Medical Image Analysis 97 (2024) 103289

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

PathoDuet: Foundation models for pathological slide analysis of H&E and IHC stains

Shengyi Hua [a], Fang Yan [b], Tianle Shen [a], Lei Ma [c], Xiaofan Zhang [a,b,*]

[a] Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai 200240, China
[b] Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China
[c] National Biomedical Imaging Center, College of Future Technology, Peking University, Beijing 100871, China

ARTICLE INFO

Keywords:
Foundation model
Pathological image
H&E and IHC

ABSTRACT

Large amounts of digitized histopathological data display a promising future for developing pathological foundation models via self-supervised learning methods. Foundation models pretrained with these methods serve as a good basis for downstream tasks. However, the gap between natural and histopathological images hinders the direct application of existing methods. In this work, we present PathoDuet, a series of pretrained models on histopathological images, and a new self-supervised learning framework in histopathology. The framework is featured by a newly-introduced pretext token and later task raisers to explicitly utilize certain relations between images, like multiple magnifications and multiple stains. Based on this, two pretext tasks, cross-scale positioning and cross-stain transferring, are designed to pretrain the model on Hematoxylin and Eosin (H&E) images and transfer the model to immunohistochemistry (IHC) images, respectively. To validate the efficacy of our models, we evaluate the performance over a wide variety of downstream tasks, including patch-level colorectal cancer subtyping and whole slide image (WSI)-level classification in H&E field, together with expression level prediction of IHC marker, tumor identification and slide-level qualitative analysis in IHC field. The experimental results show the superiority of our models over most tasks and the efficacy of proposed pretext tasks. The codes and models are available at https://github.com/openmedlab/PathoDuet.

1. Introduction

The histologic assessment stands as the gold standard for diagnosing specific cancers, predominantly depending on the expertise of pathologists. The assessment is mainly based on the analysis of Hematoxylin and Eosin (H&E) stained slides, offering fundamental structural information. Pathologists may further augment their conclusions by utilizing functional stains such as Immunohistochemistry (IHC) to provide additional diagnostic insights. As technology continues to advance, digital scanners with high throughput have revolutionized the acquisition of pathological data. Despite large amounts of data, the integration of deep learning techniques into diagnostic processes has been progressing at a relatively measured pace. This can be attributed, in part, to the limited amount of labeled data for certain tasks. Unlike the annotation of natural images, the annotation process for pathological images demands expertise, rendering it resource-intensive and time-consuming. To address this challenge, foundation models emerge as a prospective solution. These models typically exploit the potential of unlabeled data, facilitating efficient transferring to downstream tasks with reduced dependency on labeled data.

Existing foundation models mainly rely on self-supervised learning (SSL) methodologies. The essence of SSL involves the generation of supervised signals directly from the data itself. This process is often called the pretext task (Jing and Tian, 2020). As a dominant branch of SSL methods, contrastive learning (CL) has attracted significant attention (He et al., 2020; Chen et al., 2020; Oquab et al., 2023). In general, CL focuses on exploiting image similarity as a means to discern and categorize images concerning others. Another branch, featured by masked autoencoders (He et al., 2022), utilizes image generation to boost models' understanding. Compared with generative SSL methods, CL has better performance when transferred to discriminative tasks (Shekhar et al., 2023). As a result, CL is preferred in this work given the fact that quite a few pathological tasks are highly related to identification. However, the direct application of CL methods designed for natural images to histopathological images requires careful consideration. CL posits that the majority of images should possess semantic uniqueness. A routine is to contract different views of the same image and to separate different images in semantic space. Pathological whole slide images (WSI) are yet cropped into smaller patches to fit in the input

* Corresponding author at: Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai 200240, China.
E-mail address: xiaofan.zhang@sjtu.edu.cn (X. Zhang).

https://doi.org/10.1016/j.media.2024.103289
Received 13 December 2023; Received in revised form 19 July 2024; Accepted 24 July 2024
Available online 31 July 2024

size requirements of most models, restricting cropped patches to exhibit semantic distinctiveness from neighboring patches. Tough isolation of these patches may cause over-fragmentation of semantic space, thus affecting the performance of models. The conflict, consequently, requires a special design of contrastive pretext tasks concerning the essence of histopathological images. In seeking insights for the strategy of tailoring, inspiration can be drawn from the analogous process of pathological evaluation.

A typical characteristic of pathologists' working methodology is their habitual practice of zooming in and out during the examination process. Initially, they employ a low magnification level to screen overall structures and tissues, identifying regions of interest that require closer inspection. Subsequently, at a higher magnification level, pathologists analyze individual cells or clusters of cells, refining their understanding and classification of the identified regions. To simulate the zooming in and out operations performed by pathologists, we define one of our pretext tasks as a "cross-scale positioning" task, leveraging the large-scale public H&E datasets to develop a foundation model for H&E stain. In which, besides the commonly used two branches of different augmented image views in CL, we add a branch that learns the representation of a patch from its neighboring regions. This manipulation enables patches to be understood from a broader perspective, thereby alleviating the conflict between CL's requirement on semantic division and pathological patches' concentration.

Additionally, pathologists often utilize additional functional slides for a more comprehensive diagnosis. Notably, IHC markers are frequently employed, offering valuable insights into subtyping cancers. However, the effective interpretation of IHC slides cannot be derived without H&E stained slides. The H&E slides serve as a fundamental reference, providing essential contextual information and structural details to complement the specific molecular information gleaned from IHC slides. Therefore, an ideal foundation model for IHC stain should be able to assess IHC images according to markers' expression levels and align with H&E models in the semantic space in the aspect of the tissue structure. With the limited publicly available IHC data, we exploit the trained H&E foundation model and introduce the "cross-stain transferring" pretext task, to deepen comprehension of pathological images stained in a different way. Specifically, we align the IHC representation with the IHC-style transferred H&E representation via a transferer drawing on the adaptive instance normalization, AdaIN (Huang and Belongie, 2017). This alignment injects structural information that is readily accessible in H&E images, as well as preserving diagnostic information rooted in IHC images.

A pretext token mechanism is introduced to unify the two proposed tasks. Both tasks require an auxiliary input in a different form, i.e., a much smaller patch or a staining hint. Contrary to designing a separate network to handle the additional input, an extra token with the auxiliary information is fed into the Vision Transformer (ViT) model (Dosovitskiy et al., 2020) throughout the network training, and thereby combining the cross-scale or cross-stain information with original representation. The relation is subsequently exploited by a delicately designed module, termed a task raiser, to explicitly associate the two forms. This mechanism enriches the model's capacity to discover and leverage intrinsic correlation between tasks and staining modalities in a lightweight way.

The whole framework and ensuing models are collectively denoted as PathoDuet as shown in Fig. 1. This reflects the dual functionality of the framework, i.e., it offers two distinct strategies for developing foundation models: 1) by exploiting a shifted view of scales to discover broader semantic space, e.g., comprehending patches further from surrounding regions, 2) by progressing on the basis of other closely related and already exploited modalities, e.g., learning representation of IHC images from H&E model. The efficacy of our proposed methods is validated across a wide range of downstream tasks, covering both H&E and IHC images. We use a spider chart in Fig. 2 to visualize the overall performance across different tasks compared with powerful pathological models. These tasks range from classifying cancer tissue types at the patch level, categorizing WSIs, identifying cancer cells within IHC images, assessing the expression levels of IHC markers, to WSI-level IHC qualitative analysis. Our contributions are summarized as:

- We introduce an auxiliary token into the plain ViT backbone, accompanied by task raisers for the designed pretext tasks, to build up a unified SSL framework. Within this framework, we propose PathoDuet, which is the first to provide both H&E and IHC foundation models, to the best of our knowledge.
- We design the cross-scale positioning task to obtain a pretrained H&E foundation model with a broader understanding, and the cross-stain transferring task to obtain an IHC interpreter from the existing H&E foundation model with limited IHC data.
- We validate the efficacy of our methodologies on several downstream tasks, consistently demonstrating superiority across the majority of these tasks. The codes and models are open-source to facilitate future use and reproductive experiments.

2. Related works

This section reviews the literature about SSL in computer vision and histopathology, respectively.

2.1. Self-supervised learning

Self-supervised learning can be seen as another learning strategy besides supervised and unsupervised learning. The difference from the supervised one is SSL requires no labeled data, so to this point, SSL is a special form of unsupervised learning. The difference from the purely unsupervised one is SSL requires supervision generated from the data, and the generating method is called pretext tasks.

Pretext tasks can be various. Some tasks aim at predicting properties of images, like rotation angles (Gidaris et al., 2018), flipping (Srinidhi et al., 2022), etc. Some design small problems that help to learn features of images, like jigsaw puzzle solving (Noroozi and Favaro, 2016) and relative position prediction of sub-regions (Doersch et al., 2015). Some exploit the process of image generation, including image colorization (Zhang et al., 2016), super-resolution (Ledig et al., 2017), inpainting (Pathak et al., 2016), and exploiting generative adversarial networks (GANs) (Zhu et al., 2017). Besides, He et al. propose masked autoencoder (MAE) (He et al., 2022) that leads recent research into masked image modeling.

