Toward linear representations of foundation models for histopathology image retrieval

Matheus Santos*, Igor Borja*, David Lima*, Grinaldo Oliveira*, Ângelo Duarte†,
Washington Santos‡ and Luciano Oliveira*‡
*Intelligent Vision Research Lab, Institute of Computing, Federal University of Bahia
Email: matheusaugusto.igorborja.davidlima,grinaldooliveira,lrebouca{@ufba.br}
†High-Performance Computing Lab, Department of Technology, State University of Feira de Santana
Email: angeloduarte@uefs.br
‡Structural and Molecular Pathology Lab, Gonçalo Moniz Institute, Fundação Oswaldo Cruz
Email: wluis@bahia.fiocruz.br

2025 38th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI) | 979-8-3315-8951-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/SIBGRAPI67909.2025.11223348

***Abstract*—While domain-specific foundation models have accelerated advancements by leveraging large-scale histopathological datasets, they often struggle to generalize across diverse clinical scenarios and datasets. In this paper, we address this limitation with a linear-prototype framework: A 128-D projection head trained by a modified prototypical loss on a few labelled slides. Evaluated on proprietary glomerular biopsies, dermoscopic skin cancer, and ovarian-cancer WSIs, our proposal approach consistently boosts image retrieval. Across nine backbones (UNI/UNI2-h, Phikon/Phikon-v2, Virchow2, DINO/DINOv2, ViT, and a ResNet-50 baseline), few-shot tuning raises mAP by up to 15 percentage points. These gains show that our lightweight, prototype-based layer can reconcile the breadth of foundation pre-training with the depth of task-specific discrimination, enabling improvement on medical-image retrieval across diverse pathologies.**

I. Introduction

Whole-slide imaging has generated digital-pathology repositories containing millions of gigapixel slides, storage, annotation, and analysis workflows [1]. Content-based medical-image retrieval (CBMIR) eases this burden by encoding each slide as a high-dimensional feature vector and rapidly returning visually or semantically similar cases [2]. CBMIR extracts texture, colour, and morphological features from biopsy images, ranks them, typically by cosine or learned similarity, and returns slides that share lesion patterns. These matched examples provide pathologists with precedent in ambiguous cases, improving diagnostic confidence and supporting data-driven decisions [1].

Early CBMIR studies adapted handcrafted colour-, texture-, and shape-based descriptors from generic image retrieval, but these fixed features perform poorly on histopathology, where subtle intra-class variation is common [3], [4]. Deep encoders (*e.g.* autoencoders, siamese/triplet networks, and supervised contrastive models) learn more discriminative embeddings yet introduce two new bottlenecks: Very high-dimensional feature vectors that inflate storage and latency, and a heavy reliance on large, expert-labeled datasets [1], [5]. Current work therefore seeks compact, label-efficient representations, *e.g.* deep hashing [6], [7] and self-supervised frameworks such as CLIP [8], SwAV [9], and DINO [10]. Despite progress, accurately embedding certain tissue types remains difficult because their diagnostic cues are subtle and highly variable.

Whole-slide retrieval systems cut gigapixel slides into tiles, identify diagnostically relevant regions of interest in whole slides images (WSIs), and index those patches for similarity search [7], [11], [12]. Chen et al. [11] replace handcrafted features with a self-supervised vector quantisation variational autoencoder (VQ-VAE) whose discrete codebook yields semantically meaningful embeddings. Evaluated on 8,035 WSIs spanning 37 tumour subtypes from Brigham and Women's Hospital, the method reaches mAP@5 of 90% for some cancers but falls to 20% for others, averaging 50.5%. The large spread reflects class imbalance and fine-grained variations in texture, morphology, and stain, emphasising that even advanced self-supervised models still struggle with the inherent heterogeneity of histopathology data.

A. Contributions

Here we examine how well foundation models extrapolate to unseen histopathology data and whether a lightweight, prototype-based transfer learning scheme can further improve retrieval accuracy. We introduce a supervised framework that appends a 128-dimensional projection head to a frozen foundation model and trains this head with a prototypical loss [13]. For each class, a prototype is computed as the centroid of its support embeddings, and the loss function pulls every query embedding toward its own prototype while pushing it away from all other prototypes in Euclidean space. The result is a compact, linearly comparable feature space that preserves class structure yet enables fast similarity search.

