Medical Image Analysis 83 (2023) 102645

Contents lists available at ScienceDirect

Medical Image Analysis

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

RetCCL: Clustering-guided contrastive learning for whole-slide image retrieval

Xiyue Wang [a,b], Yuexi Du [c], Sen Yang [d], Jun Zhang [d], Minghui Wang [a,b], Jing Zhang [a,*], Wei Yang [d], Junzhou Huang [d], Xiao Han [d,*]

[a] College of Biomedical Engineering, Sichuan University, Chengdu 610065, China
[b] College of Computer Science, Sichuan University, Chengdu 610065, China
[c] College of Engineering, University of Michigan, Ann Arbor, MI, 48109, United States
[d] Tencent AI Lab, Shenzhen 518057, China

ARTICLE INFO

Keywords:
Histopathology
Image retrieval
Self-supervised learning
Feature extraction

ABSTRACT

Benefiting from the large-scale archiving of digitized whole-slide images (WSIs), computer-aided diagnosis has been well developed to assist pathologists in decision-making. Content-based WSI retrieval can be a new approach to find highly correlated WSIs in a historically diagnosed WSI archive, which has the potential usages for assisted clinical diagnosis, medical research, and trainee education. During WSI retrieval, it is particularly challenging to encode the semantic content of histopathological images and to measure the similarity between images for interpretable results due to the gigapixel size of WSIs. In this work, we propose a Retrieval with Clustering-guided Contrastive Learning (RetCCL) framework for robust and accurate WSI-level image retrieval, which integrates a novel self-supervised feature learning method and a global ranking and aggregation algorithm for much improved performance. The proposed feature learning method makes use of existing large-scale unlabeled histopathological image data, which helps learn universal features that could be used directly for subsequent WSI retrieval tasks without extra fine-tuning. The proposed WSI retrieval method not only returns a set of WSIs similar to a query WSI, but also highlights patches or sub-regions of each WSI that share high similarity with patches of the query WSI, which helps pathologists interpret the searching results. Our WSI retrieval framework has been evaluated on the tasks of anatomical site retrieval and cancer subtype retrieval using over 22,000 slides, and the performance exceeds other state-of-the-art methods significantly (around 10% for the anatomic site retrieval in terms of average $mMV@10$). Besides, the patch retrieval using our learned feature representation offers a performance improvement of 24% on the TissueNet dataset in terms of $mMV@5$ compared with using ImageNet pre-trained features, which further demonstrates the effectiveness of the proposed CCL feature learning method.

1. Introduction

In digital pathology, the glass slides are scanned into whole-slide images (WSIs) with high resolution and gigapixel size, which provide rich cell-level information and have been allowed for clinical diagnosis (Evans et al., 2018; Mukhopadhyay et al., 2018). However, visual inspection on the entire WSI is very labor-intensive and time-consuming. Computational pathology based on deep learning technologies has been emerged to facilitate the automation process of pathology diagnoses, such as classification of cancer types (Campanella et al., 2019; Lu et al., 2021; Xue et al., 2021), delineation of cancerous or nuclear regions (Kumar et al., 2017), survival prediction (Shao et al., 2020), image retrieval (Kalra et al., 2020a), etc. Benefiting from the increasing amount of WSIs, WSI retrieval has recently attracted growing attention (Chen et al., 2021; Kalra et al., 2020a,b), which can return a series of similar WSIs from a historically characterized database when given a WSI for a query. These retrieved WSIs with associated diagnosis information can help provide high interpretability, making it possible in clinical diagnosis, medical research, and trainee education. For example, WSI retrieval can improve diagnostic accuracy (especially for a rare case) by finding cases with similar morphological features, which may provide a possible virtual peer review to help build a computational consensus.

Content-based image retrieval (CBIR) algorithm is a potential solution for medical image retrieval which contains two stages: image feature extraction and similar image retrieval on a pre-built database (Hegde et al., 2019; Li et al., 2018). If the extracted features in

* Corresponding authors.
E-mail addresses: jing_zhang@scu.edu.cn (J. Zhang), haroldhan@tencent.com (X. Han).

https://doi.org/10.1016/j.media.2022.102645
Received 2 December 2021; Received in revised form 21 July 2022; Accepted 27 September 2022
Available online 1 October 2022
1361-8415/© 2022 Elsevier B.V. All rights reserved.

the first stage cover the descriptive visual property of the image, similar image retrieval can be regarded as a nearest-neighbor finding problem, which indicates that a descriptive and robust data representation is the core task of the CBIR task (Kalra et al., 2020a; Tizhoosh et al., 2021).

