Rotation-Agnostic Image Representation Learning for Digital Pathology

Saghir Alfasly Abubakr Shafique Peyman Nejat Jibran Khan Areej Alsaafin Ghazal Alabtah
H.R.Tizhoosh*
KIMIA Lab, Department of Artificial Intelligence & Informatics, Mayo Clinic, Rochester, MN, USA
{alfasly.saghir, tizhoosh.hamid}@mayo.edu

Abstract

This paper addresses complex challenges in histopathological image analysis through three key contributions. Firstly, it introduces a fast patch selection method, FPS, *for whole-slide image (WSI) analysis, significantly reducing computational cost while maintaining accuracy. Secondly, it presents* PathDino, *a lightweight histopathology feature extractor with a minimal configuration of five Transformer blocks and only ≈ 9 million parameters, markedly fewer than alternatives. Thirdly, it introduces a rotation-agnostic representation learning paradigm using self-supervised learning, effectively mitigating overfitting. We also show that our compact model outperforms existing state-of-the-art histopathology-specific vision transformers on 12 diverse datasets, including both internal datasets spanning four sites (breast, liver, skin, and colorectal) and seven public datasets (PANDA, CAMELYON16, BRACS, DigestPath, Kather, PanNuke, and WSSS4LUAD). Notably, even with a training dataset of ≈6 million histopathology patches from The Cancer Genome Atlas (TCGA), our approach demonstrates an average 8.5% improvement in patch-level majority vote performance. These contributions provide a robust framework for enhancing image analysis in digital pathology, rigorously validated through extensive evaluation.* [1]

1. Introduction

The advent of whole slide image (WSI) scanning in digital pathology has revolutionized the research in computational pathology [1–3]. While digital pathology enables both researchers and clinicians to enjoy the ease of access to the WSIs, processing and storing these gigapixel images are still quite challenging.

Motivation: Large image size and scarce or lack of patch-level labels (annotations) pose two main challenges in WSI analysis [4]. As a result, most state-of-the-art methods adopt Multi-instance Learning (MIL) with weak supervision [5–13]. While these approaches may eliminate the need for pixel-level annotations, MIL significantly increases computational loads and potentially lowers the quality of results compared to fully supervised approaches. While some attempts have been made to select representative patches [5, 6, 14, 15], many such methods remain computationally intensive, leaving the desire for efficient, accurate solutions an unmet need.

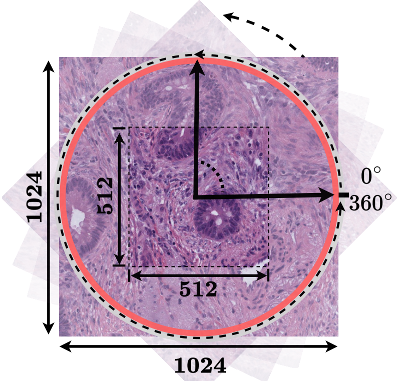


Figure 1. `HistoRotate`. A 360° rotation augmentation for training models on histopathology images. Unlike training on natural images where the rotation may change the context of the visual data, rotating a histopathology image improves the learning process for discriminative embedding learning.

The field of image analysis in digital pathology has predominantly adopted deep models designed for natural image analysis without further customization to the field [16–19]. While showing good performance on natural image analysis, pre-trained deep models may not fully exploit the unique characteristics of histopathology images. Furthermore, most current training recipes for histopathological embedding learning adopt conventional training and common augmentation techniques for natural images [18]. However, histopathology images have arguably very different features compared to natural images and even radiology images. This gap motivated us to design an improved training approach for histopathology images.

*Corresponding author

[1] The project page: https://KimiaLabMayo.github.io/PathDino-Page/

Contributions: We present a two-fold solution that encompasses selective patching and robust feature extraction. First, we propose a fast patch selection **`FPS,`** a "*divide & conquer*" algorithm that is capable of identifying a compact and yet highly representative subset of patches for analysis. This algorithm has been meticulously tuned to balance computational efficiency and diagnostic utility. Secondly, we introduce **`PathDino`** a lightweight histopathology-specific transformer consisting of just five small vision transformer blocks, customized and finely tuned to the nuances of histopathological images. It not only exhibits superior performance but also effectively reduces susceptibility to overfitting. We also propose **`HistoRotate`**, a seamless 360° rotation augmentation technique designed specifically for training histopathology models. The incorporation of this augmentation technique with the proposed lightweight histopathology-specific transformer results in a significant enhancement of embedding quality and effectively mitigates overfitting. Our model is rigorously validated through extensive evaluation on multiple datasets, showing both computational efficiency and superior performance. Overall, our **key contributions** are as follows:

- **Fast Patch Selection**: A novel and efficient patch selection mechanism curates a compact, spatially diverse subset of patches from WSI, reducing computational overhead while maintaining representational fidelity.
- **PathDino**: A lightweight histopathology-specific Vision Transformer with only 5 transformer blocks, totaling 9 million parameters, offering reduced susceptibility to overfitting.
- **Rotation-Agnostic Representation Training**: We propose **`HistoRotate`**, a 360° rotation augmentation technique designed for training histopathological image analysis models. Unlike natural images, rotating histopathological patches maintain the general context while enhancing embedding learning for improved reliability.
- **Extensive Evaluation**: Rigorous validation through comprehensive experiments across eleven datasets, demonstrating competitive to superior performance compared to existing state-of-the-art methods.

2. Related Work

WSI Patching. WSI patching is a fundamental phase in WSI analysis pipelines, although it has received limited attention in the field. Many methods employ a brute force tiling approach, where the entire WSI is divided into thousands of patches [7, 20–22], typically utilized with weakly supervised training methods like multi-instance learning [5–13]. This approach is often employed when only WSI-level labels are available, as in TCGA, instead of pixel-level annotations [23–25]. However, brute force patch processing proves very challenging in practice due to the immense computational costs and potential training instability.

Clustering-Based Patch Selection. This approach aims to address patch quality by selecting representative patches but introduces new degrees of freedom such as number of clusters. It includes both *Independent Patching Phase*, where only one method in the literature, namely Yottixel's mosaic [14], follows this independent approach. Yottixel employs a two-stage clustering process, first based on color (stain) features and then on connected regions, creating a patch set with visual and spatial diversity. At the end, it uses a guided sampling inside each cluster. It stands as the only independent patching method adaptable to various WSI analysis pipelines. In contrast, the *Integrated Patching Phase* tightly couples patching methods with specific WSI analysis methods, limiting their applicability to other uses. For example, in [5], patch clustering is performed for each WSI into k clusters, integrated with Multi-instance learning. Similarly, in [6], a similar approach is used, clustering the entire dataset patches into a few clusters and matching specific WSI patches with cluster centroids, effectively assigning patches with pseudo labels.

