# DINOHist: Fine-Tuning Self-Supervised Foundation Models for Histopathological Image Retrieval

**José Solenir L. Figuerêdo**[1,2,3] **and Rodrigo Tripodi Calumby**[1,2]

[1]University of Feira de Santana (UEFS) – Feira de Santana – BA – Brazil

[2]Postgraduate Program in Computer Science – UEFS

[3]Federal Institute of Sergipe (IFS) – Campus Lagarto – Lagarto – SE – Brazil

jose.figueredo@ifs.edu.br

rtcalumby@uefs.br

***Abstract.*** *This study investigates self-supervised learning (SSL) for feature extraction in Content-Based Histopathological Image Retrieval (CBHIR). We evaluate pre-trained, domain-specific, and fine-tuned DINOv2 models (DINOHist-S and DINOHist-B) on glomerular datasets. The analysis includes comparisons with state-of-the-art histopathology-specific models to assess domain adaptation. Results demonstrate that fine-tuning enables DINOv2 to outperform state-of-the-art specialized models, even with limited labeled data. The proposed models achieve superior effectiveness, particularly at lower-ranking positions, with statistical significance. These findings show that fine-tuned foundation models can surpass specialized approaches while reducing computational costs.*

## 1. Introduction

Histopathology focuses on the microscopic analysis of biological tissues. It supports the diagnosis of diseases such as cancer, hepatic, and renal disorders [Erfankhah et al. 2019, Abels et al. 2019, Chagas et al. 2020, L'Imperio et al. 2021]. This process is vital for therapeutic planning and medical research and is widely used in clinical practice. However, manual histological slide analysis is laborious, time-consuming, and prone to subjectivity and errors [Filiot et al. 2023]. Therefore, to reduce the workload of specialists and improve the objectivity of image analysis, Computer-Aided Diagnosis (CAD) systems have emerged as a promising application [Shi et al. 2018]. Among the different approaches employed in these systems, Content-Based Image Retrieval (CBIR) methods have gained particular attention.

CBIR is a computational approach that enables the retrieval of visually similar images from large-scale databases based on intrinsic visual features. In the medical domain, CBIR systems have been applied to histopathology to support pathologists in comparing cases under examination with previously diagnosed instances, thereby contributing to more informed and evidence-based decision-making [Cazzolato et al. 2019]. Given the continuous expansion of digital medical image repositories, such systems become particularly relevant by reducing reliance on manual search processes.

The widespread adoption of digitization technologies, particularly Whole Slide Imaging (WSI), has significantly advanced the field of digital and computational pathology [Navid Farahani et al. 2015]. This workflow typically involves multiple stages, in-

cluding slide selection and digitization, data annotation, development of artificial intelligence models, and subsequent evaluation, as illustrated in Figure 1. However, the substantial volume of data generated through this process makes manual search impractical, thus highlighting the need for automated and efficient methods for analysis and retrieval of relevant information.