Among all these pretext tasks, similarity-based CL task has demonstrated their superiority, because they focus on invariant features instead of covariant ones. CL-based methods pull together the latent representations of similar images while keeping those of dissimilar images away from each other. These methods often regard different augmented views from the same image as a similar (positive) pair, and those from different images as dissimilar (negative) pairs. Therefore, CL-based methods typically contain two network branches. MoCo v1 (He et al., 2020) is a basic one with a symmetric structure, a momentum-update mechanism, a memory bank, and an InfoNCE loss that considers positive samples and negative samples together. SimCLR (Chen et al., 2020) adds a simple multi-layer perceptron (MLP) called projector after each backbone encoder and updates two branches concurrently without a memory bank. BYOL (Grill et al., 2020) further adds another MLP, predictor, after one of the branches (the online branch), while updating the other target branch with momentum, and changing the loss to cosine similarity loss which does not involve negative samples. SimSiam (Chen and He, 2021) proposes a similar structure of BYOL but in a symmetric way with its gradient stop technique to avoid collapse. In MoCo v3 (Chen et al., 2021), it adopts the BYOL's structure but continues using the InfoNCE loss. DINO (Caron et al., 2021) and its later version DINO v2 (Oquab et al., 2023) integrate previous works and achieve another state-of-the-art performance, thereby attracting applications to other fields.

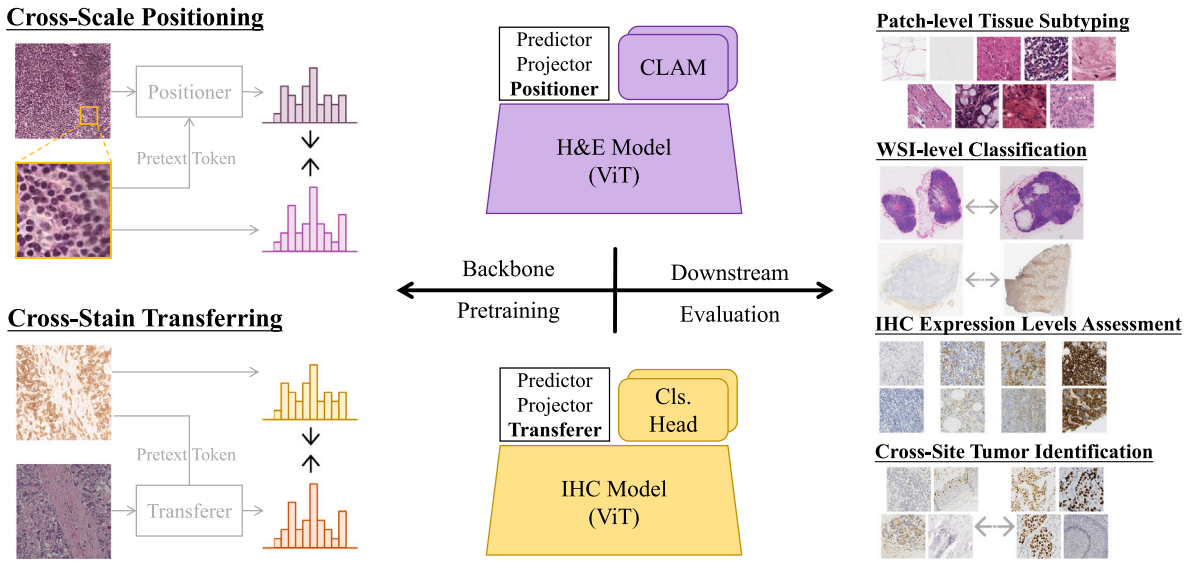


Fig. 1. An overview of PathoDuet. Left: two pretext tasks, cross-scale positioning and cross-stain transferring, are designed to develop H&E and IHC models. Right: a series of downstream tasks, covering both H&E and IHC ones, are used to evaluate models' performance in application.

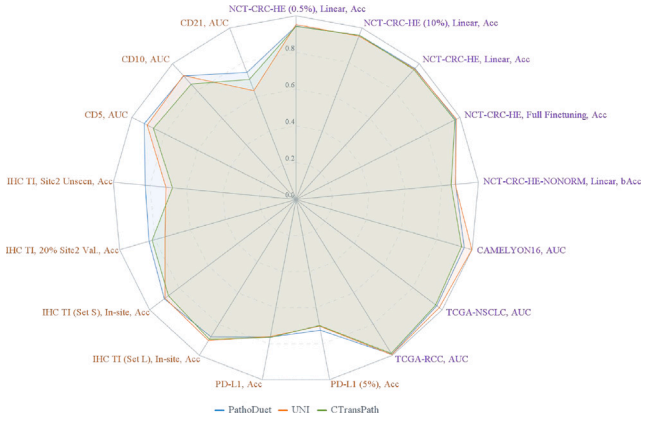


Fig. 2. An overall performance visualization. Each task is named as *training dataset, (special settings,) evaluating metric*. H&E tasks are colored purple, and IHC ones are yellow.

2.2. Self-supervised learning in digital pathology

With the development of SSL frameworks in computer vision, the concept and some existing methods have been migrated to histopathology. Besides traditional pretext tasks, some works design specific pretext tasks in digital pathology. Magnification prediction (Sahasrabudhe et al., 2020; Koohbanani et al., 2021) and stain prediction (Koohbanani et al., 2021) are two simple tasks that exploit the characteristics of pathological images. Resolution sequence prediction (Srinidhi et al., 2022) is a dedicated task with easy supervision and remarkable performance. Others leverage the stain. RestainNet (Zhao et al., 2022), as its name suggests, simulates the process of de-staining and restaining via separated Hematoxylin and Eosin channels. Ling et al. (2023) use self-supervised methods to realize arbitrary stain transfer of pathological images from different domains. These methods though take into consideration the features of histopathological images, their performance depends on the association between the pretext tasks and the downstream tasks.

Some works are based on CL frameworks. Huang et al. (2021), Li et al. (2021), Ciga et al. (2022), Kawai et al. (2023) directly migrate traditional CL methods to medical data, while some studies take into account the uniqueness of histopathological images and propose a modified SSL framework. CTransPath (Wang et al., 2022) keeps the body of MoCo v3 but adds a pseudo positive selection mechanism to avoid false penalty of semantic similar patches and uses a hybrid CNN-Transformer as its backbone encoder. Meanwhile, Wang et al. also propose a clustering-guided contrastive learning method and the produced model RetCCL (Wang et al., 2023b). CS-CO (Yang et al., 2022) separates histopathological images into H-channel and E-channel and uses a cross-stain prediction task in the first phase and a CL-based method in the second phase. Abbet et al. (2022) take advantage of domain information. Meanwhile, a branch of works has focused on the gigapixel WSIs (Vu et al., 2023; Wang et al., 2023a; Lazard et al., 2023; Aryal and Yahyasoltani, 2023; Schirris et al., 2022). Recently, pathological models pretrained with ultra-large amounts of data have shown their superiority. Chen et al. (2023) collect 100,000 slides to pretrain UNI with DINO v2 framework, and Virchow (Vorontsov et al., 2023) with 1.5 million slides. Besides, some works look into multi-modal methods like visual-language learning and pretrain models on large-scale image-text datasets (Huang et al., 2023; Pisula and Bozek, 2022; Lu et al., 2023; Zhang et al., 2023), but the focus is more on leveraging strong language models.

3. Methods

In this section, we first describe the introduction of a pretext token and subsequent task raiser module to unify the proposed two pretext tasks. The details of the tasks are discussed in the following subsections, including the real-world inspiration and the imitation with the contrastive learning framework.

3.1. Pretext token empowered SSL framework

The basic framework is based on a typical contrastive learning method, i.e., MoCo v3 (Chen et al., 2021). The difference starts from the input. To mimic pathologists, we need an extra input to bring the information from a different scale or stain. A typical practice to handle two input types is employing two networks respectively with a following interaction module. However, we aim at learning a single powerful encoder for the downstream tasks. Therefore, we propose the pretext token mechanism to uncover the relation within the same encoder, thus deepening understanding of pathological images like pathologists.

The mechanism starts from adding a *pretext token*, which is concatenated with the partitioned mini-patches' embeddings before being fed into the ViT. The token interacts with these embeddings and becomes an abstract association between the input pair in the attention blocks. After that, some simple and lightweight modules, termed as *task raisers*, utilize the association to perform pretext tasks. As illustrated in Fig. 3, the task raiser is instantiated as a *positioner* to position a local patch from its surrounding region in the cross-scale positioning task, or a *transferer* to generate IHC-style features from H&E images in the cross-stain transferring task.