While embedded clustering methods prove inflexible and unsuitable for integration into other WSI pipelines, approaches based on clustering, although enhancing the quality of the chosen patch set, concurrently introduce an additional layer of parameters and variability to the overall process. To address these challenges, we propose a new fast patch selection method that avoids the brute-force and multi-variable clustering approaches. Crucially, our **`FPS`** aligns with the independent patching phase, exemplified by Yottixel, enhancing adaptability for WSI analysis pipelines while greatly improving efficiency.

Vision Transformer in Histopathology. A prevalent trend in histopathological image analysis is the adaptation of mainstream vision transformers, especially ViT (Vision Transformer) [26, 27]. Many existing models are essentially fine-tuned versions of ViT [16–19], often overlooking the unique characteristics of histopathological images compared to natural images, leading to issues such as overfitting since ViTs are known to be data-hungry [28]. In contrast, our comparably compact ViT architecture **`PathDino`** tailored for histopathological images, achieving better results while mitigating overfitting.

Self-Supervised Learning in Digital Pathology. Self-supervised learning has gained popularity in digital pathology due to its independence from annotated histopathological images, making it possible to leverage large datasets [29, 30]. However, most self-supervised learning approaches are primarily developed for natural image analysis [29–34]. Applying these methods directly to histopathological embedding learning without considering domain-specific differences can lead to suboptimal performance. Recent studies underscore the value of domain-specific pre-training for transferability. Domain-specific self-supervised learning methods are also shown to significantly enhance performance in medical imaging tasks [16, 35–41]. Furthermore, BYOL, SimSiam, and SimCLR frameworks have been employed for image classification and patch retrieval in histopathology [16, 22, 38, 39].

Recent studies have shown promising results in enhancing model performance for downstream tasks in medical imaging through transfer learning and domain-specific self-supervised learning methods. Kang *et al.* in [18] conducted a comprehensive benchmarking study on self-supervised representation learning in histopathology images, evaluating several methods on a dataset of 32.6M patches (19M from TCGA[2] and 13.6M from TULIP which is an private dataset), including SwAV, MoCoV2, Barlow Twins, and DinoV1 [18]. Hierarchical Image Pyramid Transformer (HIPT) is a self-supervised Transformer trained on TCGA patches using Dino-based self-supervised training, whereas TransPath is a self-supervised model trained on TCGA and PAIP patches through contrastive learning [16, 17]. iBOT-Path [19], a vision transformer, was trained on 40M histopathology patches from TCGA using the self-supervised iBOT framework [33]. Additionally, models like BiomedCLIP [42] and PLIP [43], trained with image-text contrastive learning on the biomedical PMC-15M dataset and the histopathology dataset OpenPath, respectively. Virchow [44], a Transformer-based model with 632 million parameters, was trained using DinoV2-based self-supervised learning on 1.5M internal WSIs [30].

Our work differs from previous methods in the following aspects: ***WSI Patching***: Our FPS method offers superior efficiency compared to [14] without the need for patch clustering, while still maintaining competitive accuracy. ***Histopathology-specific ViT Structure***: Our PathDino is a lightweight ViT that contains only 5 small transformer blocks for effective histopathological image analysis. ***Training Recipe***: Our training recipe features *HistoRotate* augmentation that applies 360° rotation leading to rotation-invariant embedding learning.

3. Proposed Method

3.1. FPS: Fast Patch Selection

In this section, we introduce a method for the systematic selection of representative patches from WSIs for computational pathology. The algorithm aims to cater to both the diversity and relevance of the tissue structure, thus capturing the inherent complexity and heterogeneity of tissue slides as illustrated in Fig. 2-A.

Preprocessing. Given a WSI, I, with dimensions $W \times H$, a thumbnail image, T, with dimensions $w \times h$ is generated. A tissue mask, M, is obtained through binary thresholding.

Density-Proportional Selection with Kernel Density Estimation (KDE) The contours extracted [45] from the tissue mask are denoted by C, where $C = \{c_1, c_2, \dots, c_n\}$. For each contour c_i, a bounding box is defined as $R_i = [x, y, w, h]$. A set of potential patch locations, P, is constructed as follows:

$$P = \bigcup_{i=1}^{n} \{(x, y) \mid x \in [R_{i,x}, R_{i,x} + R_{i,w} - r_w], \\ y \in [R_{i,y}, R_{i,y} + R_{i,h} - r_h]\}, \tag{1}$$

where, r_w and r_h are the dimensions of the patches in the mask space. Subsequently, density-proportional KDE is employed to generate the set S of selected patches:

$$S = \text{KDE}(P, n_s), \tag{2}$$

where n_s is the predefined number of patches to be selected. Utilizing the KDE to approximate the probability density function $f(x)$ over the set P is performed as follows:

$$f(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right), \tag{3}$$

where K is the kernel function, N is the total number of points in P, and h is the bandwidth (*i.e.*, the width of the smoothing kernel).

Density-Proportional Sampling. In accordance with the density map generated by KDE, points are sampled proportionally to their density values:

$$p(x) = \frac{f(x)}{\sum_{x \in P} f(x)}. \tag{4}$$

A random sample S consisting of n_s points is extracted from P based on the probability density function $p(x)$:

$$S = \text{Rand}(P, p(x), n_s). \tag{5}$$

The resulting set S conforms to the spatial density characteristics of the tissue structures in the slide, thus capturing the tissue heterogeneity.

Spatial Constraints. To avoid oversampling from densely packed regions, a minimum Euclidean distance, e_{min}, is enforced between any two selected patches s_i and s_j:

$$\forall s_i, s_j \in S, \sqrt{(s_{i,x} - s_{j,x})^2 + (s_{i,y} - s_{j,y})^2} \geq e_{\text{min}}. \tag{6}$$

Finally, the selected patches are mapped back to the WSI coordinates at high magnification for downstream analyses. Each patch location $(x, y) \in S$ is scaled to its corresponding location in I using the ratio between W and w, as well as H and h. The patches are extracted and stored for subsequent analyses.

[2] https://portal.gdc.cancer.gov/

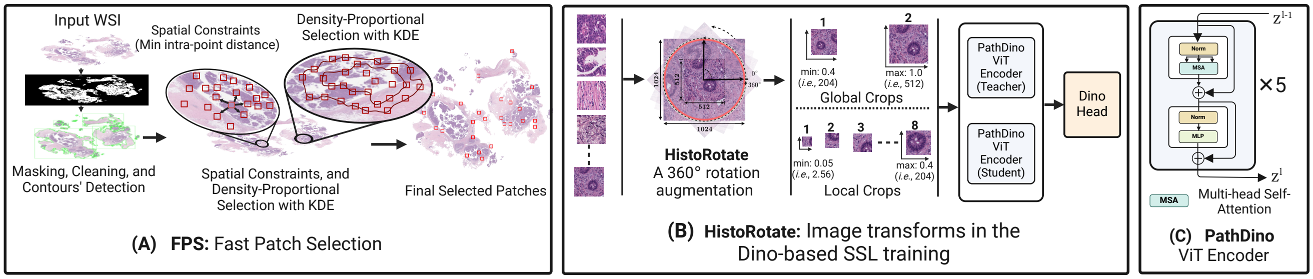


Figure 2. The WSI Analysis Pipeline. (A) The fast patch selection method, FPS, selects a set of representative patches while preserving spatial distribution. (B) HistoRotate is a 360° rotation augmentation for histopathology model training, enhancing learning without contextual information alteration. (C) PathDino is a compact histopathology Transformer with five small vision transformer blocks and ≈9 million parameters, significantly leaner than alternatives.