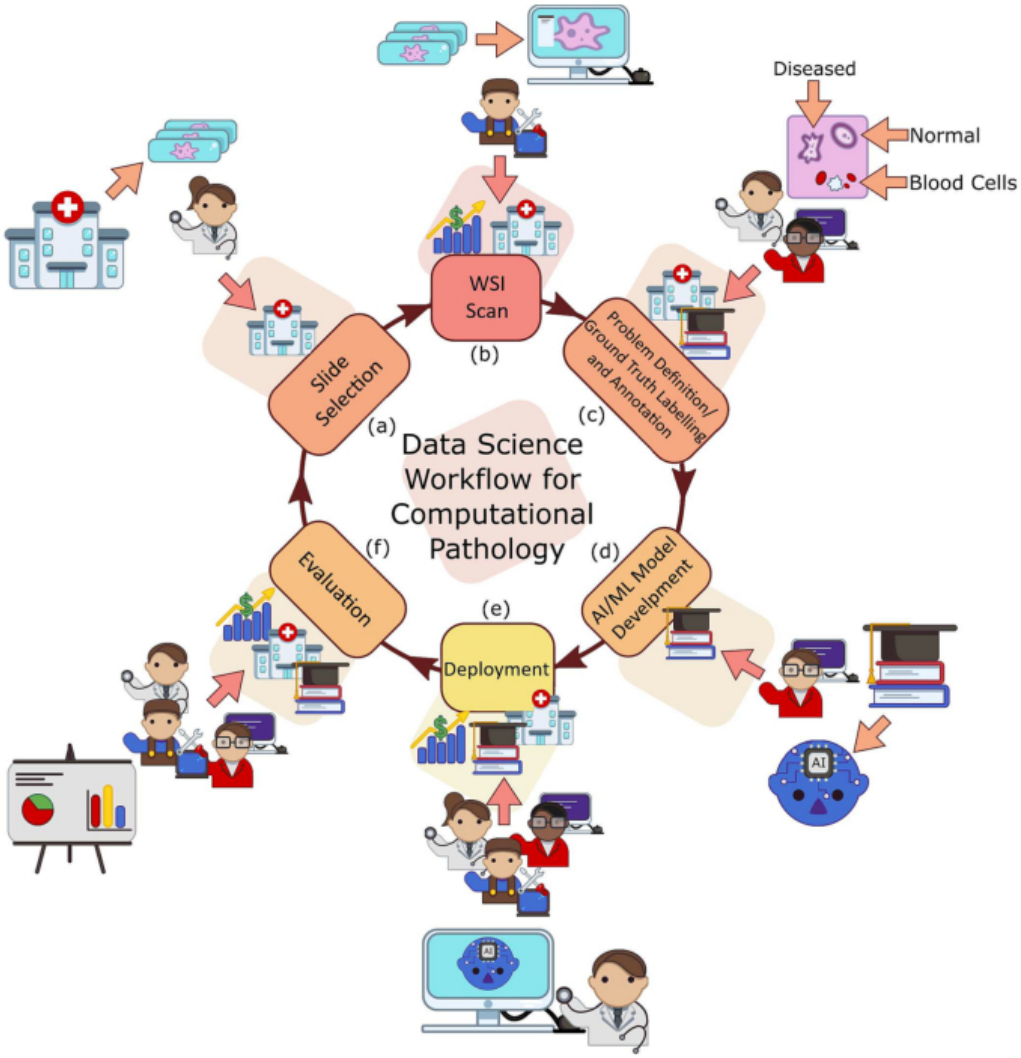


**Figure 1. Workflow commonly used in studies of computational/digital pathology [Hosseini et al. 2024].**

CBIR systems represent images through high-dimensional features, which are used to compute similarity between a query image and those stored in databases [Gudivada and Raghavan 1995]. In recent years, deep learning-based approaches have achieved significant performance improvements in this task [Kalra et al. 2020, Haq et al. 2021, Souid et al. 2023]. However, these methods rely on large volumes of labeled data, which are costly and complex to acquire in the medical domain. This challenge becomes even more critical in the context of glomerular histopathological images, where the identification and segmentation of glomeruli require meticulous expert analysis, making the process time-consuming, error-prone, and subject to high inter-observer variability. Furthermore, these images present considerable

morphological, staining, and tissue-pattern variability, which which limits conventional models from learning robust representations.

To address these limitations, strategies such as transfer learning and self-supervised learning have been increasingly investigated, as they enable the extraction of meaningful representations from unlabeled data [Chen et al. 2019, Gui et al. 2024]. In this context, methods such as SwAV, MoCo v3, iBOT, and EsVIT have demonstrated strong performance across a wide range of tasks [Caron et al. 2020, Chen et al. 2021, Zhou et al. 2022, Li et al. 2022]. Additionally, domain-specific approaches for histopathology, such as retCCL [Wang et al. 2023] and large-scale self-supervised models [Filiot et al. 2023], have further demonstrated the potential of these techniques for learning meaningful visual representations. Nevertheless, significant gaps still remain regarding their effective integration and evaluation within CBIR systems, particularly in highly specialized domains such as nephropathology.

In this context, this work investigates the effectiveness of self-supervised models for glomerular histopathological image retrieval, with an emphasis on adapting foundation models pre-trained on general domains to limited-data scenarios. Specifically, we explore self-supervised fine-tuning on small-scale glomerular image datasets to assess its ability to learn domain-specific representations. Additionally, we examine whether such models can outperform large-scale, domain-specific approaches while significantly reducing computational cost. This investigation contributes to the development of more effective, accessible, and practical CBIR systems that are better aligned with the constraints of digital pathology, particularly in settings with limited data and computational resources.

The remainder of this paper is organized as follows. Section 2 presents the related work and Section 3 describes the proposed experimental process. The results and discussions are presented in Section 4. Finally, Section 5 brings the conclusions and future works.

## 2. Related Works

CBIR systems index and retrieve images using visual attributes such as color, texture, and shape. Recent advances in Artificial Intelligence, particularly deep and self-supervised learning, have directly improved CBIR performance by enhancing the ability to extract and compare meaningful visual features, enabling more robust and efficient retrieval methods [Hegde et al. 2019, Yang et al. 2020, Kalra et al. 2020, Zheng et al. 2022, Mohammad Alizadeh et al. 2023, Wickstrøm et al. 2023, Wang et al. 2023, Filiot et al. 2023].

Hegde et al. proposed SMILY, a system using convolutional neural networks to generate image embeddings for histopathological image retrieval. In contrast, Yang et al. introduced the BoB (Bunch of Barcodes) method, which represents whole-slide images as mosaics of binary-encoded patches, differing from SMILY's embedding-based approach. Further, Zheng et al. proposed a graph-based framework for retrieving regions of interest that leverages spatial relationships between image regions, thereby distinguishing it from both SMILY and BoB.

In the context of hashing-based retrieval, Seyed et al. proposed a convolutional Siamese network with binary encoding and an optimized loss function, achieving superior

effectiveness compared to prior methods on public benchmarks. Shifting focus, Kristoffer et al. introduced a self-supervised framework for liver CT imaging that integrates explainability through the RELAX method, thereby enhancing the interpretability of retrieval results. Building upon these advancements, Wang et al. more recently presented RetCCL, a large-scale self-supervised framework trained on approximately 15 million unlabeled images. Their approach learns highly transferable representations without requiring task-specific fine-tuning, establishing state-of-the-art performance in histopathological image retrieval.

Filiot et al. presented a comprehensive analysis of Masked Image Modeling (MIM) within the histopathological domain, leveraging the iBOT architecture based on Vision Transformers. Their approach trains models to reconstruct masked regions of input images, enabling the learning of rich semantic representations without relying on labeled data. A key contribution of this work is its large-scale experimental design, which systematically analyzes the impact of dataset size, model capacity, and pre-training strategies on the quality of the learned representations.

Filiot et al. further evaluated their models across multiple downstream tasks, including cancer classification over diverse organs and datasets, demonstrating strong generalization capabilities. Their results indicate that representations learned through masked image modeling (MIM) can match or even outperform fully supervised approaches, particularly in scenarios with limited labeled data. An additional contribution is the public release of the pre-trained models, facilitating their reuse as foundation models within the histopathological domain. Although the study does not explicitly address CBIR, its findings are highly relevant, as they provide robust feature extraction mechanisms that can substantially improve image retrieval performance.