In the following two subsections, we introduce in detail two pretext tasks we propose within this framework, and how pretext token and instantiated task raisers take effect in extracting information from two inputs. If not specified, we use the patch for the cropped image from a WSI and mini-patch for the output of patch embedding in ViT. Meanwhile, if the network takes only one kind of input, e.g., patches in patch network and IHC images in IHC network, the final input is formulated as $[\mathbf{x};\epsilon]$, where $\mathbf{x} = [x_1;\dots;x_L] \in \mathbb{R}^{L\times C}$ represents the original input after patch embedding module in ViT. $\epsilon \in \mathbb{R}^C$ is a set of learnable parameters to hold the place for the pretext token so that the architecture of networks can be consistent and the usage of our models can be similar to that of a normal ViT. L is the embedding length, and C is the number of channels. When the network takes two inputs, the input changes to $[\mathbf{x}^o;x^p]$, where $\mathbf{x}^o = [x_1^o;\dots;x_L^o] \in \mathbb{R}^{L\times C}$ stands for the embeddings of the original input, e.g., a region or an H&E image, and $x^p \in \mathbb{R}^C$ for those of the pretext token input, e.g., a patch or an IHC crop.

3.2. Cross-scale positioning

If one observes how pathologists view slides, the most frequent action of them is likely to be zooming in and out. This phenomenon originates from the diagnostic process wherein pathologists first detect suspicious regions from a global view, then zoom in and define the region with its local surroundings, and subsequently zoom out to inspect the next region of interest. The underlying concept is that they obtain a global and coarse understanding of a WSI under low magnification, while a local but fine understanding of critical regions under high magnification. Inspired by this, we abstract the concept as a larger region containing a smaller patch, and design the cross-scale positioning task. Due to the abundance of H&E data, this task can be applied as a normal SSL method to pretrain a model, resulting in a three-branch architecture shown in Fig. 3(a). Two of the branches exactly form the original CL framework to provide basic understanding, while the rest one performs the cross-scale positioning task to enhance understanding of different scales.

In this task, the goal is to bridge the representations of a patch from a local view and a global one. The local understanding of a patch is directly obtained from the output features of patch networks. To extract a global understanding, we jointly input patches (as pretext token) and their surrounding regions (as main input) into the region network. We instantiate the task raiser as an explicit positioner module to get positioning weights over mini-patches of regions, and use the weights to extract the desired global features.

As shown in Fig. 3(a), the whole framework has three branches, a region network, an online patch network, and a target patch network.

Patch in a local view. The two patch networks constitute a traditional CL framework, i.e., MoCo v3, to provide a basic understanding of patches. The input is a differently augmented view of the patch for each branch. The backbone encoder is denoted as $F(\cdot)$. The outputs of the patch network's backbone are believed to represent the patch in a local view, as $[\mathbf{y};y^p] = F([\mathbf{x};\epsilon])$. Then, a typical projector $G(\cdot)$ and predictor $H(\cdot)$ are used to generate the key $\mathbf{k} = G(\mathbf{y})$ and query $\mathbf{q} = H(\mathbf{k})$ as defined in MoCo v3. The contrastive loss is regarded as the loss in a local view $\mathcal{L}_{\text{local}}$, and is defined as,

$$\mathcal{L}_{\text{local}} = -\log \frac{\exp(\mathbf{q}\cdot\mathbf{k}_+/\tau)}{\sum_{i=1}^{B}\exp(\mathbf{q}\cdot\mathbf{k}_i/\tau)}, \tag{1}$$

where $\mathbf{k}_+$ is the key of a positive sample (that from a different view of the same image), B the batch size, and τ the temperature hyper-parameter. The parameters are θ for the online branch and ξ for the target. θ is updated with gradient, and a momentum update mechanism (He et al., 2020) is applied as $\xi \to m\xi + (1-m)\theta$, where m is a momentum factor.

Patch in a global view. The region network and the online patch network perform the cross-scale positioning task. In the patch networks, the pretext token is just a learnable parameter to occupy the space as mentioned before, while in the region network, the token is fed with referenced patches, thereby defining the input as $[\mathbf{x}^o;x^p]$ with $\mathbf{x}^o$ from regions and x^p from patches. When the token is passed through the encoder, it interacts with mini-patches of the region, and the output $[\dot{\mathbf{y}};\dot{y}^p] = F([\mathbf{x}^o;x^p])$ is then regarded as the association between the patch and the region. A positioner $P(\cdot)$ of simple structure like MLP, serving as the task raiser, learns to generate positioning weights over mini-patches given the association as,

$$\mathbf{p} = [p_1,\dots,p_L] = \text{Softmax}(P(\dot{y}^p)), \tag{2}$$

where $p_i \in \mathbb{R}$.

The weighted feature $\hat{\mathbf{y}} = \sum_{j=1}^{L} p_j\dot{y}_j$ combines the inherent feature of a patch and correlating features from the corresponding region, thereby considered to be the patch representation in a global view. The key $\hat{\mathbf{k}} = G(\hat{\mathbf{y}})$ and query $\hat{\mathbf{q}} = H\left(\hat{\mathbf{k}}\right)$ are simply acquired as before.

Bridge local and global views. The global feature and the local feature are then pulled together. Considering both the positioner and the patch network should be updated, we adopt SimSiam's two-way loss to propagate gradient in two networks, i.e.,

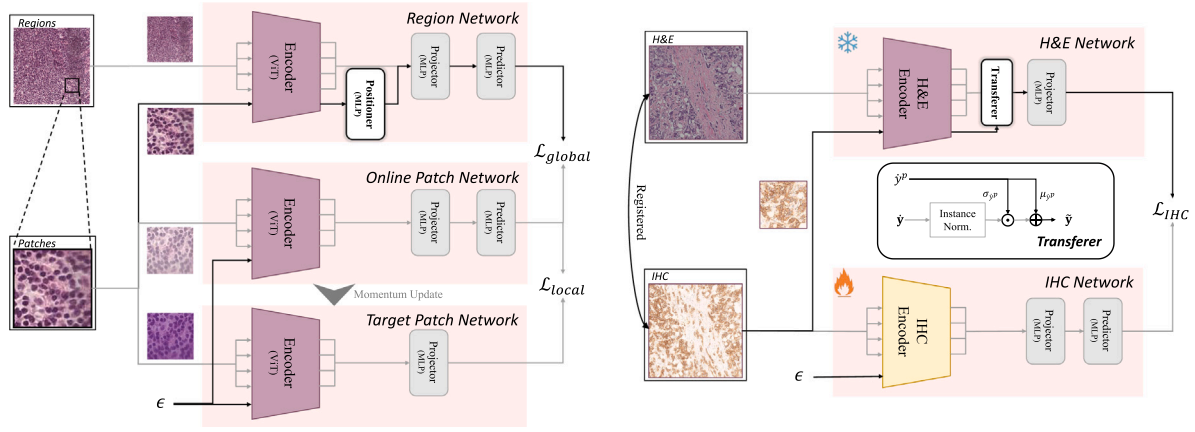
$$\mathcal{L}_{\text{global}} = D(\hat{\mathbf{q}},\text{stopgrad}(\mathbf{k})) + D\left(\mathbf{q},\text{stopgrad}\left(\hat{\mathbf{k}}\right)\right), \tag{3}$$

where $D(\mathbf{q},\mathbf{k}) = -\frac{\mathbf{q}}{\|\mathbf{q}\|_2}\cdot\frac{\mathbf{k}}{\|\mathbf{k}\|_2}$ is cosine similarity, and stopgrad$(\cdot)$ means a gradient stop manipulation (Chen and He, 2021). The local and global losses are then optimized equally.

3.3. Cross-stain transferring

As mentioned earlier, pathologists cannot always give a detailed diagnosis if only H&E slides are available. They sometimes require more evidence like IHC slides to define the disease. However, limited by the inadequacy of IHC slides, they should detect abnormal regions in H&E slides, find the corresponding regions in IHC slides, and finally derive the expected information. The underlying concept is they obtain structural information from H&E images and further diagnostic information from corresponding IHC regions. Inspired by this, we formulate the relation as paired H&E and IHC images and design the cross-stain transferring task. Considering the scarcity of desired datasets, and a priori structural knowledge in the H&E pretrained model, this task is implemented in a two-branch way without a traditional CL branch and initiated from existing H&E models. The whole process can be seen as we first acquire basic knowledge about pathological images using traditional CL method, then we enhance the global understanding using cross-scale positioning task, and now we use another cross-stain transferring task to explicitly translate H&E understanding into IHC.

In this task, the goal is to bridge the representations of a real IHC patch and a pseudo one transferred from the corresponding H&E patch. The real representations are drawn directly from the IHC network. As for the pseudo ones, we borrow the concept of adaptive instance normalization (AdaIN) (Huang and Belongie, 2017) that the style of images is rooted in some feature statistics like channel mean and variance, and utilize it to design a transferer as task raiser in the H&E network. After we jointly input IHC references (as pretext token) and their paired H&E patches (as main input) into the H&E network, the transferer replaces the mean and standard error of H&E features with those of IHC reference, thereby transferring H&E features into IHC styles. The normalized features are viewed as pseudo features of



(a) Cross-Scale Positioning **(b)** Cross-Stain Transferring

Fig. 3. Detailed networks of two pretext tasks. The flow of pretext token is represented by the black arrows, ϵ is the placeholder, and the task raisers (positioner and transferer) are presented in the white blocks. (a) Three-branch cross-scale positioning network. (b) Two-branch cross-stain transferring network and the transferer module.

the corresponding IHC patch with injected existing H&E semantics, e.g., separation of structurally normal and abnormal patches. Meanwhile, a basic CL method with negative samples ensures learning the innateness of IHC images.