3.2. HistoRotate: Rotation-Agnostic Training

In addressing the unique challenges in tissue image analysis, we introduce a new self-supervised training recipe that incorporates a rotation-agnostic scheme as depicted in Fig. 1, designed to enhance the quality of the learned representations by incorporating various angular rotations of the image during the training process. Let I denote an input image, and θ denote a randomly selected angle from a predefined set Θ. The rotation operation $\mathcal{R}$ is formally defined as:

$$\mathcal{R}(I, \theta) = I_\theta. \tag{7}$$

In our implementation, two types of rotations are considered:
(a) *Random Continuous Rotation*: θ is sampled from a continuous uniform distribution over the range $[0, 360]$ degrees.

$$\theta \sim \mathcal{U}(0, 360). \tag{8}$$

(b) *Random Discrete Rotation*: θ is selected from the set $\Theta = \{90, 180, 270, 360\}$. Each image undergoes a cropping operation $\mathcal{C}$ before and after the rotation, followed by a resizing operation $\mathcal{S}$ to generate a transformed image I'.

$$I' = \mathcal{S}(\mathcal{C}(\mathcal{R}(I, \theta))). \tag{9}$$

***HistoRotate* with Dino Framework**. As depicted in Fig. 2-B, we applied these transformations on two types of image crops used in Dino framework [29]: **Global Crops**: Images are cropped and resized to a scale s sampled from $\mathcal{U}(0.4, 1)$. **Local Crops**: Images are cropped and resized to a smaller scale s' sampled from $\mathcal{U}(0.05, 0.4)$. In the final data augmentation pipeline, we generate a set $\mathcal{I}$ of transformed images from each original image I:

$$\mathcal{I} = \{I'_1, I'_2, \ldots, I'_n\}. \tag{10}$$

The proposed rotation-agnostic representation learning scheme yields a significant advantage in obtaining more comprehensive and robust tissue image representations.

3.3. PathDino: A Histopathology-specific Vision Transformer

We introduce PathDino, a shallow and compact vision transformer designed for histopathological image analysis. This model is lightweight and less prone to overfitting. It has an embedding size of $d = 384$, 6 attention heads, and a patch size of 16×16 for input images $X \in \mathbb{R}^{H \times W \times C}$. We evaluate two input resolutions: $H = W = 512$ (PathDino-512) and $H = W = 224$ (PathDino-224). PathDino encoder comprises a total of $L = 5$ blocks. Each block consists of a multi-head self-attention (MSA) layer, LayerNorm (LN), and a multilayer perceptron (MLP):

$$\mathbf{z}_i^\ell = \mathrm{MLP}(\mathrm{LN}(\mathrm{MSA}(\mathbf{z}_i^{\ell-1}))) + \mathbf{z}_i^{\ell-1}, \tag{11}$$

where $\mathbf{z}_i \in \mathbb{R}^d$, $\ell = 1, \cdots, L$, and $i = 1, \cdots, N$ and N here represents the total input transformer patches. Fig. 2-C visualizes PathDino encoder structure, whereas Fig. 3 visually compares PathDino's performance, FLOPs, and parameter count with those of its counterparts. PathDino contains ≈9M parameters, significantly fewer than ViT-s (21M) used by DinoSSLPath [18] and HIPT [17], as well as the ViT-b (85M) used by iBOT-Path [19].

4. Experiment Setup

Hardware: All experiments have been conducted on a Dell PowerEdge XE8545 server with 4× NVIDIA A100-SXM4-80GB and 2× AMD EPYC 7413 CPUs, 1023 GB RAM.
PathDino Pretraining Dataset. We extracted a total of $6,087,558$ patches from $11,765$ diagnostic TCGA WSIs. Specifically, $3,969,490$ patches have a 1024×1024 dimension, while $2,118,068$ patches have a 512×512 dimension. The extraction was conducted at a $20\times$ magnification level, with a patch tissue area threshold of 90%.
PathDino Pretraining Details. All pretraining and evaluation processes are conducted using the *Pytorch* deep learning library and Python. We adapt DINO [29] framework in

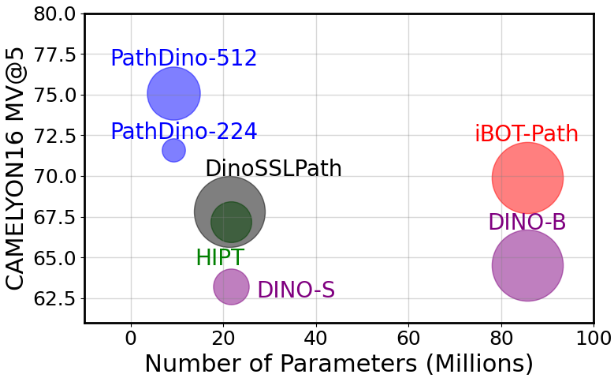


Figure 3. PathDino vs. its counterparts. Number of parameters (millions) vs. the patch-level retrieval with macro average F-1 score of majority vote (MV@5) on CAMELYON16 dataset. The bubble size represents the FLOPs.

which we integrated our augmentation method *HistoRotate* to be applied to each cropped image portion of the internal and global crops. In the pretraining phase of our study, we utilized $\approx$ 6M patches from TCGA. To ensure high-quality data selection, a tissue threshold of 90% was employed to filter the patches without enough tissue coverage from the WSIs. Our pretraining approach follows self-supervised learning, implemented on top of the DINO framework. We employed two sets of crops, comprising 2 global crops and 8 local crops. Our pretraining efforts resulted in the development of two distinct models: PathDino-224, trained on 224×224 cropped images obtained solely from the $2,118,068$ patches with size 512×512. We utilized a batch size of 384 with the AdamW optimizer and a learning rate of 0.0001 for 30 epochs. Meanwhile, PathDino-512, a model with 512×512 dimensions trained on the entire $6,087,558$ patches for 27 epochs employing a batch size of 192 and the AdamW optimizer with an initial learning rate of 0.0005.