However, for the content-based WSI retrieval (WSI-CBIR), the gigapixel size of WSI makes both the content feature extraction and interpretability of searching results challenging. (1) Effective feature extraction for semantic content in histopathological images is very challenging due to the enormous heterogeneity within WSIs and intra-/inter-class variations across WSIs. Moreover, WSI-level annotation usually targets a tiny proportion of tissues within the WSI (called a weak annotation). A pan-cancer and annotation-free feature extractor is urgently required to overcome these issues to extract robust feature representations. (2) For the WSI retrieval, it is more desirable to find WSIs in which there exist diagnosis-relevant regions/patches rather than retrieving WSIs with global similarity. Moreover, these target patches may occupy a tiny part of the gigapixel WSI. These characteristics make the task of WSI retrieval very challenging. A possible trick is to perform local patch-by-patch retrieval and then globally aggregate these patch retrieval results to return associated similar WSIs. However, due to the sheer size of WSIs and their unbalanced tissue type distribution, it is very challenging to develop a proper global aggregation algorithm.

Current histopathological image retrieval methods usually split WSIs into patches and perform the patch-level retrieval (Ma et al., 2016, 2018; Shi et al., 2017; Zhang et al., 2014; Zheng et al., 2017), which requires exhaustive annotation for these sub-regions and could not be flexibly expanded to WSI retrieval due to the lack of efficient patch aggregation methods. An early WSI retrieval method directly concatenated all the patch features as the global WSI embedding to find similar WSIs by the nearest neighbor searching. However, the overall WSI-level comparison approach equally treats tissue types and fails to focus on clinically relevant sub-regions within the WSI. Two recent studies have proposed suitable patch aggregation algorithms for WSI retrieval. The difference is that Yottixel (Kalra et al., 2020a,b) recognized WSIs through the "median-of-min" ranking approach, and FISH (Chen et al., 2021) developed a nearest neighbor approach based on the Van-Emde Boas-tree for the WSI retrieval. However, their features depend entirely or partly on the ImageNet data, which may result in suboptimal performance due to the domain difference between natural and pathological images. Thus, an effective in-domain feature extractor is urgently required to improve the feature extraction ability for histopathological images, ideally, in an unsupervised manner. Self-supervised learning (SSL) without manual annotation has become a promising method to improve the feature representation ability for the histopathological image analysis (Dehaene et al., 2020; Koohbanani et al., 2021; Li et al., 2021; Lu et al., 2019; Srinidhi et al., 2021). However, these methods have not trained on a large-scale and diverse domain-specific database. Meanwhile, their utilized standard contrastive learning methods (e.g., SimCLR (Li et al., 2021) and MoCo (Dehaene et al., 2020)) assume each sample is an individual instance. When applied to WSIs, it may cause serious bias due to the extremely unbalanced tissue type distribution and a large portion of similar tissues within/across WSIs. For histopathological images, negative pairs in the contrastive learning setting may be composed of highly related samples, which could confuse the network training process. In summary, for the broader application of WSI retrieval, there is a need for robust content feature extraction in an unsupervised manner and a global aggregation approach on the local patch retrieval results to find the most similar WSIs.

To overcome the above-mentioned problems, this work proposes a WSI retrieval framework (RetCCL) based on (1) clustering-guided contrastive learning (CCL) for feature extraction and (2) distinctive query patch selection, ranking for searched patches, and aggregation algorithm for interpretable WSI searching. In the first stage, we propose a CCL method to alleviate the effect of unfair assumption in traditional contrastive learning, where we use a subqueue-based weighted InfoNCE and a between-instance-based group-level InfoNCE to learn robust feature representations both at the instance-level and cluster-level. In the second stage, we represent the entire WSI by combining distinctive patches that are obtained by unsupervised feature-based and space-based clustering approaches. Due to the unbalanced tissue type distribution within WSIs, we perform a patch-by-patch retrieval instead of the entire WSI searching to retrieve diagnosis-relevant sub-regions/patches within similar WSIs. These retrieved patches are curated by our ranking and aggregation algorithm depending on the entropy-based uncertainty measurement and cosine-similarity-based constraint. The final retrieved patches are associated with their source WSIs to obtain the most similar WSIs. Additionally, we also show that RetCCL can perform patch retrieval to directly return a series of relevant sub-regions when pathologists provide a sub-region as a query.