As shown in Fig. 3(b), the whole framework is a typical two-branch one, including an H&E network and an IHC network.

Real IHC features. The IHC network is initiated with previous H&E models to provide a basic pathological understanding and gets updated with gradient during the transferring task. The network takes only IHC images and uses placeholders to occupy the pretext token. The outputs $[\mathbf{y}; y^p] = F([\mathbf{x}; \epsilon])$ are believed to represent the basic understanding of IHC patches and used to compute the key $\mathbf{k} = G(\mathbf{y})$ and query $\mathbf{q} = H(\mathbf{k})$.

Pseudo IHC features. The H&E network is also initiated with the previous H&E model but froze during this task to provide a stable understanding of H&E images. The network takes both H&E images as the main input and a cropped IHC reference as the pretext token. The crop is of moderate size of the original IHC image to provide adequate stain information while avoiding information leaks. The outputs of the network $[\dot{\mathbf{y}}; \dot{y}^p] = F([\mathbf{x}^o; x^p])$ are instance normalized with the crop's feature via the transferer to provide a pseudo target for IHC features.

$$\begin{aligned}\tilde{\mathbf{y}} &= \mathrm{AdaIN}(\dot{\mathbf{y}}, \dot{y}^p) \\ &= \frac{\dot{\mathbf{y}} - \mu_{\dot{\mathbf{y}}}}{\sigma_{\dot{\mathbf{y}}}} \cdot \sigma_{\dot{y}^p} + \mu_{\dot{y}^p},\end{aligned} \tag{4}$$

where $\mu_{\dot{\mathbf{y}}}$ and $\sigma_{\dot{\mathbf{y}}}$ represent the channel mean and standard error of $\dot{\mathbf{y}}$, and $\mu_{\dot{y}^p}$ and $\sigma_{\dot{y}^p}$ likewise. The key $\tilde{\mathbf{k}} = G(\tilde{\mathbf{y}})$ is computed as before.

Bridge features. The real IHC feature and the pseudo feature are then pulled together with typical InfoNCE loss $\mathcal{L}_{\mathrm{IHC}}$, which is aware of negative samples.

$$\mathcal{L}_{\mathrm{IHC}} = -\log \frac{\exp(\mathbf{q} \cdot \tilde{\mathbf{k}}_+ / \tau)}{\sum_{i=1}^{B} \exp(\mathbf{q} \cdot \tilde{\mathbf{k}}_i / \tau)}, \tag{5}$$

where $\tilde{\mathbf{k}}_+$ is the pseudo key of a positive sample, B and τ the same as mentioned before.

4. Experiments

In this section, we first introduce the datasets and experimental settings to obtain the H&E and IHC models. Next, we describe in detail the downstream tasks of H&E and IHC images sequentially. The H&E tasks include patch-level colorectal cancer (CRC) tissue typing and WSI-level classification to evaluate both basic discriminating capability and global understanding of H&E models. The IHC tasks include a typical assessment of the IHC marker's expression level, and a cross-site tumor identification to demonstrate basic IHC understanding, as well as generalization competence of models.

These results are based on quantitative experiments with other models. The models first include models using ImageNet (Deng et al., 2009) as baselines, i.e., fully-supervised ViT (Dosovitskiy et al., 2020) denoted as *ImageSup* and self-supervised ViT via MoCo v3 (Chen et al., 2021) as *ImageSSL*. Besides, some pathological models are also taken into consideration. Ciga et al. (2022) utilize SimCLR strategy to pretrain a ResNet-18 with large amounts of public pathological data (*SimCLR-ciga*). Wang et al. provide both a pretrained ResNet-50, *RetCCL* (Wang et al., 2023b), and a pretrained hybrid architecture of CNN and Swin Transformer, *CTransPath* (Wang et al., 2022), via customized methods and pathological datasets. The last is *UNI* (Chen et al., 2023), a ViT-Large pretrained with over 100 thousand slides using DINO v2 framework. These open-source models are compared throughout the experiments in this section. Pretrained model with ultra-large scale amounts of pathological data, like *Virchow* (Vorontsov et al., 2023), is postponed to Section 5.3 for a simple discussion since it is not publicly available at this time.

4.1. Pretraining stage

The pretraining stage consists of two stages, pretraining an H&E model and transferring to an IHC model. In the first stage, we perform the cross-scale positioning task under the MoCo v3 framework with our H&E dataset. In the next stage, we perform the cross-stain transferring task to the H&E model with the cross-stain dataset.

Datasets. The *H&E dataset* originates from TCGA,[1] a large-scale public dataset containing genome, epigenome, transcriptome, and image data. In this work, we collect about 30 thousand WSIs from it and select about 11 thousand diagnostic formalin-fixed paraffin-embedded (FFPE) WSIs for training. The patches are cropped under the highest magnification level with a size of 256 × 256 pixels, and the regions under the second highest level with a size of 1024 × 1024 pixels. Hence, the physical ratio of size between regions and patches is typically around 8. Finally, 1,623,258 regions and 13,166,437 patches are acquired. The *cross-stain dataset* originates from HyReCo (Lotz et al., 2022; van der Laak et al., 2021) and BCI dataset (Liu et al., 2022). HyReCo dataset consists of nine datasets of consecutive sections, each containing four slides stained with H&E, CD8, CD45, and Ki67, respectively. Additional PHH3-stained slides are re-stained from the bleached H&E slides. In total, 2,771 pairs of H&E and one stain of

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

Table 1
Linear evaluation results on NCT-CRC-HE dataset with different amounts of training data. The best performance in each column is bold, and the second best is underlined.

Methods	Percentage of training data									
	0.5%		1%		10%		50%		100%	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
ImageSSL	0.882	0.844	0.900	0.866	0.931	0.900	0.934	0.906	0.935	0.908
ImageSup	0.922	0.895	0.934	0.914	0.942	0.931	0.947	0.932	0.946	0.933
SimCLR-ciga	0.928	0.900	0.932	0.906	0.935	0.910	0.936	0.908	0.938	0.910
RetCCL	0.943	0.914	0.943	0.916	0.945	0.920	0.944	0.924	0.945	0.924
CTransPath	0.942	0.907	0.952	0.923	0.958	0.935	0.956	0.932	0.956	0.932
UNI	**0.952**	**0.934**	**0.953**	**0.935**	0.953	0.937	0.958	0.938	0.961	0.945
Ours (H&E)	0.943	0.919	0.950	0.929	**0.959**	**0.942**	**0.964**	**0.949**	**0.964**	**0.950**

IHC are acquired. BCI dataset contains 4,873 pairs of H&E and HER2 images, 3,896 pairs for train, and 977 for test. We only use the 3,896 training pairs. To obtain more training data, we crop these images under another resolution, and finally, 21,126 pairs are used in the second task.

Settings. In the training stage, we use a typical ViT-B/16 backbone and activate automatic mixed precision by PyTorch. For later consistency with the cross-scale positioning task, we use the "avg" mode instead of "token". Also, considering the gap between ImageNet and our histopathological dataset, we ignore the normalization with ImageNet's mean and variance. Following MoCo v3, the projector is a three-layer perceptron, and the predictor a two-layer perceptron with a hidden dimension of 4096. The data augmentation strategies include random cropping and scaling, Gaussian blur, color distortion, random flipping, following BYOL. The batch size is 2048, τ is set as 0.2, and the learning rate is initially 0.00015 with a scaler of BatchSize/256 and updated using a cosine decay schedule with a long warm-up of 40 epochs. AdamW is used as the optimizer with 0.1 weight decay.

We first perform the vanilla contrastive learning method with 100 epochs. It takes about 300 hours on 8 A100 GPUs. Then, we activate the positioning task. The positioner is also a two-layer perceptron with the same hidden dimension. We train an extra 20 epochs with 10 warm-up epochs, which takes another 100 hours. Then, the pretrained H&E backbone can be transferred or used directly for later downstream tasks with H&E-stained images.

After that, we perform the cross-stain transferring task based on the previous model. The batch size is 512, the learning rate is 0.0002, and the total number of epochs is 100 with 20 warm-up epochs. It takes about 6 hours on 4 GeForce RTX 3090 GPUs. After this stage, the model can be applied to downstream tasks with certain IHC images.

4.2. Downstream tasks with H&E images

To better evaluate the performance and generalizing ability of PathoDuet, we conduct a series of downstream experiments. The experiments start from H&E-related ones.

Cancer tissue identification is one of the most important tasks in computational pathology, especially with H&E images. One formulation of this task is patch-level classification, or supervised image classification, which presents a relatively simple way to test models' basic understanding of H&E patches. To both study the native ability of models and simulate the real scenarios, a linear probing strategy (Chen et al., 2020) and a full finetuning one are implemented to compare models' performance.

Another one is WSI-level classification, or weakly-supervised classification, which closely resembles real-world scenarios and is more persuasive in demonstrating models' global understanding of pathological slides. This task is usually implemented in a multiple instance learning way, and we utilize the attention-based CLAM (Lu et al., 2021) as the framework to perform the MIL process.