In contrast to the aforementioned approaches, this work investigates the use of self-supervised models fine-tuned on small, unlabeled datasets, with a particular focus on evaluating generalization in glomerular histopathological images. This domain presents significant challenges, including high morphological variability across patients, staining artifacts, intra- and inter-class heterogeneity, and the need for multi-scale structural analysis [Juan-Ferrer et al. 2026]. These factors make the extraction of discriminative representations particularly challenging and limit the effectiveness of supervised approaches that rely on large volumes of annotated data.

In this context, the proposed approach addresses the image retrieval task, focusing on realistic scenarios where limited annotations pose a fundamental challenge. Self-supervised models provide several advantages, including the ability to leverage large volumes of unlabeled data, increased robustness to domain variations, and improved generalization in low-annotation regimes. Furthermore, fine-tuning on limited datasets enables the adaptation of previously learned representations to the characteristics of the glomerulus and may outperform specialized models trained from scratch.

## 3. Experimental Pipeline

The experimental process conducted in this study is illustrated in Figure 2. It consists of 6 steps: (1) Data collection, (2) Dataset partitioning, (3) Preprocessing, (4) Adaptation of pre-trained models (fine-tuning), (5) Extraction of feature vectors and, finally, (6) Model evaluation.

### 3.1. Datasets and Partitioning

Computational pathology has been applied to a wide range of medical tasks, including content-based image retrieval [Hosseini et al. 2024]. Before analysis, the extracted tissue follow a histological preparation process. An important step in this process is the use of staining techniques to improve the visualization of specific tissue structures. Different staining methods can be used, each highlighting particular characteristics of the tissue, such as Periodic Acid–Schiff (PAS), Hematoxylin and Eosin (H&E), PAMS (*Picroaniline Methenamine Silver*), PS (*Picro-Sirius Red*), AZAN (Alcian Blue and Orange), and PICRO (Picro-Chromic Weigert). The staining method selected directly influences the contrast and visibility of cellular structures in histopathological images. Such variations can influence feature extraction and consequently affect the overall effectiveness of computational image analysis methods.

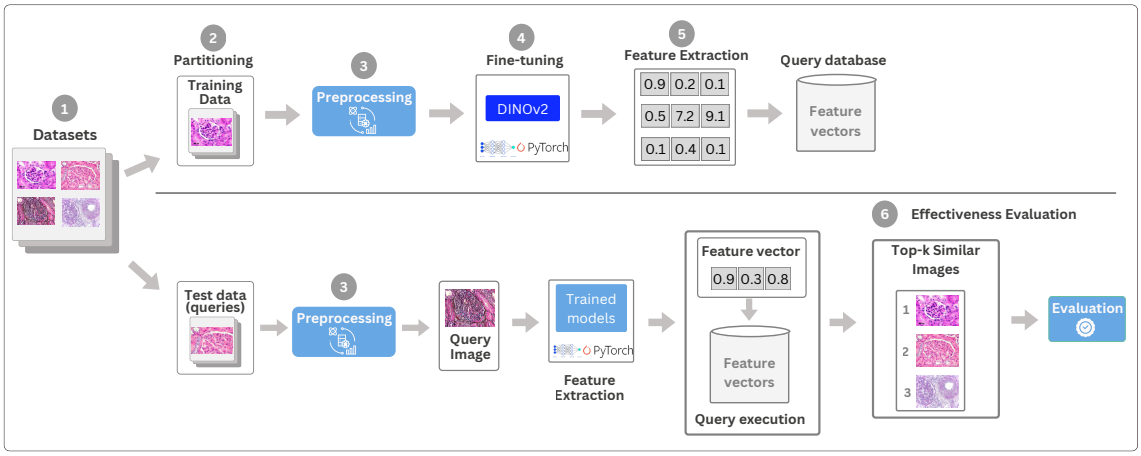


**Figure 2. Experimental Benchmark Workflow.**

Three datasets were used in the experimental process, consisting of renal glomeruli images with different types of lesions and acquired using distinct staining techniques. In this study, the datasets were named as: PathoSpotter-HE, PathoSpotter-PAS, and PathoSpotter-MultiStain. Each of these databases is detailed below.

- **PathoSpotter-HE:** It is part of the *PathoSpotter*[1] project, related to the Oswaldo Cruz Foundation (Fiocruz) of Bahia. This particular *dataset* has 4947 images, divided into six classes of interest: Normal (869), Hypercellularity (1237), Primary membranous (712), Secondary membranous (1354), Sclerosis with membranous (264) and Sclerosis without membranous (511). In order to improve the visualization of specific structures, the H&E stain was used.
- **PathoSpotter-PAS:** This dataset is also part of the *PathoSpotter* project, but unlike PathoSpotter-HE, it uses PAS as a stain. This set has 2390 images, and is also divided into six classes of interest: Normal (293), Hypercellularity (637), Primary membranous (367), Secondary membranous (609), Sclerosis with membranous (156) and Sclerosis without membranous (328).
- **PathoSpotter-MultiStain:** This dataset is also part of the *PathoSpotter* project, but uses multiple stain. In addition to having images that used H&E and PAS,

[1] https://pathospotter.bahia.fiocruz.br/

it also has images that used PAMS, PS, AZAN, and PICRO. The database is composed of 10184 images, also covering six classes: Normal (1570), Hypercellularity (2043), Primary membranous (1608), Secondary membranous (3170), Sclerosis with membranous (513) and Sclerosis without membranous (1280).

Another stage of this study involved splitting the datasets into training, validation, and test sets using a 70:20:10 ratio, respectively. Table 1 presents a detailed description of the number of samples in the sets after performing the partitioning.