The main contributions of our work can be summarized as follows:

- We propose a novel WSI retrieval algorithm called RetCCL, which includes a novel CCL-based feature extractor and a ranking and aggregation algorithm for WSI retrieval. It can also provide interpretable results by highlighting the diagnosis-relevant sub-regions within WSIs to explain the searching mechanism behind our WSI retrieval algorithm.
- Our CCL-based feature extractor is designed by integrating a subqueue-based weighted InfoNCE and a between-instance-based group-level InfoNCE into traditional contrastive learning to balance the ratio of positive/negative samples and map similar images closer.
- Our CCL pretraining is conducted using currently the largest histopathological image database (around 15 million patches cropped from more than 32,000 WSIs), covering diverse cell and tissue types, cancer diagnoses, and organs, which helps extract a pan-cancer feature extractor for WSI-CBIR.
- Benefiting from the above designs, RetCCL outperforms existing WSI retrieval methods by a large margin. Our CCL-based feature is also superior to the ImageNet pretrained feature or other SSL-based features, which is verified in the patch retrieval experiment. Our best-pretrained model has been released,[1] which has the potential to be a new feature extractor for various histopathological image applications to replace the current widely used ImageNet pretrained model.

2. Related work

This section conducts a literature review for self-supervised representation learning and histopathological image retrieval considering their relevance to our work.

2.1. Self-supervised representation learning

Self-supervised learning (SSL) aims to obtain high-level feature representation by solving various pretext tasks using the supervision of data itself. Early used pretext tasks include jigsaw puzzles (Noroozi and Favaro, 2016), context prediction (Doersch et al., 2015), image rotation prediction (Gidaris et al., 2018), image colorization (Zhang et al., 2016), input re-construction (Van Den Oord et al., 2017), etc., which encourage the model to learn covariant feature representations rather than invariant ones (Misra and Maaten, 2020). More recently, contrastive learning as a special pretext task (instance discrimination) has proven its superiority in the natural image field with the performance even better than its corresponding supervised counterpart (Chen et al., 2020b; He et al., 2020). It adopts two data augmentations from the same image as a positive pair and those from different images as negative pairs. The goal of network optimization is to pull together positive samples and repel away negative samples. Further improvement

[1] https://github.com/Xiyue-Wang/RetCCL

based on these methods tried to use a stronger backbone network like MOBY (Xie et al., 2021) with Swin transformer (Liu et al., 2021), or relax the strong assumption of positive/negative pairs by introducing more robust objective functions (Foster et al., 2021; Tsai et al., 2021).

SSL has not yet been fully investigated in the field of histopathological image analysis (Dehaene et al., 2020; Koohbanani et al., 2021; Li et al., 2021; Lu et al., 2019; Srinidhi et al., 2021). These studies include simple application of current SSL techniques (e.g., SimCLR (Li et al., 2021), MoCo (Dehaene et al., 2020), and CPC (Lu et al., 2019)) or their customized pretext tasks (e.g., magnification prediction (Koohbanani et al., 2021; Srinidhi et al., 2021), solving magnification puzzle (Koohbanani et al., 2021), and hematoxylin channel prediction (Koohbanani et al., 2021)). However, there are still three aspects that could be further improved. First, most of these studies simply transfer techniques from natural image applications to the histopathology domain, which ignores the domain shift between natural and histopathological images. Second, even if some pretexts mentioned above are specifically designed for histopathological images, they focus on the relevant semantic features within the specific tasks, resulting in limited model generalization. Defining a universal pretext task for various histopathological image analysis tasks is still challenging. Third, these works were not tested on large and diverse histopathological image datasets.

2.2. Histopathological image retrieval

Histopathological image retrieval is driven by the prosperously developed computer technology and the rapidly increasing number of histopathological images, which adopts a small image patch (cropped from WSI) or entire WSI as a query to retrieve images with similar semantics from a pre-constructed historical database. We will review current related studies from two aspects: patch-level and WSI-level image retrieval.

A majority of histopathological image retrieval studies focus on patches within WSIs (Ma et al., 2016, 2018; Shi et al., 2017; Zhang et al., 2014; Zheng et al., 2017). Traditional hand-craft features are adopted to describe morphological and textural characteristics of patches, including SIFT descriptor (Zhang et al., 2014), color histogram based feature (Ma et al., 2018), local statistical feature of nuclei (Ma et al., 2016; Zheng et al., 2017), Gabor feature (Ma et al., 2016; Zheng et al., 2017), GIST feature (Shi et al., 2017), and HOG descriptor (Shi et al., 2017). These features require transformation functions predefined by experts. Recently, the fast-developed deep learning method has facilitated automatic high-level feature extraction (e.g., SIMLY (Hegde et al., 2019)). In some above studies, these extracted features are usually compressed by supervised hashing code to reduce the calculation and storage resources (Ma et al., 2018; Shi et al., 2017; Zhang et al., 2014). Once the feature was extracted, similarity-score-based nearest neighbor matching was usually used to find the most similar images.