We use these two tasks as an evaluation of our H&E model. A more concrete description of each task is arranged in the following two subsections, while the detailed parameter settings are in the supplementary materials.

Table 2
Results on NCT-CRC-HE dataset for 2 different strategies: linear evaluation, full fine-tuning. The best performance in each column is bold, and the second best is underlined.

Methods	Linear evaluation		Full fine-tuning	
	Acc	F1	Acc	F1
ImageSSL	0.935	0.908	0.958	0.945
ImageSup	0.946	0.933	0.960	0.947
SimCLR-ciga	0.938	0.910	0.937	0.924
RetCCL	0.945	0.924	0.950	0.936
CTransPath	0.956	0.932	0.969	0.960
UNI	0.961	0.945	**0.979**	**0.967**
Ours (H&E)	**0.964**	**0.950**	0.973	0.964

4.2.1. Patch-level tissue subtyping

Identifying cancerous tissue is the main work of a pathologist, thereby playing a dominant role in computational pathology. The nature of this behavior is to recognize tissues and distinguish abnormal images. Hence, a feasible formulation is to crop the WSIs into patches and then classify them. On the other hand, patch-level classification or fully-supervised classification is one of the most basic evaluations of models. To this, this task is preferred as a good prologue of the comprehensive assessment.

In this task, the pretrained model is used as a feature extractor of cropped patches, followed by a classifying layer. The experiments are conducted under two different strategies. First, the typical *linear evaluation* strategy, i.e., only the newly-added linear layer gets updated while the rest part is frozen. Second, the *full fine-tuning* strategy, which is more likely to be applied in practice. In this way, the pretrained model gets trained together with the linear classifier. Accuracy (Acc) and F1 score are used as the evaluating metrics. It is worth noting that to better demonstrate models' performance on transfer learning, we use different proportions of training data in linear evaluation, and thus all the results are reproduced ones considering the randomness of data splits.

Datasets. We use the *NCT-CRC-HE* dataset (Kather et al., 2018) for the patch classification task. It is a collection of histopathology images specifically focused on CRC. It consists of 9 categories of tissue types, with one category representing normal tissues (NORM) and the remaining 8 categories representing colorectal cancer tissues, including adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), cancer-associated stroma (STR) and colorectal adenocarcinoma epithelium (TUM). A total number of 100,000 image patches with a size of 224 × 224 pixels at 0.5 μm per pixel (mpp) are used as the training set, while 7,180 image patches are used as the testing set. All training patches are cropped from 86 H&E WSIs and testing patches from 50 H&E WSIs.

Settings. For linear evaluation on the NCT-CRC-HE dataset, following CTransPath, we use Adam as the optimizer with a batch size of 96. The initial learning rate is set to 0.0003. Data augmentations of random horizontal, vertical, and 90-degree flips and random cropping are used.

Table 3
Results of weakly-supervised WSI classification on three public datasets. The best performance in each column is bold, and the second best is underlined.

Methods	CAM16		NSCLC		RCC	
	Acc	AUC	Acc	AUC	Acc	AUC
TransMIL	0.884	0.931	–	–	–	–
CLAM-SB	0.884	0.940	0.894	0.951	0.929	0.986
CLAM-SB+ImageSSL	0.861	0.915	0.883	0.949	0.924	0.989
CLAM-SB+ImageSup	0.853	0.890	0.870	0.940	0.943	0.990
CLAM-SB+SimCLR-ciga	0.899	0.953	0.900	0.949	0.931	0.987
CLAM-SB+RetCCL	0.868	0.919	0.851	0.927	0.932	0.987
CLAM-SB+CTransPath	0.868	0.940	0.904	0.956	0.928	0.988
CLAM-SB+UNI	**0.984**	**0.999**	**0.934**	**0.980**	**0.959**	**0.995**
CLAM-SB+Ours (H&E)	0.930	0.956	0.908	0.963	0.954	0.993

For the full fine-tuning, the learning rate for the linear classifier is set to 0.0003, while the learning rate for the rest part of the pretrained model is set to 0.00003. The maximum number of epochs is set to 50 for all models to converge.

Results. In Table 1, we evaluate our H&E model using the linear probing method under different amounts of data. From the result, we can see that our model performs well across various amounts of training data over other pretrained models. Meanwhile, it can be observed that a generally consistent increasing trend exists with the growth of amounts of training data, but the difference is relatively small for most models. A further study is conducted in Section 5.2 on the training data requirements of foundation models. Notably, the giant UNI shows a dominant performance when the training data is extremely limited, which demonstrates its general interpretability of pathological images. In Table 2, we present the evaluation of models' performance under different training strategies using the whole NCT-CRC-HE dataset. The results demonstrate that the proposed model is a good interpreter of H&E images under both a quick linear transferring manner and a thorough full fine-tuning protocol. The performance gain can be owed to the cross-scale positioning task, which enhances the model's understanding under a broader view. To verify the assumption, an ablating study is discussed in Section 5.1. UNI also provides decent performance, which shows its great understanding in pathology and powerful ViT-Large architecture.

4.2.2. WSI-level classification

Another formulation of cancer identification is closely related to real-world scenarios. The processing unit is the WSIs instead of small patches. In practice, the WSI classification task is typically weakly-supervised, since only global annotations are given, and the WSI-level labels may correspond to only small regions within. For this reason, this task challenges models' global understanding of pathological images.

Recent works (Campanella et al., 2019; Shao et al., 2021; Lu et al., 2021) have demonstrated the effectiveness of multiple instance learning (MIL) on weakly-supervised classification of WSIs. These MIL-based methods typically follow a two-step approach. First, WSIs are divided into smaller patches to generate patch-level features by utilizing a pretrained model. Second, to generate a slide-level prediction, patch-level features are aggregated using various feature fusion techniques, including recurrent neural network (RNN) or Transformer-based aggregators (Campanella et al., 2019; Shao et al., 2021) and attention-based pooling (Lu et al., 2021).

For pretrained models, we freeze the parameters and utilize the attention-based CLAM as the framework to perform the MIL process. We use Acc and area under the receiver operating characteristic curve (AUC) score to evaluate the WSI classification task. The AUC is calculated using the macro-averaged one when the number of classes is larger than 2. Besides the aforementioned methods, we further include the original CLAM (Lu et al., 2021) and TransMIL (Shao et al., 2021) as baselines. It is worth noting that except for TransMIL, other results are reproductive ones considering the update of datasets.

Datasets. We evaluate this task on three WSI-level datasets: CAMELYON16 (CAM16), TCGA non-small cell lung cancer (NSCLC), and TCGA renal cell carcinoma (RCC). *CAMELYON16* dataset was released as part of the Camelyon16 challenge (Bejnordi et al., 2017), which focused on two types of breast cancer classification: benign tissue and metastatic breast cancer. The dataset consists of a total of 399 Whole Slide Images, with 270 WSIs for training and 129 for testing. Although the dataset does provide detailed pixel-level annotations, in the context of weakly-supervised classification, we only utilize the global slide-level annotations, i.e., whether a WSI contains tumor areas or not. *TCGA-NSCLC* dataset is derived from the TCGA database for two class subtyping: lung squamous cell carcinoma (TCGA-LUSC) and lung adenocarcinoma (TCGA-LUAD). It contains a total of 1053 diagnostic WSIs, including 512 LUSC slides from 478 cases and 541 LUAD slides from 478 cases. *TCGA-RCC* dataset is another subset of TCGA which includes three subtypes of kidney tumor: kidney chromophobe renal cell carcinoma (TCGA-KICH), kidney renal clear cell carcinoma (TCGA-KIRC), and kidney renal papillary cell carcinoma (TCGA-KIRP). It contains a total of 940 diagnostic WSIs, including 121 KICH slides from 109 cases, 519 KIRC slides from 513 cases, and 300 KIRP slides from 276 cases.

Settings. For the weakly-supervised WSI classification task, following CLAM, we freeze our pretrained model and use the Adam as the optimizer with a batch size of 1 (WSI/bag) and a weight decay of 0.00001. The learning rate is set to 0.0002, and the epochs to 50. For the CAMELYON16 dataset, we adopt the official data split in the CAMELYON16 challenge. For TCGA-NSCLC and TCGA-RCC datasets, we use 5-fold Monte Carlo cross-validation to obtain a more stable result. Each WSI is cropped into non-overlapping small patches after removing the background area with a filter of saturation less than 15 in the CAMELYON16 dataset, and 8 in the TCGA dataset.

Results. In Table 3, various methods are compared using three different public datasets. UNI achieves all best across these three datasets, which demonstrate its power in understanding pathological slides. Excluding UNI, our model shows great performance over these three datasets, e.g., +3.1% accuracy gain in CAMELYON16, +0.3% in TCGA-NSCLC, and +1.1% in TCGA-RCC. This demonstrate that it is effective to use cross-scale positioning tasks to enhance the global understanding of pathological images. As for other models, pathological models surpass original CLAM and ImageNet models in most cases, but the superiority is not consistent. This might be credited to the strong generalizing ability and global understanding gained from ImageNet, especially when the visual encoder is frozen.