We evaluated the framework on three retrieval datasets using three backbone types: histopathology-specific foundation models, generic foundation models, and a ResNet baseline. Models trained on histopathology consistently outperformed ImageNet-based counterparts, though the generic DINOv2 remained competitive. Our framework further restructures foundation-model embeddings into a nearly linearly separable space, facilitating more efficient similarity computations. The consistent improvements observed across datasets and back-

979-8-3315-8951-6/25/$31.00 ©2025 IEEE

Authorized licensed use limited to: IFBA - INSTITUTO FEDERAL DA BAHIA. Downloaded on November 26,2025 at 13:24:05 UTC from IEEE Xplore. Restrictions apply.

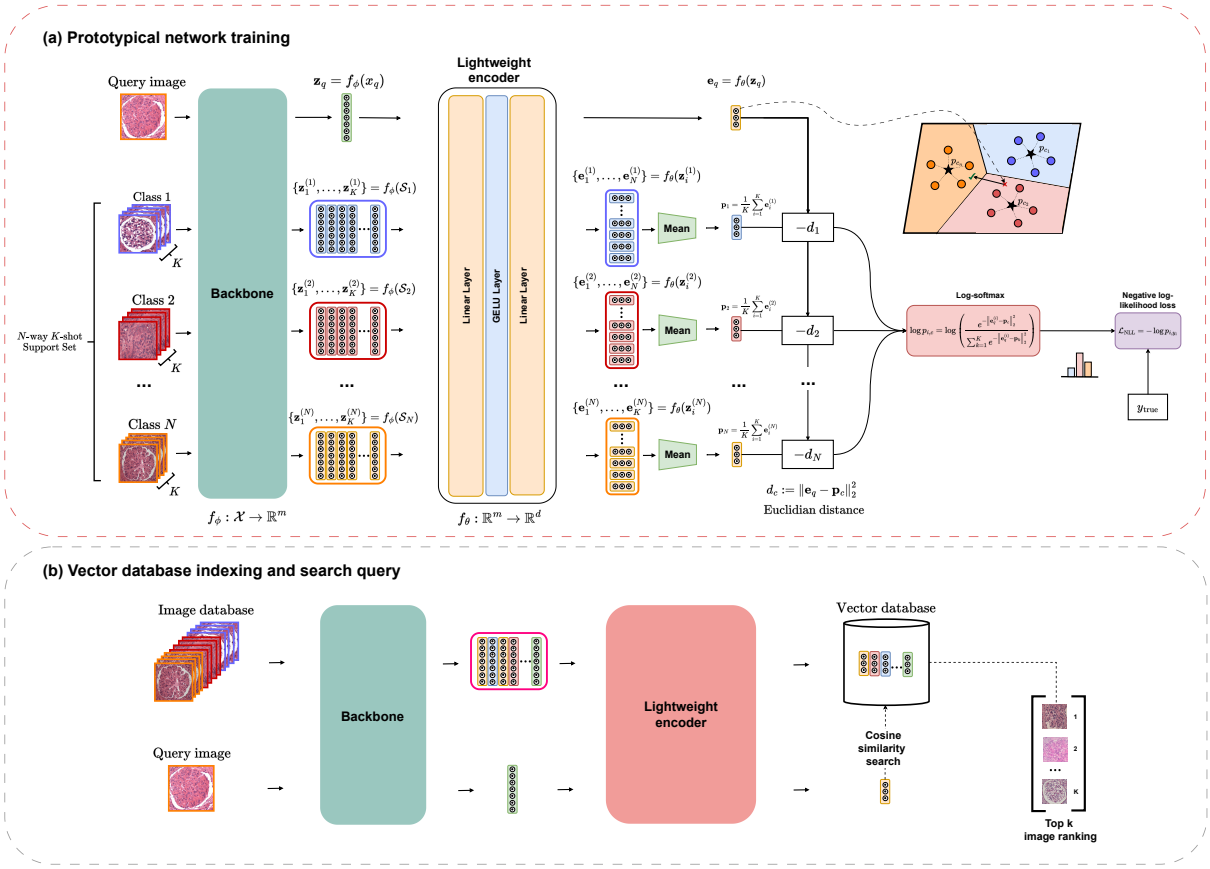


Fig. 1: Overview of prototypical image retrieval. (a) Training: the frozen backbone is followed by a lightweight encoder (two linear layers + GELU). For each class c, K samples are averaged into a prototype $\mathbf{p}_c$. A query image $\mathbf{e}_q$ is compared to all prototypes via Euclidean distance, converted to log-probabilities with log-softmax, and optimized with NLL loss. (b) Retrieval: the backbone + encoder generate embeddings for all images to form a database. A query embedding is compared using cosine similarity, and the top-k matches are ranked.