There are only a few studies to be reported for WSI retrieval since it is difficult to encode a full WSI using one feature vector (Akakin and Gurcan, 2012; Chen et al., 2021; Kalra et al., 2020a,b). Akakin and Gurcan (2012) manually picked core patches from WSI by pathologists and extracted color and texture features for each patch, such as the statistics information from various color spaces and histograms. These patch features are accumulated as the WSI representation. A support vector machine (SVM) was used to classify these WSI features into two major disease types. Once the disease type was determined, the sequential image retrieval could be performed among the specific WSIs with the same disease types. However, both the patch selection and SVM training need manual intervention. Recently emerged Yottixel (Kalra et al., 2020a,b) and FISH (Chen et al., 2021) approaches aim to choose the most distinctive patches (called a mosaic) to represent the entire WSI, which was achieved by K-means clustering based on the RGB histogram and spatial coordinate features. Yottixel utilized the ImageNet pretrained DenseNet to extract the patch embeddings and then convert them into binary barcodes for fast retrieval. Except for the same features as Yottixel, FISH pretrained a VQ-VAE on TCGA to code an integer index for each patch for retrieval. However, the simple usage of the feature extractor pretrained from natural images may result in suboptimal performance due to the domain shift between natural and histopathological images.

3. Methods

The overview of our proposed WSI retrieval framework (RetCCL) is presented in Fig. 1, which is implemented using a two-stage strategy, including the CCL-based feature extractor in Fig. 1A and the WSI retrieval process in Fig. 1B. The first stage introduces two loss functions (weighted InfoNCE and group-level InfoNCE) to help extract robust and universal features. The second stage is performed in two steps: (1) offline database construction for WSI retrieval and (2) online WSI query process. In addition to the WSI retrieval, our retrieval framework can also be used for patch-level retrieval.

3.1. Contrastive-learning-based feature extractor

3.1.1. Preliminary of contrastive learning

Given an image $\boldsymbol{x}$ and its two different augmented views: $\boldsymbol{x}_q$ and $\boldsymbol{x}_k$, a self-supervised contrastive learning method aims at pulling closer the features ($\boldsymbol{q}$ and $\boldsymbol{k}^+$) of views from the same image while repelling away the features ($\boldsymbol{q}$ and $\boldsymbol{k}^-$) of views from different images (Chen et al., 2020b; Chen and He, 2020; Grill et al., 2020; He et al., 2020; Wang et al., 2021), which is defined by the InfoNCE loss $\mathcal{L}_{\text{InfoNCE}}$ (Van Den Oord et al., 2018) and is also used as the loss function in the MoCo method (Chen et al., 2020a; He et al., 2020).

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

where $\boldsymbol{q}$ and $\boldsymbol{k}^+$ are the encoded feature vectors of $\boldsymbol{x}_q$ and $\boldsymbol{x}_k$, while $\boldsymbol{k}_i^-$ represents the ith negative sample from a memory bank. Temperature τ is the hyperparameter that adjusts the smoothness of the loss function. L denotes the length of the memory bank that is used to store negative samples. The $\cdot$ denotes a dot product between two feature vectors. f and h are shared feature extractors, where f as a momentum encoder (He et al., 2020) is updated by

$$\boldsymbol{\theta}' \leftarrow m\cdot\boldsymbol{\theta}' + (1-m)\cdot\boldsymbol{\theta} \tag{2}$$

where the $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$ are the parameters of f and h, respectively.

3.1.2. Proposed clustering-guided contrastive learning

Despite the success of self-supervised contrastive learning in recent years, its basic assumption about the negative samples is not suitable for WSI. As shown in Eq. (1), all samples in the memory bank are regarded as negative samples. However, just like Wang et al. (2021) have claimed, there may exist possible highly correlated samples, which should be considered as positive samples with respect to the anchor sample $\boldsymbol{x}_q$, while they are repelled from the anchor sample in the setting of the standard contrastive learning. Such a problem will be even more serious in the context of histopathological images, where different patches may actually come from the same tissues and have strikingly similar appearances.