4.3. Downstream tasks with IHC images

As for the other part of PathoDuet, to evaluate the IHC model's competence, we conduct three additional experiments that are closely related to real-world diagnosis, expression level assessment of certain markers, cancer cell identification and slide-level IHC qualitative analysis.

Assessing the expression level of IHC markers is a primary work for pathologists to assess an IHC slide. We formulate this task as a classification between IHC patches of different expression levels to test models' basic capability of tackling IHC images. Cancer cell identification is also critical in the analysis of IHC images because only the expressed cancer cells are of diagnostic significance. It is yet hard for pathologists to find these cells in IHC slides alone since the structure of single cells is too vague to be recognized. Hence, this task serves as an advanced challenge to models' understanding of IHC images. Meanwhile, with data from two sites available, this task is intended as a cross-site one to better evaluate models' generalizing ability. To note, these tasks are implemented under linear protocol to focus on the inherent capabilities of models. Slide-level IHC qualitative analysis tests models' performance on WSI diagnosis. We collect three different markers, and exploit multiple instance learning method to ask models to give a positive or negative prediction of certain marker. A detailed description is also delivered in the following subsections.

Table 4
Results of PD-L1 expression level assessment. The best performance in each column is bold, and the second best is underlined.

Methods	100% Training Set			5% Training Set		
	Acc	bAcc	wF1	Acc	bAcc	wF1
ImageSSL	0.754	0.721	0.753	0.686	0.698	0.695
ImageSup	0.726	0.688	0.715	0.648	0.653	0.651
SimCLR-ciga	0.754	0.754	0.744	<u>0.705</u>	0.704	<u>0.710</u>
RetCCL	0.751	0.754	0.754	0.677	0.693	0.686
CTransPath	**0.765**	**0.762**	**0.768**	0.700	<u>0.709</u>	0.703
UNI	0.760	0.747	0.755	0.703	<u>0.709</u>	0.701
Ours (IHC)	<u>0.763</u>	<u>0.755</u>	<u>0.765</u>	**0.726**	**0.721**	**0.732**

4.3.1. IHC expression levels assessment

Besides tumor cell identification, assessment of certain marker's expression levels plays a primary role in IHC diagnosis. We formulate it as a patch-level multi-class classification task, because trivial regression may lose focus on certain scores that have diagnostic meaning. We manually selected several thresholds closely related to pathologists' examination. A linear probing method is applied with metrics including accuracy (Acc), balanced accuracy (bAcc), and weighted F1 (wF1) score because of class imbalance.

Datasets. Considering the scarcity of public datasets that meet the requirements, we use an in-house dataset. We collect two groups of IHC patches with PD-L1 marker from the same medical center. After exhaustive annotation of expression scores, we select 0.05, 0.2, and 0.5 as thresholds, thereby creating a 4-class classification task of rarely-, lightly-, moderately- and severely-expressed patches. In the first group, there are 765/1,059/645/481 patches in the order of expression level, and 693/1,138/491/478 patches in the second. We use Group 1 as the training set and Group 2 as the validation set. Besides the original setting, a more difficult setting is also implemented that only 5% of training data is available, i.e., 38/46/38/25 patches respectively.

Settings. For IHC expression levels assessment, we also keep the linear classification setting in Section 4.2.1. The batch size is 128 for full training data and 64 for 5% training data. The learning rate is 0.02.

Results. In Table 4, the performance of different models is reported in different amounts of training data. Reviewing the overall results, we can suppose that pathological foundation models of H&E images can have a certain insight into IHC images because we can see superior performance over ImageNet models. When focusing on individual models, our IHC model demonstrates attracting performance on most metrics, especially in the limited training data case. Notably, CTransPath also presents excellent performance, especially when training data is sufficient, and SimCLR-ciga provides second best performance when training data is limited. This might be credited to their use of some IHC images as pretraining data. Besides, although UNI does not see IHC images during pretraining, it still presents surprisingly powerful generalizing ability to IHC images. When the difficulty of the task increases, the advantage of explicit transferring to the IHC model is more obvious. The phenomenon states that although H&E models can provide satisfactory results with adequate training data, explicit transferring is necessary for limited annotated data.

4.3.2. Cross-site tumor identification

As mentioned earlier, identifying tumor cells in IHC images is of high significance, but difficult for pathologists without auxiliary H&E images. Therefore, this task can be a further examination of the models' abilities. Meanwhile, with the help of data from two different sites, we can further investigate models' generalizing competence under out-of-distribution settings. We formulate it as a patch-level classification task and use the linear protocol as well. The metrics are the Acc and F1 score.

Datasets. The dataset is also private, consisting of IHC images from two medical sites. The annotation is simply positive (images containing tumor cells) and negative (images without any tumor cells). The first site (*Site 1*) contains two sets of data stained with ER, KI67, and PR acquired at different times, resulting in slightly different appearances. The first set has 1,365 patches (700 positive and 665 negative), and the other has 126 patches (64 positive and 62 negative), denoted as *Set L* and *Set S* respectively. In the second site (*Site 2*), there are 3,688 patches with much more domain shift, including 2,684 positive and 1,004 negative stained with ER, KI67, and EGFR.

To better evaluate the models' performance, we design four scenarios. *Set L* → *Set S* and *Set S* → *Set L* are viewed as in-site settings. The difference is we use a larger dataset (Set L) as the training set in the former one and a smaller dataset (Set S) in the latter one. Data in Site 2 can be viewed as being sampled from a new domain other than Site 1. To this, it can be used for out-of-distribution (OOD) evaluation. We also design two OOD settings. *Site 2 Seen* keeps 20% of data in Site 2, and uses it for model selection. The model is trained with Set L in Site 1, validated with 20% data in Site 2, and tested with the rest 80% data. *Site 2 unseen* is a stricter setting, which is also common in real-world scenarios. The model is trained with Set L in site 1, validated with Set S, and tested with data from absolutely unseen Site 2.

Settings. For IHC tumor identification, a linear probing method is implemented as well. In *Set L* → *Set S*, *Site 2 Seen* and *Site 2 Unseen*, the batch size is 128 and the learning rate is 0.005. In *Set S* → *Set L*, the batch size is reduced to 64. The maximum number of epochs is also set to 50 for convergence of all models.

Results. In Table 5, the performance of different models is reported under four evaluation scenarios. When applying in-site settings, ImageNet supervised ViT, CTransPath, UNI, and our IHC model all present remarkable performance. When tested on another site, ImageNet models yet display a weakening performance. Meanwhile, our model offers the best performance in a setting of partially available data from a new domain and maintains the performance in a pure OOD setting. Other pathological models perform well with some a priori information from another domain, but may not keep it when the test set is totally unseen.

4.3.3. Qualitative analysis of IHC slides

Besides patch-level IHC tasks, diagnosis directly from IHC slides is also of great importance. Therefore, we collect some IHC slides of different markers, and invite some experts to give a positive or negative label to each slide. This task further examines models' capability of assessing IHC slides given certain marker. We still apply CLAM as the training method. The evaluating metrics include AUC, Acc, F1 score, Recall, Precision and Specificity.

Datasets. The dataset includes 3 IHC markers, CD5, CD10 and CD21. For each marker, we collect over 250 slides and annotate each slide with a positive or negative label. To be detailed, we collect 124/189 positive/negative CD5 slides, 139/111 CD10 slides, and 115/150 CD21 slides, respectively. We use each marker solely and implement 5-fold cross validation.

Settings. For slide-level prediction of IHC markers, we follow the settings used in TCGA experiments as introduced in Section 4.2.2.

Results. In Tables 6–8, we present the results of slide-level prediction of CD5, CD10 and CD21. If we take three tables as a whole, we can see that ImageSSL, SimCLR-ciga, CTransPath, UNI and our models provide relatively good and stable performance. For ImageSSL, an interesting fact is that it surpasses ImageSup in all statistics. A reasonable guess is self-supervised learning helps model generalizing to other domains, especially those rarely-seen domains. For SimCLR-ciga and CTransPath, as aforementioned, they may benefit from some IHC data used for pretraining. For UNI, it still presents great performance with its surprising generalizing ability. For our model, it provides outstanding performance, achieving at least second-best AUC across all three datasets, while no other models can achieve. This should be credited to cross-stain transferring, which exploits existing H&E model and a little amount of paired IHC and H&E data. Reviewing the overall results, we can derive similar conclusions. Foundation models of H&E

Table 5
Results of patch-level tumor identification in IHC images. The best performance in each column is bold, and the second best is underlined.

Methods	Site 1 → Site 1				Site 1→ Site 2			
	Set L → Set S		Set S → Set L		Site 2 Seen		Site 2 Unseen	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
ImageSSL	0.817	0.811	0.864	0.862	0.547	0.545	0.504	0.503
ImageSup	0.881	0.879	0.875	0.875	0.677	0.667	0.537	0.535
SimCLR-ciga	0.849	0.846	0.766	0.754	0.762	0.744	0.728	0.717
RetCCL	0.802	0.793	0.818	0.811	0.727	0.717	0.629	0.625
CTransPath	0.897	0.896	0.872	0.870	0.816	0.794	0.677	0.657
UNI	**0.905**	**0.904**	0.895	0.894	0.743	0.700	0.712	0.692
Ours (IHC)	0.881	0.881	**0.900**	**0.900**	**0.833**	**0.797**	**0.826**	**0.769**

Table 6
Results of slide-level prediction of CD5. The best performance in each column is bold, and the second best is underlined.