TABLE I: Summary of models trained on ImageNet and histopathology datasets.

Model	Dataset	Learning Method	Architecture	#Params	Emb. Dim
UNI [14]	>100M pathology tiles	Self-Supervised (DINOv2)	ViT-L/16	307M	1024
UNI2-h [14]	>200M pathology tiles	Self-Supervised (DINOv2+iBOT)	ViT-H/14-reg8	632M	1536
Virchow2 [15]	1.5M H&E stained WSIs	Self-Supervised (Modified DINOv2)	ViT-H/14	632M	1280
Phikon [16]	43M pathology tiles	Self-Supervised (iBOT)	ViT-B/16	86M	768
Phikon-v2 [17]	460M pathology tiles	Self-Supervised (DINOv2)	ViT-L/16	307M	1024
DINO [10]	ImageNet-1K	Self-Supervised (DINO)	ViT-S/16	22M	384
DINOv2 [18]	LVD-142M; 142M images	Self-Supervised (DINOv2)	ViT-L/14	304M	1024
ViT [19]	ImageNet-21K; ImageNet-1K	Supervised	ViT-B/16	86M	768
ResNet50 A1 [20]	ImageNet-1K	Supervised	ResNet-50	25.6M	2048

bones demonstrate that this lightweight prototypical approach effectively enhances the latent discriminative capacity of foundation models.

II. Exploring Linear Embedding Representation to Improve Histopathology Image Retrieval

A. Foundation models

Foundation models evaluated in this study fall into two groups, generic and pathology-specific, all of them built on ViTs unless stated otherwise. Table I summarizes the main characteristics of the models studied here.

UNI adopts the DINOv2 [18] self-supervised protocol and is pre-trained on MASS-100K, a corpus of aproximately 100,000 H&E WSIs; its larger sibling UNI2-h [14] uses a ViT-h/14-reg8 [19] backbone and an even broader slide collection, but we note that both variants underperform on glomerular tissue, which is under-represented in the pretraining data. Phikon [16] employs masked-image modeling (iBOT [21]) on in-domain slides, and Phikon-v2 scales both data and capacity while switching to a DINOv2 schedule, improving feature quality for downstream tasks [16], [17]. Virchow2 [15] further enlarges the training set to 3.1M WSIs and introduces histology-aware

Authorized licensed use limited to: IFBA - INSTITUTO FEDERAL DA BAHIA. Downloaded on November 26,2025 at 13:24:05 UTC from IEEE Xplore. Restrictions apply.

architectural tweaks, jointly exploring data and model scaling in computational pathology [15].

For generic baselines, we included ViT models pre-trained on ImageNet-21K [22] and two self-supervised counterparts: DINO [10], which learns via student–teacher distillation, and its successor DINOv2, which improves data scale, model depth and training stability [10], [18]. Although not a foundation model in scope, ResNet-50 A1 [20], a CNN trained on ImageNet-1K serves as a strong convolutional baseline.

B. Proposed framework

Figure 1 outlines the complete workflow of the proposed method. In (a), the few-shot prototypical training stage is illustrated. A frozen foundation backbone $f_\phi : \mathcal{X} \to \mathbb{R}^m$ embeds a support set consisting of N classes with K examples each as well as a held-out query image. The resulting m-dimensional vectors are passed through a lightweight encoder composed of two linear layers separated by GELU activation, producing compact d-dimensional embeddings. Class prototypes are computed as the mean of the support embeddings, and the squared Euclidean distance between the query and each prototype feeds a soft-max to yield class probabilities; the network is optimised with a negative log-likelihood loss. In (b), the retrieval phase is showed. All gallery images are first embedded by the frozen backbone and the trained lightweight encoder, then stored in a vector database. A query image is processed in the same way, and cosine similarity is used to rank gallery vectors, returning the top-k most similar slides.