### 3.2. Preprocessing

After the partitioning step, a preprocessing pipeline was conducted. The main goal of this step was enhance training quality by introducing variability, while also ensuring that the data is in the correct format. For the training set, the following pipeline was followed:

- **Resizing:** All images were scaled to $224$ x $224$ *pixels*, as required by the network architecture, which was designed to operate on inputs of this size.
- **Random horizontal flip:** A random horizontal flip is applied to the images with a default probability of $50\%$, aiming to increase the robustness of the model with respect to different image orientations.
- **Conversion to RGB:** Converts the image to RGB color mode if it is not already in that format, ensuring it is processed with three color channels.
- **Normalization of pixel values:** A normalization is applied to the values of the *pixels* of the image using means and standard deviations specific to each color channel (R, G, B).

The same procedures were applied to the validation and test sets, except for the random horizontal flip.

**Table 1. Datasets Statistics**

| Dataset | Number of Classes | Stain | Train | Validation | Test | Total |
|---|---|---|---|---|---|---|
| PathoSpotter-HE | 6 | HE | 3462 | 990 | 495 | 4947 |
| PathoSpotter-PAS | 6 | PAS | 1673 | 478 | 239 | 2390 |
| PathoSpotter-MultiStain | 6 | HE, PAS, PAMS, PS, AZAN, PICRO | 7128 | 2038 | 1018 | 10184 |

### 3.3. Experimental Setup

One of the main objectives of this study is to evaluate the effectiveness of fine-tuned self-supervised models for histopathological image retrieval. Fine-tuning refers to the process of refining a model that has been pre-trained on large-scale datasets, with the goal of tailoring its representations to a specific task or domain. While fine-tuning is traditionally performed using labeled data, this work explores its applicability in scenarios characterized by limited and unlabeled datasets. In particular, we investigate whether meaningful domain adaptation can still be achieved under such constraints. For this purpose, the DINOv2 model is employed in its Small (ViT/14) and Base (ViT/14) variants [Oquab et al. 2023].

*Fine-tuning* was performed using the official implementation of DINOv2[2]. Publicly available pre-trained weights for the *backbones* were loaded into both the *teacher* and *student* networks. Model adaptation was performed separately for each *dataset* using its corresponding training split. The test set was used exclusively to evaluate the quality of the learned representations. It is important to note that all images were processed at the *patch* level.

In our experiments, we used the same training pipeline configuration as during network pre-training, with a base learning rate of $2 \times 10^{-4}$. We additionally evaluated a learning rate of $5 \times 10^{-4}$, as suggested in prior studies [Filiot et al. 2023, Oquab et al. 2023]. Training was conducted for 2000 epochs, using a batch size of 32 for both the Small (ViT/14) and Base (ViT/14) architectures. In contrast to the default configuration provided by the reference implementation, we employed a crop value of 1, corresponding to a global crop strategy. This modification was motivated by preliminary experiments indicating that preserving the entire glomerular structure was more suitable for these datasets.

Unlike natural-image domains, glomerular histopathological images rely on global morphological organization, and subtle structural differences may correspond to distinct lesions. In this context, aggressive local cropping may remove diagnostically relevant regions or disrupt important structural relationships within the glomerulus. Therefore, preserving the complete glomerular context during self-supervised adaptation proved more effective than focusing only on local regions. Although an ablation study with different crop ratios was not performed, the adopted configuration was guided by preliminary empirical observations and domain-specific characteristics. All remaining hyperparameters were kept the same as those specified in the configuration file provided in the official implementation.

The experiments were performed across all combinations of datasets, architectures, and hyperparameter settings. After the fine-tuning stage, feature vectors were extracted from the images and stored in a database for subsequent evaluation in the CBIR task. For this evaluation, the training and validation sets were merged to create the retrieval database. Image retrieval was then performed using the k-Nearest Neighbors (kNN) algorithm with $K = 50$. Finally, the test set was used as the query set to evaluate the quality of the learned feature representations.

All experiments were performed using a single GPU NVIDIA A100-SXM (80 GB). The models adapted in this study have been made publicly available at the following electronic address: `https://drive.google.com/drive/folders/11hKJL0c6xcC4AfpHSUPixoRYmoT01rbV?usp=sharing`.

### 3.4. Evaluation

The comparative analysis performed in this study includes the following feature extractors: (i) ResNet50 [He et al. 2016], pre-trained on the ImageNet dataset; (ii) DINOv2, in its Small and Base variants, pre-trained on a large-scale collection of natural images [Oquab et al. 2023]; (iii) state-of-the-art histopathology-specific feature extractors, including RetCCL [Wang et al. 2023] and Phikon ViT-B [Filiot et al. 2023], main baseline; and (iv) the proposed models, DINOHist-S and DINOHist-B. Although ResNet50 is

[2]https://github.com/facebookresearch/dinov2

a classical architecture, its inclusion remains justified given its widespread adoption as a feature extractor in the literature [Wang et al. 2023].

The models were evaluated using the *Mean Average Precision (MAP) at K*, or simply $MAP@K$. It computes the mean Average Precision (AP) across multiple queries, where AP quantifies the quality of the ranking of relevant results within a retrieved set. This measure allows evaluating both the relevance of the ranked images and the system's effectiveness in positioning relevant images at the top. In this work, the relevant images are those that belong to the class of the query image.

Image similarity was calculated using cosine similarity, which measures the degree of similarity between two vectors in multidimensional space. The models were compared using the Wilcoxon test with a 95% confidence level. Since several queries were executed, the $MAP@K$ value reported is the average across these executions.