To alleviate this problem, we propose a clustering-based contrastive learning method that includes a weighted InfoNCE ($\mathcal{L}_{\text{W-InfoNCE}}$) and a group-level InfoNCE ($\mathcal{L}_{\text{G-InfoNCE}}$) to modify the definition of positive/negative samples in the standard contrastive learning. The $\mathcal{L}_{\text{W-InfoNCE}}$ is defined based on a subqueue strategy to reduce the effect of possible false-negative samples in contrastive learning. Benefiting from Wang et al. (2021), the $\mathcal{L}_{\text{G-InfoNCE}}$ obtains distinctive group centers and encourages the anchor sample and its nearest group center to have higher similarity while enforcing the anchor sample and the remaining group centers to have a lower similarity.

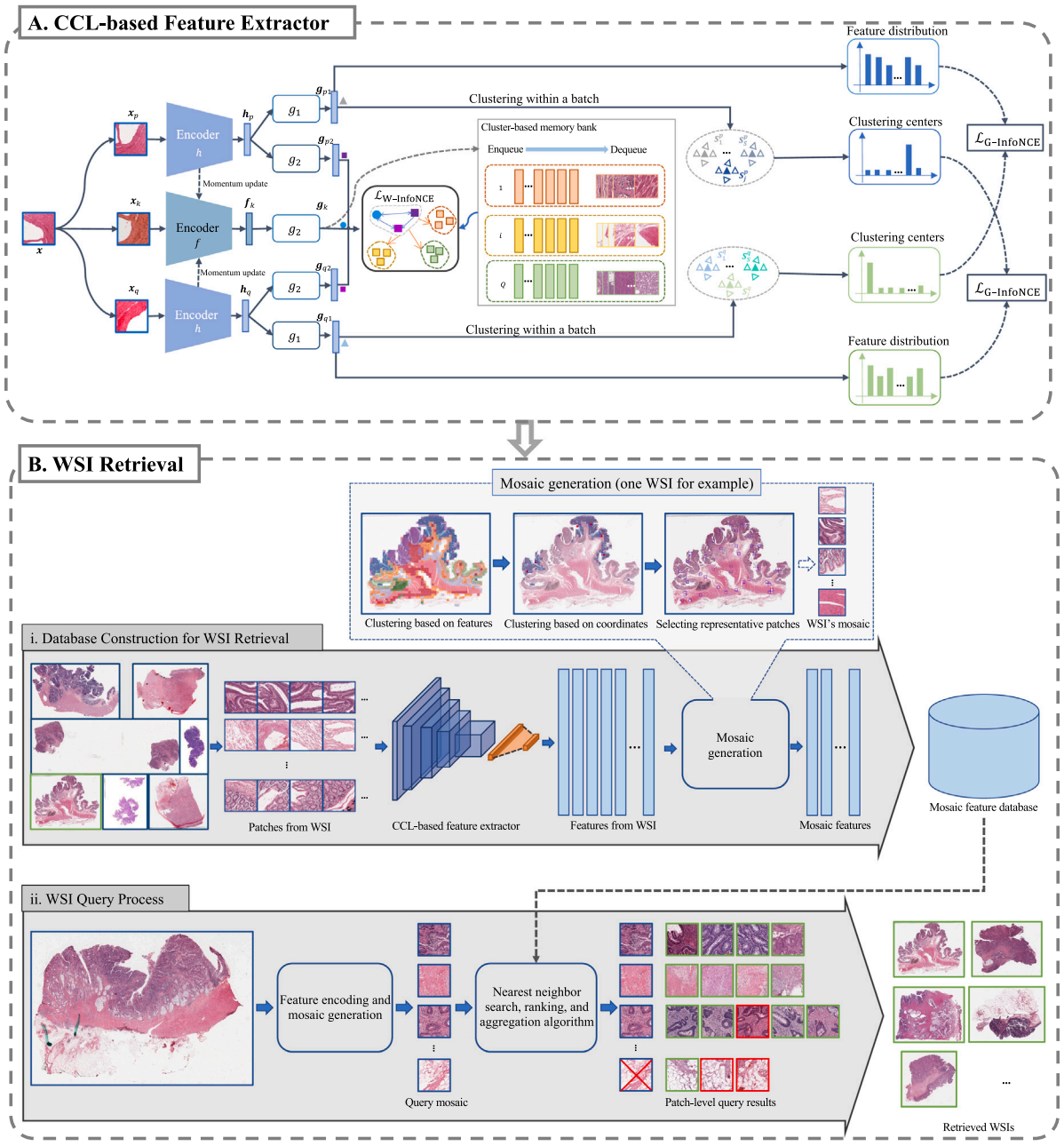

Fig. 1. Overview of our proposed WSI retrieval architecture (RetCCL), which contains two stages: (A) CCL-based feature extractor and (B) WSI retrieval. In (A), weighted InfoNCE and group-level InfoNCE loss functions are integrated for a better positive/negative ratio. The clustering-guided memory bank contains Q smaller sub-memory queues. New input features $\boldsymbol{g}_k$ will be queued into the queue (cluster) that is most similar to $\boldsymbol{g}_k$. The group-level InfoNCE loss is calculated by swapped predictions to map similar images closer. (B) The WSI retrieval procedure consists of two steps: (i) offline database construction for all reference WSIs and (ii) online WSI query process. In (i), two techniques are mainly leveraged. The first technique is the newly designed CCL-based feature extraction that makes use of existing large-scale public histopathological image databases. The second is the mosaic generation technique based on a double-clustering method applied within a WSI to obtain its most representative patches. The first K-means clustering relies on our CCL-based features. The coordinates of patches in each cluster obtained in the first clustering are used as features for the second clustering. These patches closest to the clustering centers are selected as the final WSI representation. The red dots in "Clustering based on coordinates" denote the centroids of the final clusters. The small blue boxes in "Selecting representative patches" represent the selected patches within the WSI. In (ii), the WSI query step contains similar feature extraction and mosaic generation procedures as introduced in (i). Then these selected patches within the WSI are respectively used to complete the retrieval task and return a set of similar patches from the pre-built database. Finally, the nearest neighbor search, ranking, and aggregation algorithms are applied to determine the final WSI-level retrieval results.