Methods	AUC	Acc	F1	Recall	Precision	Specificity
ImageSSL	**0.931**	**0.888**	**0.908**	**0.909**	**0.908**	**0.860**
ImageSup	0.924	0.873	0.895	0.895	0.899	0.844
SimCLR-ciga	0.862	0.798	0.837	0.850	0.826	0.722
RetCCL	0.881	0.825	0.852	0.840	0.866	0.803
CTransPath	0.868	0.814	0.844	0.834	0.856	0.784
UNI	0.907	0.863	0.887	0.888	0.887	0.828
Ours (IHC)	0.924	0.840	0.862	0.835	0.895	0.850

Table 7
Results of slide-level prediction of CD10. The best performance in each column is bold, and the second best is underlined.

Methods	AUC	Acc	F1	Recall	Precision	Specificity
ImageSSL	0.898	**0.842**	**0.813**	0.802	0.827	0.873
ImageSup	0.870	0.800	0.758	0.737	0.784	0.848
SimCLR-ciga	0.902	**0.842**	0.810	**0.814**	0.824	0.874
RetCCL	0.855	0.782	0.742	0.749	0.752	0.819
CTransPath	0.849	0.773	0.729	0.720	0.754	0.825
UNI	**0.912**	0.828	0.786	0.757	0.820	0.880
Ours (IHC)	0.909	0.835	0.795	0.766	**0.838**	**0.894**

Table 8
Results of slide-level prediction of CD21. The best performance in each column is bold, and the second best is underlined.

Methods	AUC	Acc	F1	Recall	Precision	Specificity
ImageSSL	0.677	0.614	0.667	0.690	0.662	0.535
ImageSup	0.638	0.571	0.614	0.628	0.631	0.520
SimCLR-ciga	0.659	0.582	0.638	0.660	0.627	0.492
RetCCL	0.657	0.604	0.659	0.688	0.644	0.507
CTransPath	0.701	**0.637**	0.667	0.651	**0.703**	**0.631**
UNI	0.635	0.607	0.644	0.641	0.658	0.570
Ours (IHC)	**0.740**	0.626	**0.669**	**0.695**	0.667	0.548

images can be a good choice to understand IHC images, compared with models trained with natural images. An IHC-specific model, however, can provide more insights especially when the difficulty of the task is relatively high, i.e., large domain shift and limited annotated data.

5. Discussion

In this section, we first conduct several ablation experiments to validate the efficacy of two proposed pretext tasks. Then, we study the data requirement of applying foundation models to downstream tasks, to demonstrate the necessity of developing pathological foundation models. Finally, we compare our H&E model to some giant models pretrained with ultra-large scale pathological datasets and observe that both increasing the amount of data and designing a tailored framework help to develop powerful pathological models.

Table 9
Ablation study: performance on WSI classification.

Model	CAM16		NSCLC		RCC	
	Acc	AUC	Acc	AUC	Acc	AUC
MoCo v3	0.915	0.959	0.890	0.937	0.948	0.992
+XSP	0.930	0.956	0.908	0.963	0.954	0.993

Table 10
Ablation study: performance on H&E patch classification.

Model	Linear evaluation		Full fine-tuning	
	Acc	F1	Acc	F1
MoCo v3	0.956	0.944	0.973	0.960
+XSP	0.964	0.950	0.973	0.964

Table 11
Ablation study: performance on PD-L1 expression level assessment.

	100% Training Set			5% Training Set		
	Acc	bAcc	wF1	Acc	bAcc	wF1
H&E	0.762	0.763	0.764	0.714	0.699	0.717
IHC	0.763	0.755	0.765	0.726	0.721	0.732

5.1. Efficacy of the proposed methods

In Section 3.2, we claim that the essence of the cross-scale positioning task is to bridge representations in a local view and those in a global view. To validate whether cross-scale positioning enhances models' global understanding, we choose the WSI classification task since it is more related to a global understanding of a slide. We compare the performance of a purely MoCo v3 pretrained model using our dataset and the model using cross-scale positioning, named *MoCo v3* and *+XSP*, respectively. From the results in Table 9, we can see that the latter model outperforms the purely MoCo v3 one in most metrics. The results can prove that cross-scale positioning boosts models' understanding of an image from a broader view. Besides, we also conduct ablating experiments on NCT-CRC-HE dataset, and observe a little performance gain after applying cross-scale positioning task as shown in Table 10. The margin in full finetuning setting barely exists, which is reasonable since the two models share the same model architecture. If we take the results in Tables 9 and 10 as a whole, we can derive the conclusion that cross-scale positioning helps model understand H&E images better.

In Sections 4.3.1 and 4.3.2, we prove that our IHC model can surpass other pathological models, and we append an ablation study to further prove the efficacy of cross-stain transferring. We use the PD-L1 expression level assessing task. From the results in Table 11, it is obvious that explicit transferring benefits the understanding of IHC images on the condition of a strong H&E base, especially when the training date is limited.

Table 12
Comparative study on NCT-CRC-HE and NCT-CRC-HE-NONORM dataset. SwinT* means a hybrid model of CNN and Swin Transformer. The best performance in each column is bold, and the second best is underlined.

	Architecture	#WSIs	CRC			CRC (no norm)		
			Acc	bAcc	wF1	Acc	bAcc	wF1
Ours (H&E)	ViT-Base	~11K	<u>0.964</u>	<u>0.952</u>	<u>0.964</u>	0.888	0.875	0.894
Ours (H&E)*			–	–	–	<u>0.901</u>	<u>0.888</u>	<u>0.906</u>
CTransPath	SwinT*	~32K	0.958	0.931	0.955	0.879	0.852	0.883
UNI	ViT-Large	~100K	–	–	–	–	0.874	0.875
Virchow	ViT-Huge	~1.5M	**0.968**	**0.956**	**0.968**	**0.948**	**0.938**	**0.950**

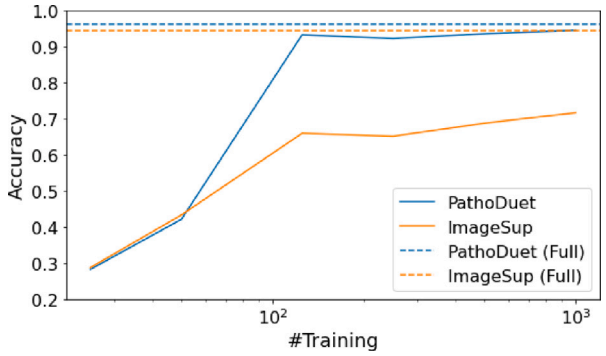


Fig. 4. Data requirement study on NCT-CRC-HE dataset. PathoDuet is compared with ImageSup and the performance with the full dataset as an upper bound is represented by the dotted line with the same color.

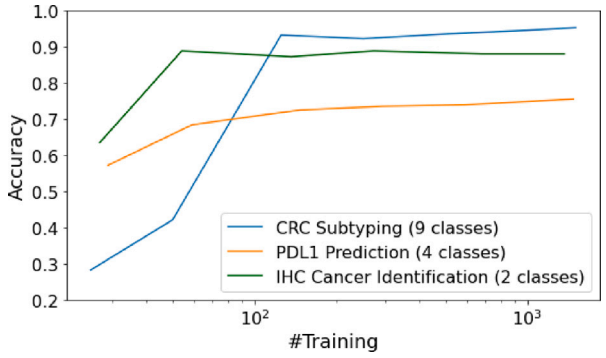


Fig. 5. Data requirement study on different datasets using PathoDuet.

5.2. Data requirements of downstream tasks

To prove the necessity of building up pathological foundation models, we further study the amounts of training data to effectively apply a foundation model to specific downstream tasks.

In Fig. 4, we compare our H&E model with ImageNet-supervised ViT using the NCT-CRC-HE dataset. We can observe that although these two models can present close performance with the full dataset (10^6 images), when the amount is limited to only 10^2, the performance gap is large. This can be used to validate that using pathological foundation models can reduce data requirement of downstream tasks in pathology.

To gain a more general insight in a direct way, we present the performance of PathoDuet over three different patch-level datasets, as introduced in Section 4.2.1, 4.3.1 and 4.3.2 respectively. From the results in Fig. 5, it is obvious that PathoDuet reduces the data requirements to about 10^2 over different datasets. However, when there are only less than 50 training images, the models fail to tackle the downstream tasks and present a performance like random guess. Through these two experiments, we look further into the value of pathological foundation models in that they highly relieve the burden of data annotation.