The problem of image retrieval via feature extraction is formally defined as follows: Given a support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{NK}$, composed of K labeled examples per class from N classes, and a query set $\mathcal{Q} = \{q_j\}_{j=1}^{M}$, composed of M randomly selected image instances per batch, the objective is to learn a lightweight embedding function $f_\theta : \mathbb{R}^m \to \mathbb{R}^d$ that maps features generated from a pre-trained encoder $f_\phi : \mathcal{X} \to \mathbb{R}^m$, where $\mathcal{X}$ is the image space, into a discriminative d-dimensional embedding space. The overall embedding for a given input image $x \in \mathcal{X}$ is thus obtained by a composition of both encoders:

$$e = f_\theta(f_\phi(x)) \,. \tag{1}$$

The lightweight encoder f_θ is trained using the labeled support set to produce embeddings that allow effective few-shot classification. For each class $c \in \{1, \ldots, N\}$, a prototype vector $\mathbf{p}_c \in \mathbb{R}^d$ is computed as the mean of the embeddings of the support examples belonging to that class:

$$\mathbf{p}_c = \frac{1}{|\mathcal{S}_c|} \sum_{(x_i, y_i) \in \mathcal{S}_c} f_\theta(f_\phi(x_i)) \,, \tag{2}$$

where $\mathcal{S}_c = \{(x_i, y_i) \in \mathcal{S} \mid y_i = c\}$.

For each query sample $x_q \in \mathcal{Q}$, its embedding $f_\theta(f_\phi(x_q))$ is compared to the class prototypes using squared Euclidean distance, given by

$$d(f_\theta(f_\phi(x_q)), \mathbf{p}_c) = \|f_\theta(f_\phi(x_q)) - \mathbf{p}_c\|_2^2 \,. \tag{3}$$

To enable interpretation as classification scores, distances are transformed into negatives before applying the softmax. The model is trained by minimizing the negative log-likelihood of the true class:

$$\mathcal{L}(x_q, y_q) = -\log \frac{\exp(-d(f_\theta(f_\phi(x_q)), \mathbf{p}_{y_q}))}{\sum_{c'=1}^{N} \exp(-d(f_\theta(f_\phi(x_q)), \mathbf{p}_{c'}))} \,. \tag{4}$$

While the training objective is based on few-shot classification, the ultimate goal of the model is image retrieval: Learning a representation space where semantically similar images are close together. The prototype-based clustering mechanism used during training imposes a strong structural prior, encouraging the formation of well-defined, semantically coherent clusters in the embedding space. This facilitates retrieval by enabling efficient similarity comparisons based on simple distance metrics, such as Euclidean distance or cosine similarity.

After training, the composite encoder $f_\theta(f_\phi(\cdot))$ is utilized in the subsequent two stages, as illustrated in Figure 1(b). First, during vector database indexing, the model processes a fixed dataset $\mathcal{C} = \{x_j\}_{j=1}^{J}$, where J denotes the total number of reference images. Each image x_j is encoded into a latent representation, forming the set of database embeddings stored in a vector list. Second, in the search stage, a query image x_q is embedded via $f_\theta(f_\phi(x_q))$, and similarity-based retrieval is performed by computing the cosine similarity between e_q and each database vector e_j as

$$\text{sim}(e_q, e_j) = \frac{e_q \cdot e_j}{\|e_q\| \, \|e_j\|}, \quad \forall j \in \{1, \ldots, J\} \,. \tag{5}$$

Unlike the original prototypical network architecture, which employs four convolutional blocks (each comprising a 64-filter 3×3 convolution, batch normalization, a ReLU activation, and 2×2 max-pooling [13]), our proposed framework leverages a pre-trained encoder, typically a vision foundation transformer-based model, which remains frozen during training. This is followed by a lightweight trainable encoder consisting of a linear layer projecting to 512 dimensions, a GELU activation, and a subsequent linear layer reducing the embedding to 128 dimensions. The GELU activation function was selected due to its differentiable nature, which supports stable gradient flow and efficient optimization [23].

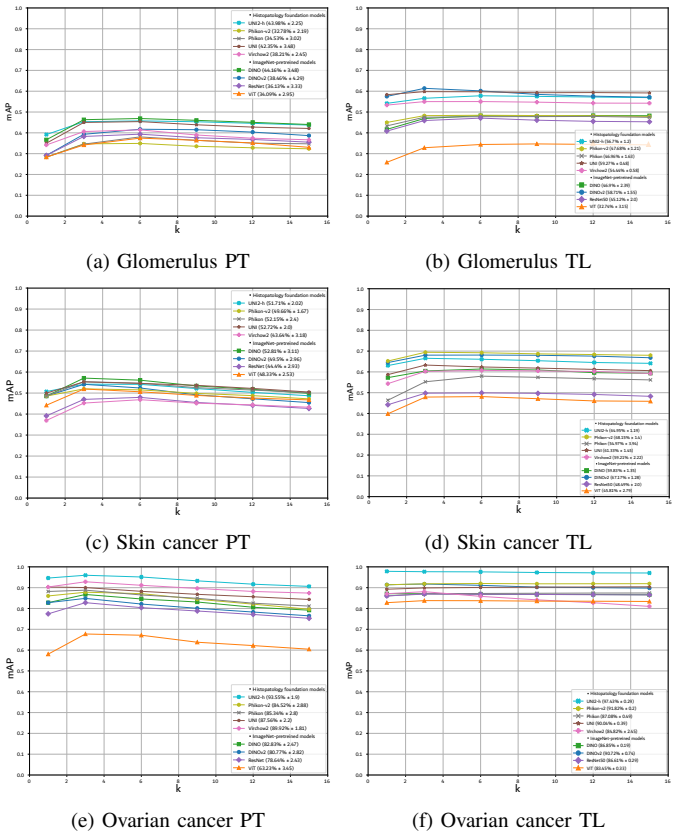
Our lightweight encoder generates embeddings with dimensions different from those of the benchmark backbones (384–2048). This design reflects deliberate trade-offs between representational capacity, computational efficiency, and storage, enabling faster inference and reduced memory usage while preserving sufficient fidelity for retrieval.