The overview of our proposed CCL method is shown in Fig. 1A. It augments input image $\boldsymbol{x}$ three times and obtains $\boldsymbol{x}_p$, $\boldsymbol{x}_k$, and $\boldsymbol{x}_q$. Then, two encoders f and h with the same structures but slightly different parameters are applied to transfer these input images to high-level semantic space $\mathbb{R}^d$, resulting in $\boldsymbol{h}_p$, $\boldsymbol{f}_k$, and $\boldsymbol{h}_q$. The encoder branch f is updated by moving average as mentioned above. Then, two MLP heads g_1 and g_2 are used to project the $\boldsymbol{h}_p$ into $\{\boldsymbol{g}_{p_1}, \boldsymbol{g}_{p_2}\}$ and the $\boldsymbol{h}_q$ into $\{\boldsymbol{g}_{q_1}, \boldsymbol{g}_{q_2}\}$, respectively. g_2 is also used to project $\boldsymbol{f}_k$ into $\boldsymbol{g}_k$. The $\boldsymbol{g}_{p_2}$, $\boldsymbol{g}_{q_2}$, and $\boldsymbol{g}_k$ are further combined to calculate our weighted InfoNCE loss, while $\boldsymbol{g}_{p_1}$ and $\boldsymbol{g}_{q_1}$ are used to cluster within a batch for the calculation of group-level InfoNCE loss. The final loss function is the combination of the two loss functions, which will be introduced in the following.

Online Clustering-guided Memory Bank Construction. Based on the basic contrastive learning framework mentioned earlier (He et al., 2020), we introduce an online clustering-guide memory bank to reduce the influence of potential false-negative samples. A weighted InfoNCE loss is proposed to give less weight on these false-negative-like samples

with respect to the anchor feature embedding $\boldsymbol{g}_{p_2}$ or $\boldsymbol{g}_{q_2}$ as shown in Fig. 1A. The detailed calculation process will be introduced as follows.

In each training epoch, as shown in Fig. 1A, we respectively take $\{\boldsymbol{g}_{p_2}, \boldsymbol{g}_k\}$ and $\{\boldsymbol{g}_{q_2}, \boldsymbol{g}_k\}$ as two positive pairs to perform contrastive learning with a shared memory bank. Negative samples are weighted using a clustering process. Specifically, all negative samples within the memory bank are first clustered into Q classes using the K-means approach, which are called Q sub-memory queues. Their centroids are represented as $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_j, \ldots, \boldsymbol{c}_Q\}$. Next, the similarity scores between the input feature $\boldsymbol{g}_k$ and each centroid $\boldsymbol{c}_j (j = 1, 2, \ldots, Q)$ are calculated as $\{Sim_1, \ldots, Sim_j, \ldots, Sim_Q\}$. The maximum of these similarity scores can be obtained as Sim_{max}, which corresponds to the cluster Q_{max} whose centroid is most similar to $\boldsymbol{g}_k$. Then, the weight $\phi(\boldsymbol{g}_{k_i^-})$ for each negative sample in the memory bank can be calculated by