## 4. Results and discussion

Table 2 presents the effectiveness of the models evaluated in this study at various ranking depths ($MAP@5$, $MAP@10$, $MAP@20$, $MAP@30$, $MAP@40$, $MAP@50$). The "Histo" column indicates whether the model was trained on histopathological data. DINOHist-S and DINOHist-B achieved the highest $MAP@k$ scores, especially at lower ranks, demonstrating that fine-tuning to the target domain produced the best results. Fine-tuned DINOHist-S and DINOHist-B models exceeded their non-fine-tuned variants by over 20%. Although improvements over the baselines (RetCCL and Phikon ViT-B) are not large in absolute terms, they are meaningful in this context. Since $MAP@k$ is highly sensitive to the ordering of relevant instances, modest numerical gains may reflect meaningful improvements in the discriminative quality of the learned representations and in the ranking consistency of the retrieval system.

Furthermore, in highly specialized medical imaging domains such as glomerular histopathology, visual patterns are often subtle and characterized by high inter-class similarity and intra-class variability. Under these conditions, achieving consistent improvements over strong domain-specific baselines becomes particularly challenging. Therefore, the observed gains suggest that the proposed fine-tuning strategy enabled the models to better capture domain-specific morphological characteristics, leading to more accurate and semantically coherent retrieval results.

The experimental results are particularly relevant, considering that the models were fine-tuned using a single GPU and small-scale datasets. In addition, they highlight that adapting foundation models to task-specific data can outperform domain-specific models trained at a large scale, while requiring only a fraction of the computational resources and training time. For instance, on the *PathoSpotter-HE* dataset, the DINOHist-S model required approximately 40 minutes of training. In comparison to Phikon ViT-B, trained on datasets of similar size, only $0.05\%$ of GPU time was required (40 minutes vs. 1,216 hours) to achieve the results reported in Table 2. This finding is highly significant, as it underscores the potential of self-supervised models in scenarios characterized by limited labeled data and restricted computational resources.

Table 3 presents the results of the statistical test, comparing the models proposed in this study with Phikon ViT-B, adopted as the main *baseline*. The results show that

**Table 2. Effectiveness of the models used in this study.**

| Dataset | Model | Histo | MAP@5 | MAP@10 | MAP@20 | MAP@30 | MAP@40 | MAP@50 |
|---|---|---|---|---|---|---|---|---|
| PathoSpotter-HE | ResNet50 | | 0.5571 | 0.4933 | 0.4341 | 0.3938 | 0.3711 | 0.3606 |
| | DINOv2 (ViT-S) | | 0.5608 | 0.4845 | 0.4292 | 0.3919 | 0.3792 | 0.3602 |
| | DINOv2 (ViT-B) | | 0.5460 | 0.4982 | 0.4427 | 0.4070 | 0.3846 | 0.3696 |
| | RetCCL | ✓ | 0.6804 | 0.5865 | 0.5149 | 0.4734 | 0.4428 | 0.4235 |
| | Phikon ViT-B | ✓ | **0.7257** | 0.6206 | 0.5148 | 0.4654 | 0.4269 | 0.4044 |
| | DINOHist-S | ✓ | 0.7135 | 0.6320 | **0.5685** | 0.5136 | 0.4762 | 0.4469 |
| | DINOHist-B | ✓ | 0.7131 | **0.6369** | 0.5623 | **0.5231** | **0.4856** | **0.4577** |
| PathoSpotter-PAS | ResNet50 | | 0.5366 | 0.4873 | 0.4319 | 0.3954 | 0.3767 | 0.3574 |
| | DINOv2 (ViT-S) | | 0.5527 | 0.4933 | 0.4203 | 0.3868 | 0.3582 | 0.3369 |
| | DINOv2 (ViT-B) | | 0.5455 | 0.5009 | 0.4231 | 0.3935 | 0.3688 | 0.3488 |
| | RetCCL | ✓ | 0.6224 | 0.5378 | 0.4656 | 0.4219 | **0.3911** | **0.3815** |
| | Phikon ViT-B | ✓ | 0.6532 | 0.5528 | 0.4608 | 0.4126 | 0.3909 | 0.3689 |
| | DINOHist-S | ✓ | **0.6718** | **0.5678** | 0.4714 | 0.4211 | 0.3810 | 0.3490 |
| | DINOHist-B | ✓ | 0.6589 | 0.5826 | **0.4886** | **0.4289** | 0.3900 | 0.3652 |
| PathoSpotter MultiStain | ResNet50 | | 0.5390 | 0.4815 | 0.4271 | 0.3924 | 0.3736 | 0.3571 |
| | DINOv2 (ViT-S) | | 0.5490 | 0.4996 | 0.4371 | 0.4099 | 0.3799 | 0.3621 |
| | DINOv2 (ViT-B) | | 0.5576 | 0.4880 | 0.4286 | 0.3989 | 0.3797 | 0.3641 |
| | RetCCL | ✓ | 0.6490 | 0.5595 | 0.4837 | 0.4438 | 0.4151 | 0.3990 |
| | Phikon ViT-B | ✓ | **0.6839** | 0.5828 | 0.4929 | 0.4427 | 0.4141 | 0.3912 |
| | DINOHist-S | ✓ | 0.6599 | 0.5919 | 0.5235 | **0.4948** | **0.4711** | **0.4515** |
| | DINOHist-B | ✓ | 0.6720 | **0.5977** | **0.5315** | 0.4946 | 0.4674 | 0.4444 |

DINOHist-S and DINOHist-B were statistically superior or equivalent to Phikon ViT-B across multiple ranking levels. This finding is particularly relevant, as it highlights the potential of DINOHist models in challenging scenarios, such as those involving the nephropathology domain. This superiority is highlighted at lower-ranking positions, indicating that DINOHist models have a greater ability to retrieve a broader set of relevant images across the result list. This behavior suggests improved coverage of the semantic space, although not always with substantial gains at the top-ranked positions. In CBIR tasks, this characteristic can be particularly advantageous in exploratory scenarios, where specialists aim to analyze multiple similar instances beyond the most immediate matches.