Downstream Datasets. ***Private Skin:*** Contains 660 WSIs primarily capturing cutaneous squamous cell carcinoma (cSCC) biopsies in various differentiation stages including a class of normal skin biopsy. Demographic features indicate a median patient age of 77, with females making up 35% of the dataset. ***Private Liver:*** Includes 150 WSIs of alcoholic steatohepatitis (ASH), 158 WSIs of non-alcoholic steatohepatitis (NASH), and 18 WSIs of normal cases predominantly sourced from liver biopsies. ***Private CRC:*** Features 209 WSIs, categorized into Cancer Adjacent Polyp (CAP), Non-recurrent Polyp (POP-NR), and Recurrent Polyp (POP-R) classes. ***Private Breast:*** Consists of 73 WSIs classified into 16 tumor subtypes and one class of normal tissue, encapsulating a variety of pathological conditions such as Adenoid Cystic Carcinoma (ACC), Ductal Carcinoma In Situ (DCIS), among others. ***PANDA [23]:*** A public dataset of $12,625$ WSIs of prostate biopsies stained with H&E, collected from diverse international sites for comprehensive evaluation. ***CAMELYON16 [25]:*** Provides 399 meticulously annotated WSIs of lymph node sections collected from breast cancer patients across two hospitals in the Netherlands. ***BRACS [24]:*** Encompasses 547 WSIs from 189 patients, annotated into seven distinct lesion subtypes by board-certified pathologists. ***DigestPath [46]:*** Comprises two specialized datasets for diagnosing gastrointestinal histopathology features: the Signet Ring Cell Detection Dataset (SRC) and the Colonoscopy Tissue Segmentation and Classification Dataset (TSCC). ***PanNuke [47]:*** A semi-automatically generated nuclei instance segmentation and classification dataset containing exhaustive nuclei labels across 19 different tissue types. ***Kather-7K [48]:*** Features $7,180$ non-overlapping image patches sourced from 50 patients with colorectal adenocarcinoma, serving as an ideal validation set for model evaluation. ***WSSS4LUAD [49]:*** Specifically built for segmentation tasks in lung adenocarcinoma histopathology, including over $10,091$ patch-level annotations. Additional details for each dataset are available in the Suppl-Tables [S6, S7].

Evaluation Metrics. For the evaluation of WSI-level and patch-level retrievals, we used Top-1, the majority vote among Top-3 (MV@3), and the majority vote among Top-5 (MV@5) metrics within the leave-one-out evaluation scheme. To assess the patch classification task, we trained a linear classifier using the extracted feature embeddings and computed accuracy and macro average F-1 score. Embedding variances were analyzed using Principal Component Analysis, as illustrated in Figure 5. Additionally, the quality of the Vision Transformer (ViT) is visually assessed using activation maps, as shown in Figure 4. An extensive evaluation, both qualitative and quantitative, is presented in the subsequent sections and the supplementary file.

5. Experimental Results

5.1. FPS Effectiveness

Table 1 provides an in-depth comparative assessment between Yottixel's mosaic and our FPS patching method across 3 private and 3 public histopathology datasets, utilizing BiomedCLIP [50], PLIP [43], and PathDino as backbones. Across internal datasets, FPS consistently exhibits competitive to superior performance. For example, in Private-Breast dataset, FPS achieves a top-1 accuracy of 58% with PLIP and 68% with PathDino, outperforming Yottixel's corresponding values of 55% and 63%. In Private-Liver dataset, FPS integrated with PathDino achieves an 83% top-1 accuracy, markedly higher than Yottixel's accuracy of 81%. This trend is corroborated in the Private-Skin and Private-CRC datasets, where FPS surpasses Yottixel's mosaic in all metrics, most notably achiev-

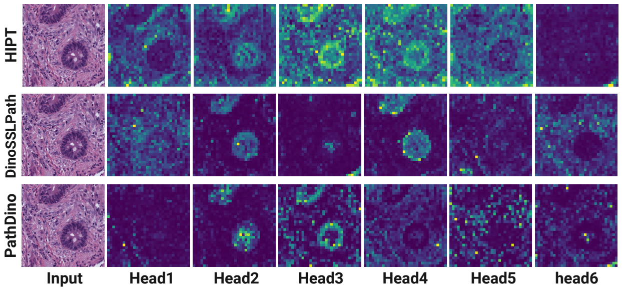


Figure 4. Attention Visualization. When visualizing attention maps, our PathDino transformer outperforms HIPT-small and DinoSSLPath, despite being trained on a smaller dataset of 6M TCGA patches. In contrast, DinoSSLPath and HIPT were trained on much larger datasets, with 19 million and 104 million TCGA patches, respectively.

Table 1. Performance accuracy of the proposed FPS against Yottixel's mosaic using BiomedCLIP, PLIP and PathDino backbones.

	Dataset		BiomedCLIP [50]		PLIP [43]		PathDino	
			Yottixel	FPS	Yottixel	FPS	Yottixel	FPS
Internal Data	Private-Breast	Top 1	**47**	**47**	55	**58**	63	**68**
	Private-Liver	Top 1	70	**74**	70	**73**	81	**83**
		MV@3	75	**77**	**76**	74	86	**86**
		MV@5	74	**77**	73	**76**	**87**	85
	Private-Skin	Top 1	68	**75**	72	**75**	**79**	78
		MV@3	73	**78**	77	**79**	**81**	80
		MV@5	76	**78**	80	**82**	81	**82**
	Private-CRC	Top 1	55	**58**	60	**64**	57	**63**
		MV@3	60	**63**	61	**67**	60	**65**
		MV@5	59	**65**	62	**69**	61	**65**
Public Data	PANDA [23]	Top 1	33	**34**	53	**56**	**59**	58
		MV@3	**36**	**36**	53	**55**	58	**58**
		MV@5	**38**	**38**	53	**54**	**58**	56
	CAMELYON16 [25]	Top 1	60	**61**	70	**73**	**76**	73
		MV@3	58	**67**	71	**77**	77	**78**
		MV@5	64	**69**	70	**75**	**78**	77
	BRACS [24]	Top 1	**56**	55	**62**	60	**65**	64
		MV@3	58	**62**	**64**	63	65	**66**
		MV@5	59	**61**	**66**	64	66	**67**

ing an MV@5 of 82% in Private-Skin with PLIP, but lower performance on Top1 and MV@3. The results in public datasets demonstrate on par performance rather than superiority. For example, in the PANDA dataset, FPS, when paired with PLIP, records a top-1 accuracy of 56%, which is 3% higher than Yottixel's mosaic. In summary, the empirical evidence overwhelmingly supports the efficacy of FPS as compared to Yottixel's mosaic. More results are reported in Suppl-Tables [S2, S3].

5.2. FPS Efficiency

Table 2 elucidates the computational efficiency and processing capabilities of both patching methods when paired with the PathDino backbone. Remarkably, FPS demonstrates higher computational efficiency in most scenarios. For instance, FPS processes Private-Breast and Private-Skin datasets in significantly less time, requiring only 13.1 and 132.0 minutes in total, respectively, as opposed to Yottixel's 20.4 and 171.3 minutes. Additionally, FPS succeeds in processing more WSIs with fewer failures; in the PANDA dataset, FPS processes 138 missed WSIs compared to Yottixel's 268. This efficiency extends to other datasets, such as Private-CRC and BRACS, where FPS outperforms Yottixel's mosaic in both speed and the number of processed WSIs. These empirical findings not only validate the robustness and efficacy of FPS but also its computational advantages, underscoring its suitability for large-scale, time-sensitive histopathological image analysis.