$$\phi(\boldsymbol{g}_{k_i^-}) = \begin{cases} w, \text{ if } \boldsymbol{g}_{k_i^-} \in Q_{max} \\ 1, \text{ otherwise} \end{cases}, \text{ where } w \in [0, 1) \tag{3}$$

That is, the weighted InfoNCE will assign a smaller weight ($w \in [0, 1)$) for these negative samples within the cluster Q_{max} that is most similar to $\boldsymbol{g}_k$. For other samples in the memory bank, a higher weight of 1 is set. Accordingly, the weighted InfoNCE loss $\mathcal{L}_{\text{W-InfoNCE}}$ can be defined as

$$\begin{aligned} &\mathcal{L}_{\text{W-InfoNCE}} = \\ &-\frac{1}{2}\log\frac{\exp(\boldsymbol{g}_{p_2}\cdot\boldsymbol{g}_k/\tau)}{\exp(\boldsymbol{g}_{p_2}\cdot\boldsymbol{g}_k/\tau)+\sum_{i=1}^{L}\exp(\phi(\boldsymbol{g}_{k_i^-})\cdot\boldsymbol{g}_{p_2}\cdot\boldsymbol{g}_{k_i^-}/\tau)} \\ &-\frac{1}{2}\log\frac{\exp(\boldsymbol{g}_{q_2}\cdot\boldsymbol{g}_k/\tau)}{\exp(\boldsymbol{g}_{q_2}\cdot\boldsymbol{g}_k/\tau)+\sum_{i=1}^{L}\exp(\phi(\boldsymbol{g}_{k_i^-})\cdot\boldsymbol{g}_{q_2}\cdot\boldsymbol{g}_{k_i^-}/\tau)} \end{aligned} \tag{4}$$

where the two terms correspond to taking $\{\boldsymbol{g}_{p_2}, \boldsymbol{g}_k\}$ and $\{\boldsymbol{g}_{q_2}, \boldsymbol{g}_k\}$ as the positive pairs, respectively.

At the beginning of the training process, to initialize the Q centroids of these sub-memory queues, we randomly feed T histopathological images into our encoder that is initialized by its pretrained weights on the ImageNet data. These features are then clustered into Q initial clusters, which produces Q different classes. During each iteration, each cluster centroid is updated by

$$\boldsymbol{c}_{j^*} \leftarrow m_c\boldsymbol{c}_j + (1 - m_c)\cdot\frac{1}{\left|\mathcal{B}_j\right|}\sum_{\boldsymbol{g}_k^i\in\mathcal{B}_j}\boldsymbol{g}_k^i \tag{5}$$

where $\boldsymbol{c}_{j^*}$ denotes the updated jth cluster centroid, $m_c \in [0, 1]$ represents a weighting factor, $\mathcal{B}_j$ denotes the feature set of the jth class (cluster) in the current mini-batch, $\boldsymbol{g}_k^i$ represents the ith feature vector in the mini-batch, and $1/\left|\mathcal{B}_j\right|(\sum_{\boldsymbol{g}_k^i\in\mathcal{B}_j}\boldsymbol{g}_k^i)$ calculates the mean value of features for the jth class in the current mini-batch. At each epoch, all clustering centroids will be updated by re-clustering all negative samples in the memory bank.

Group-level Discrimination. Inspired by the cross-level discrimination (CLD) loss function (Wang et al., 2021), we add the CLD as an auxiliary branch to further mitigate the unbalanced positive/negative sample ratio brought by the basic assumption of instance discrimination in standard contrastive learning. As shown in Fig. 1A and explained earlier, an MLP head g_1 is attached to the encoder h to obtain new embeddings $\boldsymbol{g}_{p_1}$ and $\boldsymbol{g}_{q_1}$, corresponding to the two augmented views, x_p and x_q, respectively. Embeddings from all samples in one mini-batch are then clustered into S clusters for each of the two augmented view branches, and their centroids are denoted respectively as $\boldsymbol{S}_j^p$ and $\boldsymbol{S}_j^q$, where $j \in [1, 2, \ldots, S]$. Then, given a query instance, we compare its augmented view embedding ($\boldsymbol{g}_{p_1}$ or $\boldsymbol{g}_{q_1}$) with cluster centroids in the opposite branch to define positive and negative pairs for computing the group-level InfoNCE loss. For example, given $\boldsymbol{g}_{p_1}$, its positive sample is defined as the closest centroid (denoted as $\boldsymbol{S}^{q+}$) in the x_q branch. The remaining S-1 centroids are defined as negative samples, denoted as $\boldsymbol{S}_i^{q-}$, where $i \in [1, 2, \ldots, S-1]$. In summary, the group-level InfoNCE loss $\mathcal{L}_{\text{G-InfoNCE}}$ (CLD loss) is given by

