is deployable and robust: even with imperfect classification, prediction consistency tends to align with morphology, reinforcing coherent neighborhoods and improving both accuracy and NDCG.

Comparative analysis across encoders underscores the importance of pathology-adapted pretraining for retrieval. UNI consistently delivers the strongest performance across datasets and ranks, suggesting that large, domain-tuned ViTs capture stable, semantically meaningful features crucial for slide level similarity. Using NDCG, adds essential sensitivity to rank positions that accuracy alone cannot capture. The observed NDCG gains parallel, and often amplify, the accuracy improvements, showing that our pipeline not only identifies more correct matches but also prioritizes them near the top of the list.

The system's automation and flexibility carry practical benefits. Unlike Yottixel, our approach removes manual choices for cluster counts and sampling rates and is invariant to patch count, simplifying deployment across diverse archives. The same pipeline supports slide and patch level queries, enabling indexing of heterogeneous datasets without retooling. Although we evaluate with a static validation memory bank, the architecture supports continual growth, allowing the bank to densify over time as more queries are processed. Beyond whole slide images, our framework generalizes to whole slide images derived patches, where initial patch-level tests show promising results.

Limitations include the size and variability of datasets. Further, SR filtering depends on a reasonably calibrated classifier; integrating uncertainty aware filtering, open set retrieval, and out of distribution detection would enhance robustness. Prospective user studies are needed to assess interpretability and impact on diagnostic workflows.

8. Conclusion

We introduce a fully automated, patch count independent WSI retrieval framework that combines FM embeddings, AttentionMIL aggregation, and SR label-consistency filtering. Across CAMELYON16 and BRACS datasets, the method establishes new SOTA performance, improving AccMV@5 from 77.49% to 89.92% on CAMELYON16, from 54.12% to 75.86% on BRACS level-1, and from 36.47% to 51.72% on BRACS level-2, with an increases in NDCG that confirm better early-rank relevance. By using NDCG in WSI retrieval, we provide a more informative assessment of ranking quality that aligns with clinical needs. The resulting system is general purpose, annotation-free at retrieval time, dataset agnostic, and publicly available, supporting both slide and patch level queries and designed for continual memory bank growth.

Future work will pursue large scale indexing with approximate nearest neighbor search, uncertainty-aware filtering and open-set handling, and prospective clinical studies to quantify benefits for diagnostic decision support.

Acknowledgments

This work was supported by the German BMFTR-funded SATURN3 project (01KD2206C).

References

J. Barker, A. Hoogi, A. Depeursinge, and D. L. Rubin. Automated classification of brain tumor type in whole slide digital pathology images using local representative tiles. *Medical Image Analysis*, 30:60–71, 2016.

B.E. Bejnordi, M. Veta, P.J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J.A. Van Der Laak, M. Hermsen, Q.F. Manson, M. Balkenhol, and O. Geessink. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

A. Bellogín, P. Castells, and I. Cantador. Precision-oriented evaluation of recommender systems: An algorithmic comparison. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, pages 333–336. ACM, 2011.

B. Böttcher, M. van Assen, R. Fari, P.L. von Knebel Doeberitz, E.Y. Kim, E.A. Berkowitz, F.G. Meinel, and C.N. De Cecco. Evaluation of a content-based image retrieval system for radiologists in high-resolution CT of interstitial lung diseases. *European Radiology Experimental*, 9(1):4, 2025.

N. Brancati, A.M. Anniciello, P. Pati, D. Riccio, G. Scognamiglio, G. Jaume, G. De Pietro, Maurizio Di B., A. Foncubierta, G. Botti, M. Gabrani, F. Feroce, and M. Frucci. BRACS: A dataset for BReAst Carcinoma Subtyping in H&E histology images. *Database*, 2022: baac093, 10 2022. ISSN 1758-0463. doi: 10.1093/database/baac093.

W. Bulten, K. Kartasalo, P.H.C. Chen, P. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D.F. Steiner, H. Van Boven, R. Vink, and C. Hulsbergen-van de Kaa. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the PANDA challenge. *Nature medicine*, 28(1):154–163, 2022.

C. Chen, M.Y. Lu, D. FK Williamson, T.Y. Chen, A.J. Schaumberg, and F. Mahmood. Fast and scalable search of whole-slide images via self-supervised deep learning. *Nature Biomedical Engineering*, 6(12):1420–1434, 2022a.

R. Chen, T. Ding, M. Lu, D. Williamson, G. Jaume, B. Chen, A. Zhang, D. Shao, A. Song, M. Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

R.J. Chen, M.Y. Lu, D.F. Williamson, T.Y. Chen, J. Lipkova, Z. Noor, M. Shaban, M. Shady, M. Williams, B. Joo, and F. Mahmood. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer cell*, 40(8):865–878, 2022b.

X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

O. Ciga, T. Xu, and A. Martel. Self-supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022.

T. Ding, S.J. Wagner, A.H. Song, R.J. Chen, M.Y. Lu, A. Zhang, A.J. Vaidya, G. Jaume, M. Shaban, A. Kim, and D.F. Williamson. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, pages 1–13, 2025.