III. Materials and Methods

A. Datasets

We assessed the performance of our proposed approach on three datasets representing distinct imaging domains: A private glomerulus, the public ovarian cancer [24], and the

Authorized licensed use limited to: IFBA - INSTITUTO FEDERAL DA BAHIA. Downloaded on November 26,2025 at 13:24:05 UTC from IEEE Xplore. Restrictions apply.



(a) Glomerulus PT (b) Glomerulus TL

(c) Skin cancer PT (d) Skin cancer TL

(e) Ovarian cancer PT (f) Ovarian cancer TL

Fig. 2: Mean average precision (mAP) at rank-k for performance comparison of 9 pre-trained (PT) and transfer learning (TL) models, along with a ResNet baseline.

HAM10000 dermatoscopic skin cancer [25]. The glomerulus dataset consists of 7,582 training images and 116 test images of biopsy patches of renal tissue, categorized into five glomerular lesion subtypes and one normal class. The ovarian cancer dataset includes 405 training and 93 test histopathological images representing four ovarian cancer subtypes, along with non-cancerous tissue samples. The HAM10000 dataset comprises 6,906 training and 138 test dermatoscopic images from multiple sources, covering common pigmented skin lesions.

Query sets were built by drawing class-balanced images that were disjoint from the training and test galleries. For the private glomerulus corpus, these images came from a separate out-of-distribution collection, whereas for HAM10000 and ovarian cancer, they were randomly sampled from unused portions of the public datasets.

B. Implementation details

All fine-tuning and evaluation pipelines were implemented in PyTorch and executed on a single NVIDIA RTX 3060 GPU (12 GB). For each foundation model, the projection head was fine-tuned for 100 epochs using the Adam optimizer with a learning rate of 1×10^{-3} and weight decay of 1×10^{-4}. Input images were resized to 224×224 pixels and normalized with ImageNet mean and standard deviation values. During training, data augmentation included horizontal and vertical flips, color jitter, Gaussian blur, random rotations, affine transformations, optical distortion, and coarse dropout. The support set size was fixed at $K = 5$ samples per class.

C. Evaluation protocol

To evaluate the effectiveness of few-shot learning for CBMIR in histopathology datasets, we adopted mAP@k as the primary performance metric. For each query, precision is computed based on the top-k retrieved items, with mAP averaged across all queries. We report results for k = 1, 3, 6, 9, 12, and 15.

Statistical significance between the best pre-trained (PT) and transfer learning (TL) models is assessed using a two-sided paired permutation test on mAP@k, with 10,000 iterations. The null hypothesis assumes no performance difference across paired queries. Confidence intervals and p-values are estimated via bootstrapping from the empirical null distribution.

Qualitative assessment of the learned feature space was performed by projecting embeddings onto the first three principal components, revealing clear inter-class separation.