**Table 3. Wilcoxon test results (95% confidence) comparing our models with the best baseline. Green indicates statistically significant superiority, while white indicates no significant differences. No inferiority was observed.**

| Dataset | Models | MAP@5 | MAP@10 | MAP@20 | MAP@30 | MAP@40 | MAP@50 |
|---|---|---|---|---|---|---|---|
| PathoSpotter-HE | DINOHist-S vs Phikon ViT-B | | | | | | |
| | DINOHist-B vs Phikon ViT-B | | | | | | |
| PathoSpotter-PAS | DINOHist-S vs Phikon ViT-B | | | | | | |
| | DINOHist-B vs Phikon ViT-B | | | | | | |
| PathoSpotter MultiStain | DINOHist-S vs Phikon ViT-B | | | | | | |
| | DINOHist-B vs Phikon ViT-B | | | | | | |

Additionally, the consistent effectiveness at deeper ranking positions indicates that the descriptors learned by DINOHist models are more robust to the intrinsic variability of histopathological data, such as staining variations, complex morphological structures, and inter-sample heterogeneity. This finding reinforces the notion that fine-tuning self-supervised models leads to more generalizable representations in the target domain.

Finally, these results suggest that, although improvements at top-ranked positions may be limited, DINOHist models stand out for their ability to maintain relevance throughout the entire *ranking*, making them particularly promising for practical applications in medical image retrieval, where the analysis of multiple similar cases is often required.

Figures 3, 4, and 5 present illustrative examples of retrieval for each *dataset*, using DINOv2 (ViT-B) and the best-performing histopathological model. Queries were chosen to highlight the improvements from the adapted models. The top ten-ranked images are shown, with DINOv2 (ViT-B) results at the top and DINOHist-B results at the bottom. Relevant images are highlighted in blue, and irrelevant images are highlighted in red. In these scenarios, histopathological models show high effectiveness across ranking levels. For these models, the first retrieved image is consistently relevant in all *datasets*, and relevance is maintained among top-ranked images. This indicates strong performance throughout the rankings. In contrast, DINOv2 (ViT-B) retrieves more irrelevant images under the same conditions. For example, on the PathoSpotter-HE dataset, DINOv2 (ViT-B) retrieved only 1 image from the class of interest, whereas its fine-tuned version retrieved 9 relevant images.

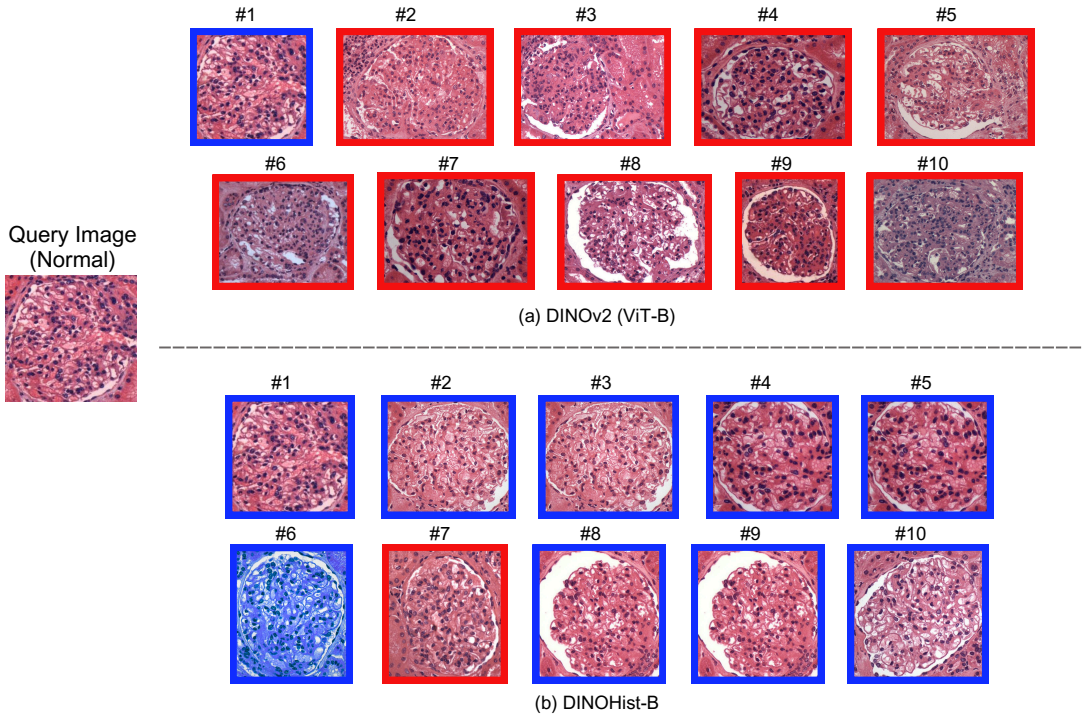


**Figure 3. Top-10 results for the specified query image from the PathoSpotter-HE dataset. At the top we have the result of DINOv2 (ViT-B) (MAP@10 = 0.1000). At the bottom the result of DINOHist-B (MAP@10 = 0.9060).**