5.3. Comparison with giant pathological models

Recently, there has been a trend in pretraining pathological models with astronomical numbers of diagnostic slides. UNI, proposed by Chen et al. (2023), utilizes 100,000 slides to train a ViT-Large with DINO v2 framework. Virchow (Vorontsov et al., 2023) further extends this number to 1.5 million and employs a ViT-Huge model with DINO v2 as well. In this section, we compare our model to those giants and CTransPath as a baseline in the CRC patch classification task. Notably, except our models, all the figures are copied from Virchow's experiment, so the results of CTransPath are different from that in Section 4.2.1, which is as small as 0.002 in accuracy, and 0.001 in balanced accuracy and weight F1 score, and can be owed to different training strategies and randomness. Here, we provide two versions of our model, a normal ViT used throughout the aforementioned experiments, and a ViT with our positioner network to aggregate the features instead of using the vanilla average. The latter one is marked with a *.

Reviewing the results in Table 12, we can see a small gap when using the normalized version of the dataset. The value of data is much more obvious with the non-normalized dataset since Virchow shows an impressive and dominant performance. However, the normal use of our 11 K-slide-trained ViT-Base model can surpass the 100 K-slide-trained UNI, which demonstrates the power of combining field knowledge. Hence, we claim that it is also of vital importance to employ a proper training strategy carefully tailored to pathological characteristics.

6. Conclusion

We introduce PathoDuet, a series of foundation models on computational pathology, covering both H&E and IHC images, and propose a new self-supervised learning framework with two pretext tasks in pathology. The key to this framework is the introduction of a pretext token and following task raisers. It consists of both a model pretraining task, cross-scale positioning, and a model adaptation task, cross-stain transferring. In cross-scale positioning, we bridge the local and global representations of H&E patches to enhance pathological image understanding in various magnifications. In cross-stain transferring, we utilize adaptive instance normalized H&E features to provide pseudo-IHC features injected with structural information. The original H&E model is therefore transferred to an interpreter of IHC images. We evaluate the performance of our models over a wide variety of downstream tasks, and the experimental results show the efficacy of our models on most tasks. Besides, we also investigate the downstream data requirements and comparison with giant pathological models, to discover the power of data and delicately designed SSL methods tailored to pathological images. PathoDuet highlights the importance of training strategy, while the giants, UNI and Virchow, point out the advantage of preparing sufficient training data. Hence, we will take all efforts to collect more data to iterate and upgrade our models in the future.

CRediT authorship contribution statement

Shengyi Hua: Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Fang Yan:** Writing – review & editing, Resources, Investigation, Data curation. **Tianle Shen:** Validation, Data curation. **Xiaofan Zhang:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62301311).

References

Abbet, C., Studer, L., Fischer, A., Dawson, H., Zlobec, I., Bozorgtabar, B., Thiran, J.P., 2022. Self-rule to multi-adapt: Generalized multi-source feature learning using unsupervised domain adaptation for colorectal cancer tissue detection. Med. Image Anal. 79, 102473.

Aryal, M., Yahyasoltani, N., 2023. Context-aware self-supervised learning of whole slide images. arXiv preprint arXiv:2306.04763.

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

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25 (8), 1301–1309.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. arXiv:2104.14294.

Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., Williams, M., Vaidya, A., Sahai, S., Oldenburg, L., Weishaupt, L.L., Wang, J.J., Williams, W., Le, L.P., Gerber, G., Mahmood, F., 2023. A general-purpose self-supervised model for computational pathology. arXiv:2308.15474.

Chen, X., He, K., 2021. Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 15750–15758.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. ICML, PMLR, pp. 1597–1607.

Chen, X., Xie, S., He, K., 2021. An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 9640–9649.

Ciga, O., Xu, T., Martel, A.L., 2022. Self supervised contrastive learning for digital histopathology. Mach. Learn. Appl. 7, 100198.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 248–255.

Doersch, C., Gupta, A., Efros, A.A., 2015. Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1422–1430.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Adv. Neural Inf. Process. Syst. 33, 21271–21284.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 16000–16009.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 9729–9738.

Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 1501–1510.

Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T., Zou, J., 2023. Leveraging medical Twitter to build a visual–language foundation model for pathology AI. bioRxiv.

Huang, Z., Chai, H., Wang, R., Wang, H., Yang, Y., Wu, H., 2021. Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In: Medical Image Computing and Computer Assisted Intervention. MICCAI, Springer, pp. 561–570.

Jing, L., Tian, Y., 2020. Self-supervised visual feature learning with deep neural networks: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 43 (11), 4037–4058.

Kather, J.N., Halama, N., Marx, A., 2018. 100,000 histological images of human colorectal cancer and healthy tissue.

Kawai, M., Ota, N., Yamaoka, S., 2023. Large-scale pretraining on pathological images for fine-tuning of small pathological benchmarks. arXiv preprint arXiv:2303.15693.

Koohbanani, N.A., Unnikrishnan, B., Khurram, S.A., Krishnaswamy, P., Rajpoot, N., 2021. Self-path: Self-supervision for classification of pathology images with limited annotations. IEEE Trans. Med. Imaging 40 (10), 2845–2856.

Lazard, T., Lerousseau, M., Decencière, E., Walter, T., 2023. Giga-SSL: self-supervised learning for gigapixel images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 4304–4313.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 4681–4690.

Li, B., Li, Y., Eliceiri, K.W., 2021. 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. CVPR, pp. 14318–14328.

Ling, Y., Tan, W., Yan, B., 2023. Self-supervised digital histopathology image disentanglement for arbitrary domain stain transfer. IEEE Trans. Med. Imaging.

Liu, S., Zhu, C., Xu, F., Jia, X., Shi, Z., Jin, M., 2022. BCI: breast cancer immunohistochemical image generation through pyramid pix2pix. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 1815–1824.

Lotz, J., Weiss, N., van der Laak, J., Heldmann, S., 2022. Comparison of consecutive and re-stained sections for image registration in histopathology. arXiv:2106.13150.

Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Zhang, A., Le, L.P., et al., 2023. Towards a visual-language foundation model for computational pathology. arXiv preprint arXiv:2307.12914.

Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed. Eng. 5 (6), 555–570.

Noroozi, M., Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In: The European Conference on Computer Vision. ECCV, Springer, pp. 69–84.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2023. DINOv2: learning robust visual features without supervision. arXiv:2304.07193.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2536–2544.

Pisula, J.I., Bozek, K., 2022. Language models are good pathologists: using attention-based sequence reduction and text-pretrained transformers for efficient WSI classification. arXiv preprint arXiv:2211.07384.

Sahasrabudhe, M., Christodoulidis, S., Salgado, R., Michiels, S., Loi, S., André, F., Paragios, N., Vakalopoulou, M., 2020. Self-supervised nuclei segmentation in histopathological images using attention. In: Medical Image Computing and Computer Assisted Intervention. MICCAI, Springer, pp. 393–402.

Schirris, Y., Gavves, E., Nederlof, I., Horlings, H.M., Teuwen, J., 2022. DeepSMILE: contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer. Med. Image Anal. 79, 102464.

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al., 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Adv. Neural Inf. Process. Syst. 34, 2136–2147.

Shekhar, S., Bordes, F., Vincent, P., Morcos, A., 2023. Objectives matter: understanding the impact of self-supervised objectives on vision transformer representations. arXiv preprint arXiv:2304.13089.

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

van der Laak, J., Lotz, J., Weiss, N., Heldmann, S., 2021. HyReCo-Hybrid re-stained and consecutive histological serial sections.

Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., Mathieu, P., van Eck, A., Lee, D., Viret, J., Robert, E., Wang, Y.K., Kunz, J.D., Lee, M.C.H., Bernhard, J., Godrich, R.A., Oakley, G., Millar, E., Hanna, M., Retamero, J., Moye, W.A., Yousfi, R., Kanan, C., Klimstra, D., Rothrock, B., Fuchs, T.J., 2023. Virchow: a million-slide digital pathology foundation model. arXiv:2309.07778.

Vu, Q.D., Rajpoot, K., Raza, S.E.A., Rajpoot, N., 2023. Handcrafted Histological Transformer (H2T): Unsupervised representation of whole slide images. Med. Image Anal. 85, 102743.

Wang, H., Ahn, E., Kim, J., 2023a. A dual-branch self-supervised representation learning framework for tumour segmentation in whole slide images. arXiv preprint arXiv:2303.11019.

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

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X., 2022. Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. 81, 102559.

Yang, P., Yin, X., Lu, H., Hu, Z., Zhang, X., Jiang, R., Lv, H., 2022. CS-CO: a hybrid self-supervised visual representation learning method for h&e-stained histopathological images. Med. Image Anal. 81, 102539.

Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., Wang, D., 2023. Text-guided foundation model adaptation for pathological image classification. arXiv preprint arXiv:2307.14901.

Zhang, R., Isola, P., Efros, A.A., 2016. Colorful image colorization. In: The European Conference on Computer Vision. ECCV, Springer, pp. 649–666.

Zhao, B., Han, C., Pan, X., Lin, J., Yi, Z., Liang, C., Chen, X., Li, B., Qiu, W., Li, D., et al., 2022. RestainNet: a self-supervised digital re-stainer for stain normalization. Comput. Electr. Eng. 103, 108304.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 2223–2232.