D. Experimental results

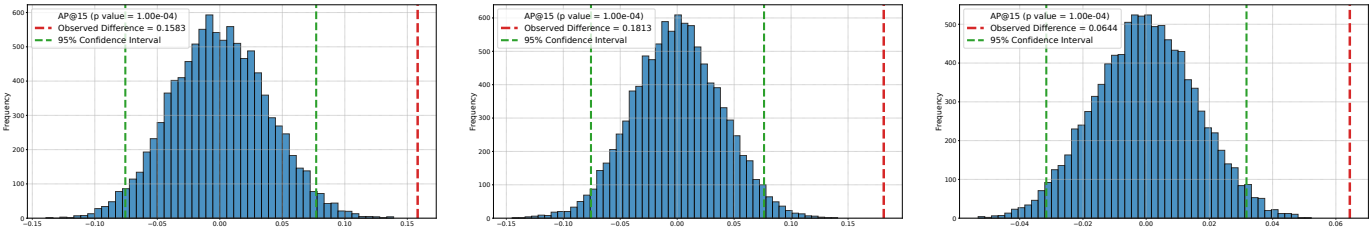
Among pre-trained models, glomerulus retrieval performance (mAP@k) ranged from 32.78% (Phikon-v2) to 44.16% (DINO), with UNI2-h achieving the second-best result at 43.98%. On skin cancer, performance varied between 43.64% (Virchow2) and 52.81% (DINO), with UNI and Phikon close behind. For ovarian cancer, PT models performed notably better, ranging from 63.23% (ViT) to 93.55% (UNI2-h), with Virchow2 (89.92%) and UNI (87.56%) also standing out. These PT baselines serve as reference points in Figure 2 (a), (c), and (e).

Transfer learning led to substantial gains across all datasets (Fig. 2 (b), (d), and (f)). On glomerulus, the best scores came from UNI (59.27%), DINOv2 (58.71%), and UNI2-h (56.70%), with improvements of +12.72 to +15.11 percentage points (pp) over their respective PT baselines. On skin cancer, Phikon-v2 reached 68.15%, followed by DINOv2 (67.17%) and UNI2-h (64.95%), yielding gains of over +13 pp. For ovarian cancer, which already had strong PT results, transfer learning pushed UNI2-h to 97.43% (+3.88 pp over PT), with Phikon-v2 and DINOv2 also above 90%. Averaging results across datasets, transfer learning consistently improved retrieval by over 10 pp, underscoring the benefit of task-specific adaptation.

Figure 3 depicts permutation tests comparing the best PT and transfer-learned models per dataset, confirming that the improvements are statistically significant: Glomerulus (+15.83 pp), skin cancer (+18.13 pp), and ovarian cancer (+3.88 pp), all with $p = 1.00 \times 10^{-4}$ and narrow 95% confidence intervals.

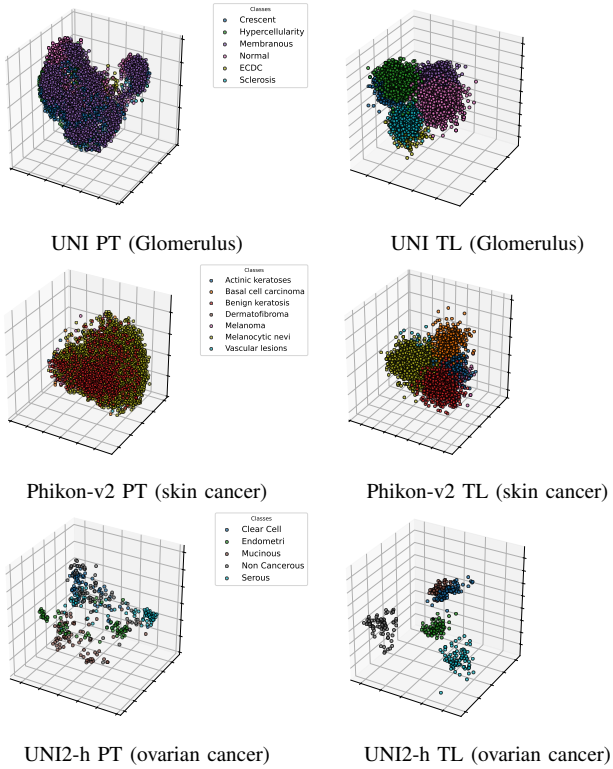
In Figure 4, PCA projections illustrate that TL reorganizes the embedding space into tighter, more discriminative class clusters, enhancing both intra-class compactness and inter-class separability. The projections in Fig. 4 correspond to the best-performing TL model in the retrieval benchmark

Authorized licensed use limited to: IFBA - INSTITUTO FEDERAL DA BAHIA. Downloaded on November 26,2025 at 13:24:05 UTC from IEEE Xplore. Restrictions apply.



(a) UNI TL x DINO PT (glomerulo) (b) Phikon-v2 TL × DINO PT (skin cancer) (c) UNI2-h TL x UNI2-h PT (ovarian cancer)

Fig. 3: Permutation test comparing transfer-learned (TL) and zero-shot pre-trained (PT) models. Blue histograms show the bootstrap distribution of the mean mAP gain (TL – PT); green ticks mark the 95% confidence interval (CI); the red line is the observed gain. In all datasets the observed value falls outside the CI ($p = 1.00 \times 10^{-4}$), confirming a significant TL advantage.