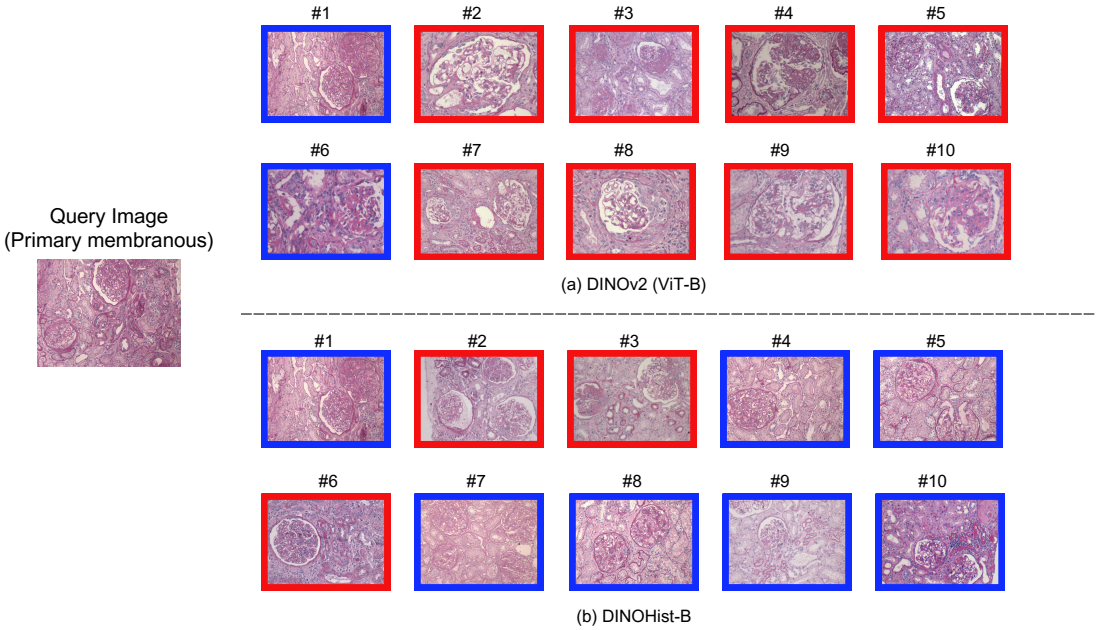


**Figure 4. Top-10 results for the specified query image from the PathoSpotter-PAS dataset. At the top we have the result of DINOv2 (ViT-B) (MAP@10 = 0.2000). At the bottom the result of DINOHist-B (MAP@10 = 0.9129.**

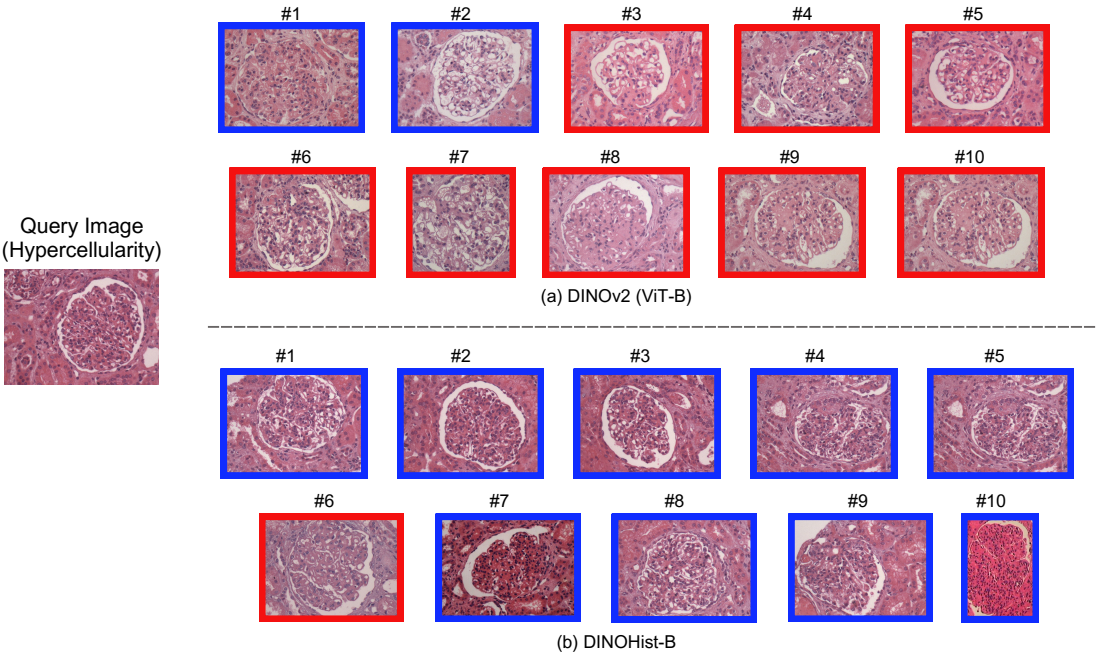


**Figure 5. Top-10 results for the specified query image from the PathoSpotter-MultiStain dataset. At the top we have the result of DINOv2 (ViT-B) (MAP@10 = 0.1556). At the bottom the result of DINOHist-B (MAP@10 = 0.9283.**

Overall, the results highlight the potential of self-supervised architectures, particularly in scenarios with limited datasets and challenges in obtaining reliable annotations. These issues highlight the relevance of this study, which investigates self-supervised mod-

els as a strategy to overcome them. In clinical practice, the ability to rapidly retrieve similar histopathological images from large databases can assist pathologists in case comparison, accurate diagnosis, and informed decision-making, thereby enhancing the practical impact of these models in clinical workflows.

## 5. Conclusion

Histopathological image retrieval aims to identify similar cases within large collections using visual descriptors. Traditional approaches typically rely on supervised learning methods, which require annotated datasets to train effective feature extractors. However, obtaining labeled histopathological data is a costly, time-consuming process and, in many scenarios, becomes impractical due to the need for expert knowledge and the inherent complexity of medical images. In this context, self-supervised models have emerged as a promising alternative, reducing dependence on labeled data while enabling the learning of rich, discriminative representations. Accordingly, this study developed and evaluated self-supervised models for image retrieval tasks in histopathological contexts.