S. Foersch, C. Glasner, A.C. Woerl, M. Eckstein, D.C. Wagner, S. Schulz, F. Kellers, A. Fernandez, K. Tserea, M. Kloth, and A. Hartmann. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nature medicine*, 29 (2):430–439, 2023.

Y. Fu, A. Jung, R. Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature cancer*, 1(8):800–810, 2020.

J. Galaro, V. Chaudhary, and A. Madabhushi. Comparative evaluation of feature extraction methods for classifying histopathological tissue. *Biomedical Engineering Online*, 10(1): 1–17, 2011.

D. Gutman, J. Cobb, D. Somanna, Y. Park, F. Wang, T. Kurc, J. Saltz, D. Brat, L. Cooper, and J. Kong. Cancer digital slide archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *Journal of the American Medical Informatics Association*, 20(6):1091–1098, 2013.

X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T. Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, pages 173–182. ACM, 2017.

X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 639–648. ACM, 2020.

N. Hegde, J.D. Hipp, Y. Liu, M. Emmert-Buck, E. Reif, D. Smilkov, M. Terry, C.J. Cai, M.B. Amin, C.H. Mermel, and P.Q. Nelson. Similar image search for histopathology: SMILY. *NPJ digital medicine*, 2(1):56, 2019.

K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

S. Kalra, H. Tizhoosh, C. Choi, S. Shah, P. Diamandis, C. Campbell, and L. Pantanowitz. Yottixel–an image search engine for large archives of histopathology whole slide images. *Medical Image Analysis*, 65:101757, 2020.

J.N. Kather, L.R. Heij, H.I. Grabsch, C. Loeffler, A. Echle, H.S. Muti, J. Krause, J.M. Niehues, K.A. Sommer, P. Bankhead, and L.F. Kooreman. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature cancer*, 1(8):789–799, 2020.

I. Lahr, S. Alfasly, P. Nejat, J. Khan, L. Kottom, V. Kumbhar, A. Alsaafin, A. Shafique, S. Hemati, G. Alabtah, N. Comfere, D. Murphree, A. Mangold, S. Yasir, C. Meroueh, L. Boardman, V. Shah, J. Garcia, and H. Tizhoosh. Analysis and validation of image search engines in histopathology. *IEEE Reviews in Biomedical Engineering*, pages 1–19, 2024. doi: 10.1109/RBME.2024.3425769.

S. Li, Y. Zhao, J. Zhang, T. Yu, J. Zhang, and Y. Gao. High-order correlation-guided slide-level histology retrieval with self-supervised hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11008–11023, 2023.

D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1150–1157, 1999.

M. Lu, D. Williamson, T. Chen, R. Chen, M. Barbieri, and F. Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

J. Mao, X. Liu, X. He, D. Jin, and Y. Li. Simplex: A simple and strong baseline for collaborative filtering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 103–112. ACM, 2021.

S. Mehta, A. Sethi, and S. Shah. Content-based sub-image retrieval system for pathology slides. In *International Conference on Image Analysis and Processing*, pages 129–138, 2009.

S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 452–461. AUAI Press, 2009.

H. Sharma, N. Zerbe, D. Heim, S. Wienert, O. Hellwich, and P. Hufnagl. Content-based image retrieval of histopathological images based on local representative tiles. *International Journal of Computer Assisted Radiology and Surgery*, 7(3):451–457, 2012.

X. Shi, M. Sapkota, F. Xing, F. Liu, L. Cui, and L. Yang. Pairwise based deep ranking hashing for histopathology image classification and retrieval. *Pattern Recognition*, 81: 14–22, 2018.

C.L. Srinidhi, S.W. Kim, F.D. Chen, and A.L. Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Medical Image Analysis*, 75:102256, 2022.

H.R. Tizhoosh and L. Pantanowitz. On image search in histopathology. *Journal of Pathology Informatics*, 15:100375, 2024.

C. Tommasino, F. Merolla, C. Russo, S. Staibano, and A.M. Rinaldi. Histopathological image deep feature representation for CBIR in smart PACS. *Journal of Digital Imaging*, 36(5):2194–2209, 2023.

H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, 2021.

J. A. Vanegas, S. Randhawa, and I. K. Sethi. Automatic classification of histopathology images using color and texture features. *Journal of Pathology Informatics*, 5(1):23, 2014.

X. Wang, Y. Du, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han. RetCCL: clustering-guided contrastive learning for whole-slide image retrieval. *Medical Image Analysis*, 83:102645, 2023.

H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu, Y. Xu, M. Wei, W. Wang, S. Ma, F. Wei, J. Yang, C. Li, J. Gao, J. Rosemon, T. Bower, S. Lee, R. Weerasinghe, B. Wright, A. Robicsek, B. Piening, C. Bifulco, S. Wang, and H. Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 2024.

H. Xue, X. Dai, J. Zhang, S. Huang, and J. Chen. Deep matrix factorization models for recommender systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3203–3209. IJCAI, 2017.

L. Zheng, J. Zhang, and G. Lu. A CBIR system for pathology images. In *Proceedings of SPIE*, volume 5033, pages 152–160, 2003.