UNI PT (Glomerulus) UNI TL (Glomerulus)

Phikon-v2 PT (skin cancer) Phikon-v2 TL (skin cancer)

UNI2-h PT (ovarian cancer) UNI2-h TL (ovarian cancer)

Fig. 4: PCA visualization of image embeddings. Each line of plots displays embeddings from best TL models of Fig. 2 and its PT model counterpart. The improved cluster compactness and separation observed after TL suggest enhanced discriminative capability.

(Figure 2), highlighting how the embedding structure improves after the TL procedure.

IV. Discussion

In few-shot retrieval, domain-specific foundation models such as UNI, Phikon, and Virchow2—despite being trained on large-scale histopathology datasets—did not consistently outperform DINO, a general-purpose vision model. This was particularly evident in datasets with greater clinical heterogeneity, suggesting that these models still struggle to generalize across institutions with unseen staining protocols or acquisition conditions. Conversely, their notably strong performance on the ovarian cancer dataset may reflect distributional alignment with pretraining samples or intrinsic visual regularities, such as distinct background colors and more class-separable morphological features. These results highlight the need for deeper examination of robustness and generalization boundaries in histopathology-specific foundation models across diverse clinical settings [26]. The glomerulus dataset exemplifies these challenges. Glomeruli are small, morphologically complex structures whose diagnostic interpretation relies on subtle textural cues and fine-grained spatial patterns that are difficult to capture—even for human experts [27]. Variability in staining compounds the difficulty: visually distinct lesions may overlap in color distributions under HE, while identical lesions can appear differently across sites due to differences in staining protocols and imaging equipment [28]. Such heterogeneity complicates both human and machine-based analysis. A parallel phenomenon is observed in dermoscopic skin cancer data. Although these lesions are generally larger and more visually distinct, intra- and inter-class color distributions tend to be relatively homogeneous. As a result, lesion shape and texture become the primary discriminative features, placing high demands on the feature extraction capacity of retrieval models. The inability of pathology-specific foundation models to consistently surpass a generic model like DINO suggests that current pretraining objectives have not adequately captured the breadth of real-world clinical image distributions. This underscores the importance of designing foundation models that explicitly address domain shift and cross-site variability in histopathology datasets [26]. Finally, while all backbones benefited from our proposed training scheme, ViT-based foundation models consistently outperformed the convolutional baseline (ResNet-50). This advantage can be attributed in part to their architectural capacity: ViTs contain more trainable parameters and are better suited to capture long-range dependencies than CNNs. Whereas CNNs rely on strong inductive biases such as spatial locality and translation invariance, ViTs employ weaker inductive biases, making them more flexible but also more dependent on large-scale pretraining [29]. This flexibility, when paired with extensive pretraining, enables ViTs to learn rich global representations, further explaining

Authorized licensed use limited to: IFBA - INSTITUTO FEDERAL DA BAHIA. Downloaded on November 26,2025 at 13:24:05 UTC from IEEE Xplore. Restrictions apply.

their superior performance in our experiments.

V. Concluding remarks

We showed that augmenting frozen vision foundation models with a 128-dimensional, prototype-trained head markedly sharpens their embedding geometry and boosts histopathology-image retrieval accuracy. Although domain-specific pre-training provides a strong baseline, its effectiveness varies with the target cohort; lightweight, few-shot adaptation reliably closes the gap, raising mAP by up to 15 pp. The method therefore offers a practical path for translating large, general-purpose or pathology-oriented models into high-performance retrieval systems tailored to specific clinical datasets.

Future work will focused on threefold paths: (i) adaptive sampling to stabilize very large backbones, (ii) continual prototypical updates so the index can evolve as new classes appear, (iii) joint optimization with approximate-nearest-neighbour indices to accelerate large-scale slide retrieval in routine pathology workflows and (iv) inclusion of additional retrieval baselines such as supervised hashing and metric-learning methods for a broader performance comparison.

Acknowledgment

This work was sponsored by Terumo Foundation from Japan. David Lima is sponsored by UFBA under the grant no. 51802. Angelo Duarte is sponsored by FAPESB and UEFS, under the grants PET 0017/2024 and FINAPESQ 115/2024. Luciano Oliveira and Washington LC dos-Santos are sponsored by CNPq under grants 301789/2025-8 and 406141/2023.

