Domain-Aware Self-Supervised Learning Across Digital Pathology Datasets Using Vision Transformers

Aneesh Chatrathi

Dartmouth Health

ABSTRACT

- Motivation: Self-supervised learning (SSL) enables representation learning without manual labels, but digital pathology datasets are heterogeneous and naive multi-dataset training risks feature collapse.
- Approach: We apply a DINO-style teacher–student framework with a lightweight specimen token that encodes dataset identity, balancing domain-specific separation with cross-domain alignment.
- Data: Experiments were conducted on thyroid cell images, bile duct patches, and cervical cancer patches.
- Results: Embeddings showed strong clustering (silhouette ≈ 0.99, ARI/NMI = 1.0), perfect linear probe accuracy (100%), and clear dataset-level groupings in UMAP

INTRODUCTION

- Challenge in pathology: Digital pathology datasets vary widely in scale, staining, and morphology, making it difficult for models to generalize across domains. Self-supervised learning (SSL) has shown promise, but models trained on a single dataset often fail to transfer due to strong domain effects.
- Limitations of prior work: Most SSL studies evaluate only one dataset at a time. When multiple datasets are pooled, models risk domain collapse, overfitting to larger sources and losing biologically meaningful variation.
- Our approach: We adapt DINO SSL with a specimen token that encodes dataset identity, enabling representations that maintain domain-specific structure while aligning across datasets. We evaluate this framework on three digital pathology datasets (thyroid, bile duct, and cervical cancer) to test whether unified training can capture both separation and consistency.

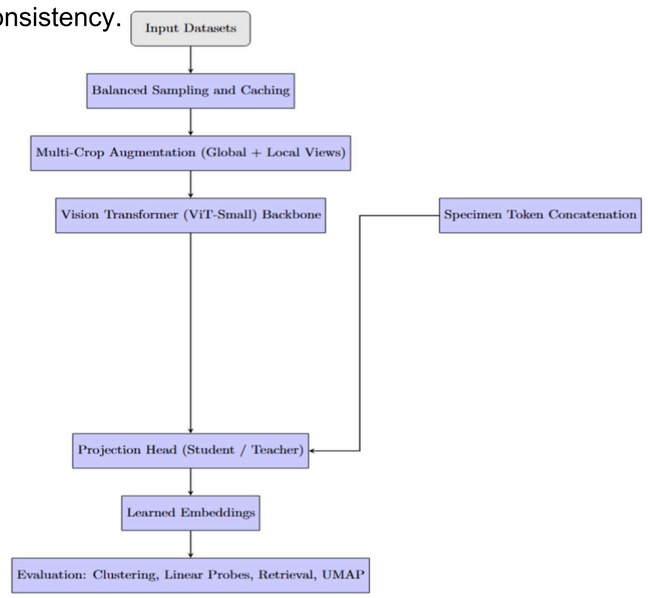
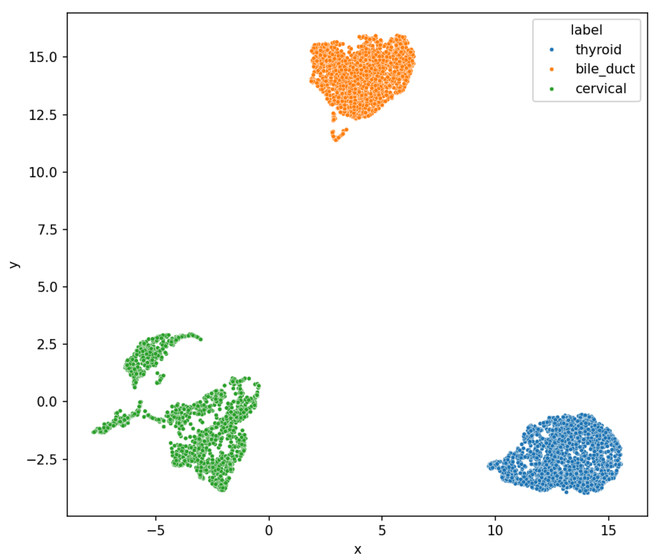


Figure 2: Overall workflow of the self-supervised learning framework. Input datasets are balanced and cached, processed through multi-crop augmentations, and encoded by a ViT-Small backbone. A domain-aware specimen token is concatenated before the projection head. Learned embeddings are evaluated through clustering, linear probes, retrieval, and UMAP visualization.

METHODOLOGY

- Datasets: Three digital pathology datasets — thyroid, bile duct, and cervical cancer. Each training epoch used balanced sampling (30k images per dataset, 90k total), with 2k per dataset reserved for evaluation.
- Preprocessing & Augmentation: All images were resized to 224×224 and normalized. We applied DINO-style multi-crop augmentation with 2 global crops (224×224) and 6 local crops (96×96). Augmentations included color jitter, Gaussian blur, flips, and random crops to promote scale and stain invariance.
- Model & Training: A Vision Transformer (ViT-S/16) backbone was used. A learnable “specimen token” (one per dataset) was concatenated with the [CLS] embedding to condition learning on dataset identity. Training followed the DINO teacher–student distillation framework with exponential moving average (EMA) updates, AdamW optimizer, cosine learning rate schedule, mixed-precision, and was run on NVIDIA V100/L40S GPUs.
- Evaluation: Embeddings were assessed with k-means clustering (silhouette, ARI, NMI), linear probes (multinomial logistic regression across domains), UMAP visualizations, and cross-domain transfer (training probes on one dataset, testing on another).



Uniform Manifold Approximation and Projection (UMAP) is a dimensionality reduction method that projects high-dimensional embeddings into 2D or 3D for visualization. It preserves both local neighborhoods and global structure, making it useful for showing whether datasets form distinct clusters or overlap in the learned representation space.

RESULTS

Table 1: Quantitative evaluation of learned embeddings across three digital pathology datasets. Metrics include linear probe accuracy, clustering indices, centroid similarity, and retrieval performance.

Evaluation Metric	**Result**
Linear Probe Performance	
Domain-level accuracy (thyroid, bile duct, cervical)	100%
Clustering Metrics	
Silhouette score	0.99
Davies–Bouldin index	0.13
Calinski–Harabasz index	188,862
Adjusted Rand Index (ARI)	1.0
Normalized Mutual Information (NMI)	1.0
Embedding Similarity	
Centroid cosine similarity (across datasets)	> 0.99
Retrieval Analysis	
Precision@k	≈ 1.0
Recall@k	0.0025

- Linear probes reached 100% accuracy within each dataset, showing perfectly separable embeddings.
- Clustering metrics confirmed strong structure (silhouette ≈ 0.99, ARI/NMI = 1.0).
- UMAP visualization revealed three compact, non-overlapping clusters aligned with dataset origin.
- Centroid similarity was high (>0.99), indicating global alignment across domains.
- Retrieval analysis showed high precision but low recall, reflecting compact domain-specific clusters.

CONCLUSION

- A DINO-based self-supervised framework with a Vision Transformer backbone produced robust embeddings across three digital pathology datasets.
- Representations were domain-separable (perfect linear probes, strong clustering) yet globally aligned (high centroid similarity).
- The specimen token offered a simple mechanism for incorporating dataset identity without collapsing representations.
- This framework provides a strong baseline for multi-dataset representation learning, supporting future work on generalizable pathology AI.

edit.