Table 2. Comparison of FPS against Yottixel's mosaic in terms of the dataset properties such as number of extracted patches and average processing speed. For fair comparison, both frameworks use PathDino as the backbone.

Dataset	# WSI	Extracted Patches		Patching Speed (m)		# missed WSI	
		Yottixel	FPS	Yottixel	FPS	Yottixel	FPS
Private-Breast	74	1,141	2,033	20.4	**13.1**	**1**	**1**
Private-Liver	326	2,974	8,297	**45.4**	64.6	**2**	3
Private-Skin	660	8,388	16,491	171.3	**132.0**	1	**0**
Private-CRC	209	4,619	6,068	79.4	**46.0**	**0**	**0**
PANDA	10617	87,451	112,763	251.9	**192.1**	268	**138**
CAMELYON16	129	2,864	3,870	84.5	**12.9**	1	**0**
BRACS	547	12,946	15,352	261.5	**117.7**	24	**12**

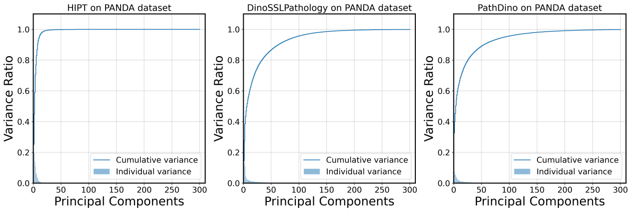


Figure 5. Embedding variance analysis of three selected Transformer-based histopathological feature extractors with the output vector size of 384 including HIPT, DinoSSLPath, and our PathDino on PANDA dataset [23].

Table 3. WSI-level top-1 accuracy using the proposed FPS patching method and "*median of minimum*" Euclidean distances as proposed in Yottixel [14].

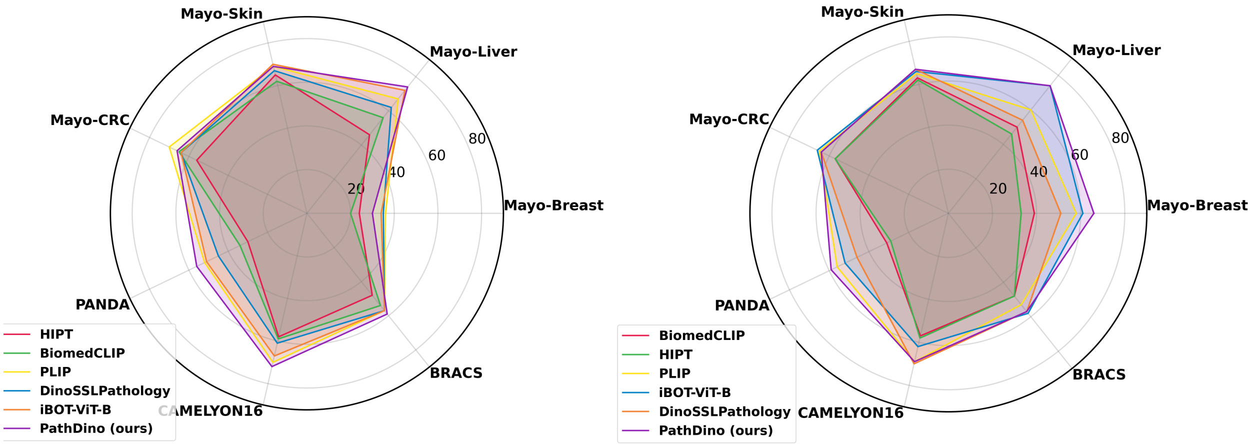
	Internal Datasets			Public Datasets		
	Breast	Liver	Skin	PANDA	CAMELYON16	BRACS
ResNet50 [51]	0.48	0.67	0.73	0.32	0.54	0.53
DenseNet121 [52]	0.48	0.64	0.69	0.30	0.67	0.52
EfficientNet-b3-288 [53]	0.41	0.66	0.73	0.32	0.59	0.55
EfficientNet-b5 [53]	0.51	0.71	0.71	0.37	0.57	0.54
ConvNext-b-224 [54]	0.56	0.75	0.74	0.34	0.62	0.58
ConvNext-xlarge [54]	0.56	0.76	0.74	0.35	0.61	0.58
ViT-b16-224 [26]	0.41	0.7	0.72	0.31	0.6	0.54
DinoV1-ViT-s16 [29]	0.48	0.71	0.74	0.36	0.67	0.6
DinoV1-ViT-b16 [29]	0.55	0.72	0.73	0.37	0.63	0.59
DinoV2-ViT-b14 [30]	0.53	0.71	0.72	0.31	0.61	0.51
CLIP - ViT-B/16 [55]	0.49	0.67	0.75	0.36	0.67	0.58
MuDiPath-ResNet50 [56]	0.44	0.7	0.72	0.35	0.63	0.51
MuDiPath-DenseNet-101 [56]	0.51	0.68	0.74	0.36	0.65	0.56
KimiaNet [57]	0.51	0.78	0.75	0.57	**0.76**	0.62
BiomedCLIP - [50]	0.47	0.74	0.75	0.34	0.61	0.55
HIPT-ViT-s16 [17]	0.44	0.68	0.73	0.32	0.62	0.52
PLIP [43]	0.58	0.73	0.75	0.56	0.73	0.60
iBOT-Path [19]	0.64	0.79	0.76	0.53	0.67	**0.64**
DinoSSLPathology-8 [18]	0.58	0.74	**0.78**	0.47	0.74	0.61
PathDino-224 (ours)	0.53	0.75	0.74	0.46	0.72	0.61
PathDino-512 (ours)	**0.68**	**0.83**	**0.78**	**0.58**	0.73	**0.64**

5.3. PathDino - WSI-Level Search

Table 3 highlights the performance of several feature extractors across various private and public datasets using the

Table 4. PathDino's performance, assessed for patch-level search accuracy and MV@5 macro average F-1 score, compared to various feature extractors. The lower-right section (grey values) indicates datasets that have been partially or fully included in the pretraining dataset TCGA.