References

[1] A. Kumar, J. Kim, W. Cai, M. Fulham, and D. Feng, "Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data," *Journal of Digital Imaging*, vol. 26, pp. 1025–1039, 2013.

[2] W. Chen, Y. Liu, W. Wang, E. M. Bakker, T. Georgiou, P. Fieguth, L. Liu, and M. S. Lew, "Deep learning for instance retrieval: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7270–7292, 2022.

[3] X. Li, J. Yang, and J. Ma, "Recent developments of content-based image retrieval (cbir)," *Neurocomputing*, vol. 452, pp. 675–689, 2021.

[4] S. R. Dubey, "A decade survey of content-based image retrieval using deep learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2687–2704, 2021.

[5] Y. Cai, Y. Li, C. Qiu, J. Ma, and X. Gao, "Medical image retrieval based on convolutional neural network and supervised hashing," *IEEE Access*, vol. 7, pp. 51 877–51 885, 2019.

[6] M. Sapkota, X. Shi, F. Xing, and L. Yang, "Deep convolutional hashing for low-dimensional binary embedding of histopathological images," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 805–816, 2018.

[7] S. Kalra, H. R. Tizhoosh, C. Choi, S. Shah, P. Diamandis, C. J. Campbell, and L. Pantanowitz, "Yottixel - an image search engine for large archives of histopathology whole slide images," *Medical Image Analysis*, vol. 65, pp. 1–12, 2020.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PmLR, 2021, pp. 8748–8763.

[9] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.

[10] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660.

[11] C. Chen, M. Y. Lu, D. F. 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*, vol. 6, no. 12, pp. 1420–1434, 2022.

[12] 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*, vol. 83, 2023.

[13] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *CoRR*, 2017. [Online]. Available: http://arxiv.org/abs/1703.05175

[14] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 850–862, 2024.

[15] E. Zimmermann, E. Vorontsov, J. Viret, A. Casson, M. Zelechowski, G. Shaikovski, N. Tenenholtz, J. Hall, D. Klimstra, R. Yousfi *et al.*, "Virchow2: scaling self-supervised mixed magnification models in pathology," *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2408.00738

[16] A. Filiot, R. Ghermi, A. Olivier, P. Jacob, L. Fidon, A. Camara, A. Mac Kain, C. Saillard, and J.-B. Schiratti, "Scaling self-supervised learning for histopathology with masked image modeling," *MedRxiv*, pp. 2023–07, 2023.

[17] A. Filiot, P. Jacob, A. Mac Kain, and C. Saillard, "Phikon-v2, a large and public feature extractor for biomarker prediction," *arXiv*. [Online]. Available: https://arxiv.org/abs/2409.09173

[18] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: learning robust visual features without supervision," *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2304.07193

[19] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[20] R. Wightman, H. Touvron, and H. Jégou, "Resnet strikes back: An improved training procedure in timm," *arXiv*, 2021. [Online]. Available: https://arxiv.org/abs/2110.00476

[21] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "ibot: Image bert pre-training with online tokenizer," *arXiv*, 2021. [Online]. Available: https://arxiv.org/abs/2111.07832

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: transformers for image recognition at scale," *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[23] M. Lee, "Mathematical analysis and performance evaluation of the gelu activation function in deep learning," *Journal of Mathematics*, vol. 2023, no. 1, p. 4229924, 2023.

[24] K. Kasture, "Ovarian cancer & subtypes," Mendeley Data, Tech. Rep., 2021.

[25] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, pp. 1–9, 2018.

[26] E. D. de Jong, E. Marcus, and J. Teuwen, "Current pathology foundation models are unrobust to medical center differences," *arXiv*, 2025. [Online]. Available: https://arxiv.org/abs/2501.18055

[27] C.-A. Weis, J. N. Bindzus, J. Voigt, M. Runz, S. Hertjens, M. M. Gaida, Z. V. Popovic, and S. Porubsky, "Assessment of glomerular morphological patterns by deep learning algorithms," *Journal of Nephrology*, vol. 35, no. 2, pp. 417–427, 2022.

[28] L. Barisoni, K. J. Lafata, S. M. Hewitt, A. Madabhushi, and U. G. Balis, "Digital pathology and computational image analysis in nephropathology," *Nature Reviews Nephrology*, vol. 16, no. 11, pp. 669–685, 2020.

[29] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 116–12 128, 2021.

Authorized licensed use limited to: IFBA - INSTITUTO FEDERAL DA BAHIA. Downloaded on November 26,2025 at 13:24:05 UTC from IEEE Xplore. Restrictions apply.