The results demonstrate that adapting self-supervised models through domain-specific fine-tuning is an effective strategy for glomerular histopathological image retrieval. Although histopathology-oriented foundation models already provide substantial advantages over general-domain models, the fine-tuned models proposed in this study achieved superior performance, surpassing even state-of-the-art domain-specific feature extractors across several evaluation scenarios. Moreover, these improvements were consistently observed for different datasets and $MAP@K$ levels.

These findings indicate that self-supervised fine-tuning enables the learning of more discriminative and semantically meaningful visual representations tailored to the nephropathology domain, facilitating the recognition of subtle morphological patterns in glomerular images. Despite the inherent challenges of image retrieval in this highly specialized setting, the results highlight the potential of this approach to enhance content-based image retrieval systems and support diagnostic and clinical decision-making workflows.

For future work, this study proposes investigating novel self-supervised architectures, including DINOv3, to assess potential improvements in histopathological image retrieval. Additional datasets will be included, with analyses of hyperparameter effects and alternative fine-tuning methods. The study will also apply dimensionality reduction techniques and evaluate the performance of extractors using various classifiers to assess model robustness and generalization in different scenarios.

## Acknowledgement

This work was partially supported by UEFS-AUXPPG 2026 and CAPES-PROAP 2026 grants. We also thank the PathoSpotter project for providing the databases used in this study.

## References

Abels, E. et al. (2019). Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *J Pathol*, 249(3):286–294.

Caron, M. et al. (2020). Unsupervised learning of visual features by contrasting cluster assignments. NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Cazzolato, M., Rodrigues, L., Scabora, L., Zaboti, G., Vasconcelos, G., Chino, D., Jorge, A., Cordeiro, R., Traina-Jr, C., and Traina, A. (2019). A dbms-based framework for content-based retrieval and analysis of skin ulcer images in medical practice. In *Anais do XXXIV Simpósio Brasileiro de Banco de Dados*, pages 109–120, Porto Alegre, RS, Brasil. SBC.

Chagas, P., Souza, L., Araújo, I., Aldeman, N., Duarte, A., Angelo, M., Dos-Santos, W., and Oliveira, L. (2020). Classification of glomerular hypercellularity using convolutional features and support vector machine. *Artificial Intelligence in Medicine*, 103:101808.

Chen, L. et al. (2019). Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58:101539.

Chen, X., Xie, S., and He, K. (2021). An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9620–9629, Los Alamitos, CA, USA. IEEE Computer Society.

Erfankhah, H. et al. (2019). Heterogeneity-aware local binary patterns for retrieval of histopathology images. *IEEE Access*, 7:18354–18367.

Filiot, A. et al. (2023). Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*.

Gudivada, V. and Raghavan, V. (1995). Content based image retrieval systems. *Computer*, 28(9):18–22.

Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., and Tao, D. (2024). A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071.

Haq, N. F. et al. (2021). A deep community based approach for large scale content based x-ray image retrieval. *Medical Image Analysis*, 68:101847.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hegde, N. et al. (2019). Similar image search for histopathology: Smily. *npj Digital Medicine*, 2(1):56.

Hosseini, M. S., Bejnordi, B. E., Trinh, V. Q.-H., Chan, L., Hasan, D., Li, X., Yang, S., Kim, T., Zhang, H., Wu, T., Chinniah, K., Maghsoudlou, S., Zhang, R., Zhu, J., Khaki, S., Buin, A., Chaji, F., Salehi, A., Nguyen, B. N., Samaras, D., and Plataniotis, K. N. (2024). Computational pathology: A survey review and the way forward. *Journal of Pathology Informatics*, 15:100357.

Juan-Ferrer, P., Pérez-Sánchez, N., Pla, F., Mollineda, R. A., Roselló-Sastre, E., and Tajahuerce, E. (2026). Glomeruli detection and classification in histopathological images using deep learning semantic segmentation. *BMC Medical Imaging*, 26(1):103.

Kalra, S. et al. (2020). Yottixel – an image search engine for large archives of histopathology whole slide images. *Medical Image Analysis*, 65:101757.

Li, C. et al. (2022). Efficient self-supervised vision transformers for representation learning.

L'Imperio, V. et al. (2021). Digital pathology for the routine diagnosis of renal diseases: a standard model. *Journal of Nephrology*, 34(3):681–688.

Mohammad Alizadeh, S. et al. (2023). A novel siamese deep hashing model for histopathology image retrieval. *Expert Systems with Applications*, 225:120169.

Navid Farahani, A. et al. (2015). Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International*, 7:23–33.

Oquab, M. et al. (2023). Dinov2: Learning robust visual features without supervision.

Shi, X. et al. (2018). Pairwise based deep ranking hashing for histopathology image classification and retrieval. *Pattern Recognition*, 81:14–22.

Souid, A. et al. (2023). Improving diagnosis accuracy with an intelligent image retrieval system for lung pathologies detection: a features extractor approach. *Scientific Reports*, 13(1):16619.

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

Wickstrøm, K. K. et al. (2023). A clinically motivated self-supervised approach for content-based image retrieval of ct liver images. *Computerized Medical Imaging and Graphics*, 107:102239.

Yang, P. et al. (2020). A deep metric learning approach for histopathological image retrieval. *Methods*, 179:14–25. Interpretable machine learning in bioinformatics.

Zheng, Y. et al. (2022). Encoding histopathology whole slide images with location-aware graphs for diagnostically relevant regions retrieval. *Medical Image Analysis*, 76:102308.

Zhou, J. et al. (2022). ibot: Image bert pre-training with online tokenizer.