			Internal Datasets								Public Datasets													
			Private-Breast		Private-Liver		Private-Skin		Private-CRC		PANDA [23]		CAMELYON16 [25]		BRACS [24]		DigestPath [46]		Kather [48]		PanNuke [47]		WSSS4LUAD [49]	
			Acc	MAF1	Acc	MAF1	Acc	MAF1	Acc	MAF1	Acc	MAF1	Acc	MAF1	Acc	MAF1	Acc	MAF1	Acc	MAF1	Acc	MAF1	Acc	MAF1
Pretrained on Natural Data	CNN-based	ResNet50 [51]	32.5	19.0	63.8	42.8	68.9	53.0	47.1	47.3	31.0	26.0	62.5	56.4	47.8	40.6	86.7	82.0	97.5	97	74.7	59	75.2	45.2
		DenseNet121 [52]	31.4	19.2	65.0	43.6	68.7	53.5	47.6	47.6	31.6	26.7	62.6	57.3	49	42.6	88.7	86	98.5	98.1	77.8	63.1	77.7	48.5
		EfficientNet-b3-288 [53]	29.7	16.6	64.5	46.7	67.4	51.8	46.5	46.7	30.8	25.8	61.5	55.4	49	42	89.8	87.5	96.1	95.5	70.8	53.3	78.0	48.1
		EfficientNet-b5 [53]	38.2	25.6	68.1	48.6	71.9	55.9	50.6	51.0	34.7	29.4	62.4	56.1	50.6	43.9	92.2	91	98.6	98.3	79.8	64.5	79.7	49.1
		ConvNext-b-224 [54]	39.7	28	68.3	50.1	72.6	58.2	48.7	48.8	33.2	28.2	64.5	59.8	49.9	43.4	92.2	90.7	99.2	99.1	85.6	73.9	83.2	52.3
		ConvNext-xlarge [54]	42.8	28.7	70.0	51.8	74.7	60.3	51.3	51.6	34.2	29.1	63.4	57.7	51.3	44.6	93.2	92	99.5	99.5	90.4	81.6	84.2	53.5
	Transformer	ViT-b16-224 [26]	29.6	16.7	67.7	49.2	71.7	55.5	46.8	46.9	32.1	26.8	62.0	55.8	49.1	42.2	89.3	80.7	98.2	97.8	79.4	67	70.4	51.3
		DinoV1-ViT-s16 [29]	36.6	25.0	70.3	49.6	71.4	56.6	49.9	50.2	34.1	29.7	63.2	57.3	51.3	44.6	92	90.1	99.5	99.4	89.8	81.3	83.9	53.4
		DinoV1-ViT-b16 [29]	38.1	27.2	71.3	52.1	72.2	57.8	50.8	51	34.7	30.1	64.5	58.4	51.3	44.4	91.9	90	99.7	99.6	91.5	83	84.7	54.2
		DinoV2-ViT-b14 [30]	31.8	20.9	68.4	48.7	69.8	54.4	48.1	48.3	31.4	26.4	60.2	53.3	50.0	42.6	89.8	86.5	98.6	98.4	76.6	64.5	76.1	66.5
		CLIP - ViT-B/16 [55]	36.4	26.8	69.4	49.8	72.7	57.7	52.3	52.6	35.8	31.0	62.8	56.7	52.5	45.5	90.0	87.8	98.4	98.2	79.1	63.7	79.2	48.6
Pret. On Histopathology Data	CNN-based	Barlow-Twins-ResNet50 [18]	50.8	37.5	76.0	55.5	72.2	56.7	56.1	56.9	46.0	43.5	63.9	58.0	54.8	47.1	95.2	94.4	99.7	99.6	91.8	85.3	86.2	54.9
		SwAV-ResNet50 [18]	50.2	37.5	77.4	60.1	74.2	59.6	56.2	56.9	45.0	42.1	68.6	63.2	55.8	48.4	95.3	94.7	99.6	99.5	90.6	82.5	82.8	51.5
		MoCoV2-ResNet50 [18]	51.9	37.5	76.7	57.9	72.9	56.3	54.6	55.3	45.2	42.3	65.0	58.9	54.6	47.4	94.7	94.0	99.7	99.6	90.8	83.7	84.6	53.7
		MuDiPath-ResNet50 [56]	32.5	20.9	68.0	47.2	71.5	55.6	47.0	47.2	31.8	27.0	62.1	57.0	49.0	42.0	89.4	87.7	98.9	98.5	80.6	68.9	81.0	50.7
		MuDiPath-DenseNet-101 [56]	36.6	25.9	69	47.5	72.0	56.2	49.4	49.8	33.3	28.8	62.3	56.4	50.5	43.5	91.6	89.3	99.4	99.2	88.8	79.7	82.8	52.5
		KimiaNet [57]	46.8	37.2	78.2	61.2	76.3	61.6	56.0	56.7	45.1	42.4	71.9	67.7	56.8	50.6	95.0	94.2	99.4	99.3	94.3	88.6	82.4	51.4
	Transformers	BiomedCLIP [50]	34.1	22.7	67.8	49.7	72.1	56.3	47.6	47.7	32.5	27.4	61.3	55.4	50.6	43.6	92.8	91.3	98.6	98.3	79.8	66.8	84.1	53.6
		HIPT-ViT-s16 [17]	37.8	25.0	70.6	50.3	71.5	56.3	49.2	49.4	33.8	28.9	67.2	62	50.1	43.2	89.3	87.5	98.7	98.3	88.6	78.2	81.0	50.5
		PLIP [43]	44.1	34.9	72.0	54.1	75.2	61.6	**57.8**	**58.4**	43.0	39.3	68.8	62.9	55.4	48.2	94.7	93.7	97.2	97.0	82.3	68.6	78.2	48.5
		iBOT-Path [19]	50.2	42.1	78.0	65.2	76.8	62.4	55.9	56.5	41.6	37.9	69.9	64.4	57.8	51.2	95.2	94.3	99.9	99.9	97.7	93.6	87.1	55.7
		DinoSSLPathology-8 [18]	47.1	36.3	77.0	59.7	76.1	61.4	56.0	56.6	39.8	35.3	67.8	60.8	56.0	49.0	95.7	95.2	99.9	99.9	96.6	92.2	88.1	56.7
		PathDino-224 (ours)	44.5	38.7	77.2	61.6	76.0	61.4	52.7	53.2	40.1	36.0	71.6	66.9	55.1	48.6	95.8	95.0	99.9	99.8	96.3	90.7	86.9	55.7
		PathDino-512 (ours)	**55.1**	**49.1**	**82.7**	**69.5**	**77.2**	**63.6**	57.4	58.1	**48.3**	**46.3**	**75.1**	**70.4**	**59.3**	**52.6**	**96.8**	**96.2**	99.9	99.9	96.6	91.1	86.7	55.4



(A) Patch-Level Retrieval MV@5 Macro Avg

(B) WSI-Level Retrieval Top-1 Macro Avg

Figure 6. Performance of selected Transformer-based histopathological feature extractors including HIPT, BiomedCLIP, PLIP, DinoSSLPath, iBOT, and PathDino. The performance is represented as the macro average of the $F1$ score for the MV@5: (A) the performance of patch-level retrieval, (B) the performance of WSI-level retrieval.

proposed FPS patching method and the *median of minimum* Euclidean distances proposed in Yottixel [14]. Across both private and public datasets, PathDino-512 demonstrates competitive to superior performance. PathDino-512 achieves an exceptional 83% top-1 accuracy in the dataset Private-Liver, outperforming other models like HIPT, and iBOT-Path (student), which attain 68% and 79%, respectively. Even in a difficult case like Private-Skin, PathDino-512 reaches a 78% top-1 accuracy, competing with DinoSSLPathology which provides 78%. Notably, in the public dataset PANDA, PathDino-512 achieves a 58% top-1 accuracy, significantly outperforming both CNN-based and Transformer-based models like HIPT which only reach 32%. The macro average $F1$ score also consistently favors PathDino-512. These empirical findings prove PathDino-512 is a robust and highly efficient model for WSI-level retrieval. More results for the macro average $F1$ score of Top1, MV@3, MV@5, along with accuracy of MV@3, and MV@5 are reported in Suppl-Tables S9, S11, S13, S10, and S12, respectively.

5.4. PathDino - Patch-Level Search

The results presented in Table 4 provide an extensive comparative analysis of models in patch-level histopathology image search. The standout performer is our proposed model, PathDino-512. The model not only outperforms others in terms of accuracy but also establishes new benchmarks in the macro average $F1$ score, a critical metric for robust evaluation. For private datasets such as Private-Breast and Private-Liver, PathDino-512 achieves

Table 5. 5-Fold Cross-Validation: Macro-F1 in Histopathology. Right side: TCGA-related datasets (see the Supplementary File).

		Internal Datasets				Public Datasets						
		Private-Breast	Private-Liver	Private-Skin	Private-CRC	PANDA [23]	CAMELYON16 [25]	BRACS [24]	DigestPath [46]	Kather [48]	PanNuke [47]	WSSS4LUAD [49]
Transformers	BiomedCLIP [50]	38.82±1.64	48.44±1.15	56.62±0.61	55.89±2.57	25.97±0.34	58.19±4.99	41.89±0.93	92.07±2.33	94.89±0.84	37.81±2.12	69.98±8.12
	HIPT-ViT-s16 [17]	43.08±6.27	59.31±5.08	59.47±2.85	48.69±4.62	25.65±1.38	61.56±4.25	42.02±8.82	84.66±7.17	96.81±0.69	42.17±3.80	64.26±7.50
	PLIP [43]	46.07±3.20	50.78±1.48	62.48±1.11	64.11±2.70	31.53±0.47	69.67±1.45	46.72±1.05	92.07±2.91	90.90±1.63	27.77±2.54	61.51±7.19
	iBOT-Path [19]	85.12±1.74	84.37±1.31	**73.09±0.39**	68.38 ±0.52	**32.95±0.86**	73.76±1.67	**56.52 ±1.96**	95.67±2.17	99.81±0.17	95.76±1.78	73.31±5.91
	DinoSSLPathology-8 [18]	77.59±2.17	74.25±4.56	66.98±1.00	58.17±4.77	28.75±2.31	70.61±1.81	46.42±5.59	94.50±1.91	99.68±0.11	86.17±2.61	76.30±9.60
	PathDino-224 (ours)	78.06±4.03	74.34±4.98	64.89±2.14	60.65±2.23	27.74±2.44	69.26±4.94	46.58±3.78	94.03±3.06	99.66±0.19	81.03±2.51	74.47±9.05
	PathDino-512 (ours)	**88.57±3.08**	**86.35±5.33**	71.36±1.64	**70.47±2.47**	32.08±2.57	**79.61±1.00**	52.59±3.21	**95.82±2.26**	99.65±0.11	84.79±3.14	72.69±7.60

the highest accuracy rates of 55.1% and 82.7%, respectively. More remarkably, it tops the macro average $F1$ score with 49.1% and 69.5% in the same datasets. These findings extend to public datasets like PANDA and CAMELYON16, where PathDino-512 records accuracy and macro average $F1$ scores of 48.3% and 46.3%, and 75.1% and 70.4%, respectively.

While it is important to note the strong performance of models like iBOT-Path and DinoSSLPathology, especially for public datasets, PathDino-512 consistently outperforms them across multiple metrics and datasets. We analyzed the patch embedding variance as shown in Fig. 5. We compare PathDino against HIPT and DinoSSLPathology as they have the same embedding size (*i.e.*, 384). Notably, PathDino capitalizes on an expanded set of components within the feature vector to accurately represent the inferred histopathology patch. Fig. 4 visually compares their attention performance in which PathDino shows better attentions.

5.5. PathDino - Patch-level 5-Fold Cross-Validation

In Table 5 detailing 5-fold cross-validation results, a thorough quantitative comparison of macro-averaged $F1$ scores is presented for an assortment of models across multiple private and public datasets. We only report the performance of histopathology Transformer-based models here. The detailed measurements of macro average $F1$ scores and accuracy values are available in Suppl-Tables S5, S4, respectively.

On the internal datasets like Private-Breast, Private-Liver, and Private-CRC, our proposed model, PathDino-512, achieves standout performance with $F1$ scores of 88.57±3.08, 86.35±5.33, and 70.47±2.47, respectively. These scores are markedly higher than the next best models, such as iBOT-Path, which reaches $F1$ scores of 85.12±1.74 in Private-Breast and 84.37 ± 1.31 in Private-Liver. In the realm of public datasets, PathDino-512, and iBOT-Path show competitive results where PathDino leads with an $F1$ score of 79.61 ± 1.00 in CAMELYON16, outperforming iBOT-Path, which scores 73.76 ± 1.67 in the same dataset. Interestingly, iBOT-Path excels in Private-Skin with an $F1$ score of 73.09 ± 0.39, the highest among all models for that specific dataset.

6. Conclusions

This paper presented a new approach to WSI analysis, addressing two pivotal challenges that have long stymied advancements in this field—computational efficiency and diagnostic fidelity. We introduced a *fast patch selection (FPS)* algorithm that reliably identifies a compact yet highly informative subset of patches, thereby significantly reducing computational overhead without compromising diagnostic inclusion. Additionally, we unveiled a new Transformer-based model structure for histopathological image analysis, *PathDino*, that only contains 5 small transformer blocks. Finally, we presented a rotation-agnostic self-supervised learning, *HistoRotate*, tailored for histopathological representation learning. Through training the proposed *PathDino* using the proposed *HistoRotate* and rigorously validating them with 12 diverse datasets, we showed that our lightweight transformer along with our training recipe effectively mitigates issues of overfitting that are prevalent in this domain. Our dual-pronged approach has demonstrated competitive to superior performance compared to the state-of-the-art methods.

Limitations: In contrast to natural images, magnification plays an important role in histopathological images. Our training dataset only included patches in $20X$ magnification from TCGA. Thus, more tuning for multi-resolution training may provide better results.

Broader Impacts: The proposed methods for whole slide image analysis have the potential to improve the diagnosis and prognosis of various diseases by providing accurate and reliable information on tissue morphology and cellular characteristics. With the widespread use of digital pathology workflows in clinical practice, these methods can reduce the workload and human errors of pathologists. Furthermore, quantifying tissue morphologies through accurate and valid image analysis method help with reducing intra- and inter-observer variability within the medical field. The proposed methods can also contribute to the advancement of histopathological image analysis by providing robust image representations.

Acknowledgments: The authors thank Mark Zarella, Wenchao Han, Sobhan Hemati, Nneka Comfere, Dennis Murphree, Chady Meroueh, Saba Yasir, Aaron Mangold, Lisa Boardman, Vijay H. Shah, and Joaquin J. Garcia for their valuable insights, discussions, and suggestions.

References

[1] Xintong Li, Chen Li, Md Mamunur Rahaman, Hongzan Sun, Xiaoqi Li, Jian Wu, Yudong Yao, and Marcin Grzegorzek. A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. *Artificial Intelligence Review*, 55(6):4809–4878, 2022. 1

[2] Neeta Kumar, Ruchika Gupta, and Sanjay Gupta. Whole slide imaging (wsi) in pathology: current perspectives and future directions. *Journal of digital imaging*, 33(4):1034–1040, 2020. 1

[3] Zizhao Zhang, Pingjun Chen, Mason McGough, Fuyong Xing, Chunbao Wang, Marilyn Bui, Yuanpu Xie, Manish Sapkota, Lei Cui, Jasreman Dhillon, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, 1(5):236–245, 2019. 1

[4] Liangrui Pan, Zhichao Feng, and Shaoliang Peng. A review of machine learning approaches, challenges and prospects for computational tumor pathology. *arXiv preprint arXiv:2206.01728*, 2022. 1

[5] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020. 1, 2

[6] Chensu Xie, Hassan Muhammad, Chad M Vanderbilt, Raul Caso, Dig Vijay Kumar Yarlagadda, Gabriele Campanella, and Thomas J Fuchs. Beyond classification: Whole slide tissue histopathology analysis by end-to-end part learning. In *Medical Imaging with Deep Learning*, pages 843–856. PMLR, 2020. 1, 2

[7] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021. 1, 2

[8] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022. 1, 2

[9] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19830–19839, 2023. 1, 2

[10] Conghao Xiong, Hao Chen, Joseph Sung, and Irwin King. Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. *arXiv preprint arXiv:2301.08125*, 2023. 1, 2

[11] Tsai Hor Chan, Fernando Julio Cendra, Lan Ma, Guosheng Yin, and Lequan Yu. Histopathology whole slide image analysis with heterogeneous graph representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15661–15670, 2023. 1, 2

[12] Honglin Li, Chenglu Zhu, Yunlong Zhang, Yuxuan Sun, Zhongyi Shui, Wenwei Kuang, Sunyi Zheng, and Lin Yang. Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7454–7463, 2023. 1, 2

[13] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4025, 2021. 1, 2

[14] Shivam Kalra, Hamid R Tizhoosh, Charles Choi, Sultaan Shah, Phedias Diamandis, Clinton JV Campbell, and Liron Pantanowitz. Yottixel–an image search engine for large archives of histopathology whole slide images. *Medical Image Analysis*, 65:101757, 2020. 1, 2, 3, 6, 7

[15] Xinliang Zhu, Jiawen Yao, and Junzhou Huang. Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 544–547, 2016. 1

[16] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 186–195. Springer, 2021. 1, 2, 3

[17] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 1, 2, 3, 4, 6, 7, 8

[18] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3344–3354, June 2023. 1, 2, 3, 4, 6, 7, 8

[19] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, pages 2023–07, 2023. 1, 2, 3, 4, 6, 7, 8

[20] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient

and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 2

[21] Yonghang Guan, Jun Zhang, Kuan Tian, Sen Yang, Pei Dong, Jinxi Xiang, Wei Yang, Junzhou Huang, Yuyao Zhang, and Xiao Han. Node-aligned graph convolutional network for whole-slide image representation and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18813–18823, 2022. 2

[22] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 2, 3

[23] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022. 2, 5, 6, 7, 8

[24] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, Maria Gabrani, Florinda Feroce, and Maria Frucci. BRACS: A Dataset for BReAst Carcinoma Subtyping in H&E Histology Images. *Database*, 2022:baac093, 10 2022. 2, 5, 6, 7, 8

[25] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 2, 5, 6, 7, 8

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 6, 7

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[29] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 4, 6, 7

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3, 6, 7

[31] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2

[32] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[33] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2, 3

[34] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 2

[35] Christos Matsoukas, Johan Fredin Haslum, Moein Sorkhei, Magnus Söderberg, and Kevin Smith. What makes transfer learning work for medical images: Feature reuse & other factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9225–9234, 2022. 3

[36] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3478–3488, 2021. 3

[37] Joseph Boyd, Mykola Liashuha, Eric Deutsch, Nikos Paragios, Stergios Christodoulidis, and Maria Vakalopoulou. Self-supervised representation learning using visual field expansion on digital pathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 639–647, 2021. 3

[38] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. 3

[39] Jacob Gildenblat and Eldad Klaiman. Self-supervised similarity learning for digital pathology. *arXiv preprint arXiv:1905.08139*, 2019. 3

[40] Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, pages 728–744. PMLR, 2021. 3

[41] Jiawei Yang, Hanbo Chen, Yuan Liang, Junzhou Huang, Lei He, and Jianhua Yao. Concl: Concept contrastive learning for dense prediction pre-training in pathology images. In

European Conference on Computer Vision, pages 523–539. Springer, 2022. 3

[42] Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2023. 3

[43] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, pages 1–10, 2023. 3, 5, 6, 7, 8

[44] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, et al. Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*, 2023. 3

[45] Satoshi Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing*, 30(1):32–46, 1985. 3

[46] Qian Da, Xiaodi Huang, Zhongyu Li, Yanfei Zuo, Chenbin Zhang, Jingxin Liu, Wen Chen, Jiahui Li, Dou Xu, Zhiqiang Hu, et al. Digestpath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Medical Image Analysis*, 80:102485, 2022. 5, 7, 8

[47] Jevgenij Gamper, Navid Alemi Koohbanani, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. Pannuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778*, 2020. 5, 7, 8

[48] Jakob Nikolas Kather, Frank Gerrit Zöllner, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Cleo-Aron Weis. Collection of textures in colorectal cancer histology. *Zenodo*, may 2016. 5, 7, 8

[49] Chu Han, Xipeng Pan, Lixu Yan, Huan Lin, Bingbing Li, Su Yao, Shanshan Lv, Zhenwei Shi, Jinhai Mai, Jiatai Lin, et al. Wsss4luad: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. *arXiv preprint arXiv:2204.06455*, 2022. 5, 7, 8

[50] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023. 5, 6, 7, 8

[51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[52] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 6, 7

[53] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6, 7

[54] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 6, 7

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6, 7

[56] Romain Mormont, Pierre Geurts, and Raphaël Marée. Multi-task pre-training of deep neural networks for digital pathology. *IEEE journal of biomedical and health informatics*, 25(2):412–421, 2020. 6, 7

[57] Abtin Riasatian, Morteza Babaie, Danial Maleki, Shivam Kalra, Mojtaba Valipour, Sobhan Hemati, Manit Zaveri, et al. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Medical Image Analysis*, 70:102032, 2021. 6, 7