$$\begin{aligned} &\mathcal{L}_{\text{G-InfoNCE}} = \\ &-\frac{1}{2}\log\frac{\exp(\boldsymbol{g}_{p_1}\cdot\boldsymbol{S}^{q+}/\tau)}{\exp(\boldsymbol{g}_{p_1}\cdot\boldsymbol{S}^{q+}/\tau)+\sum_{i=1}^{S-1}\exp(\boldsymbol{g}_{p_1}\cdot\boldsymbol{S}_i^{q-}/\tau)} \\ &-\frac{1}{2}\log\frac{\exp(\boldsymbol{g}_{q_1}\cdot\boldsymbol{S}^{p+}/\tau)}{\exp(\boldsymbol{g}_{q_1}\cdot\boldsymbol{S}^{p+}/\tau)+\sum_{i=1}^{S-1}\exp(\boldsymbol{g}_{q_1}\cdot\boldsymbol{S}_i^{p-}/\tau)} \end{aligned} \tag{6}$$

Our final loss function is calculated as the sum of the weighted InfoNCE loss and the group-level InfoNCE loss by

$$\mathcal{L} = \mathcal{L}_{\text{W-InfoNCE}} + \lambda\mathcal{L}_{\text{G-InfoNCE}} \tag{7}$$

where λ is a hyperparameter that controls the contribution of the two loss functions.

3.2. WSI retrieval method

3.2.1. Preliminary of WSI retrieval

Due to the unique WSI characteristics, WSI-CBIR is usually implemented in two stages: offline WSI feature extraction and similar WSI searching. For feature extraction, a DenseNet pre-trained on ImageNet data in a supervised manner is used in the two existing WSI retrieval methods (Yottixel (Kalra et al., 2020a) and FISH (Chen et al., 2021)). The Yottixel method solely uses the pretrained DenseNet to extract patch features. In the FISH method, the pre-trained DenseNet is used as the feature encoder to generate discrete latent code, whereas another encoder of VQ-VAE pre-trained on TCGA data is used to generate texture features. In both methods, the extracted features are further compressed into binary codes for faster retrieval. Once the patch feature extractor is determined, the subsequent WSI searching can be performed in two steps: database construction and runtime searching. First, each WSI is reduced to a set of patches, for which a dual clustering method is used in both the Yottixel and FISH methods. The feature vectors of these patches are used to represent each WSI and a searching database can then be constructed. During the runtime searching process, a query WSI is also converted into a set of patches and a patch-by-patch matching is performed to retrieve the most similar patches in the database, which are then projected to find corresponding most relevant WSIs. To reduce false matches, Yottixel proposes a simple "median-of-min" approach and computes a Hamming distance as the matching score for each retrieved patch. Those retrieved patches with Hamming distances below a threshold score will be kept, and others are removed. The FISH method develops a more sophisticated searching and ranking approach that combines a Van Emde Boas tree with an uncertainty-based ranking algorithm to achieve fast and accurate WSI retrieval. Retrieved patches with high uncertainty and low cosine similarity scores are removed before selecting corresponding WSIs.

3.2.2. Proposed WSI retrieval method

Our WSI retrieval algorithm differs from the previous Yottixel and FISH methods in two aspects. First, we employ our histopathology-image-pretrained real-valued features for patch representation, whereas previous methods mainly use ImageNet-pretrained binary features. The simple use of out-of-domain data may lead to suboptimal performance due to the large domain difference between histopathological and natural images. Our in-domain features ensure more informative and histopathology-specific feature representations. Second, we propose a simple but effective searching, ranking, and aggregation algorithm for WSI retrieval. Unlike the simple "median-and-min" strategy used by Yottixel, our method fully considers the characteristics of WSIs to remove uninformative patches, which ensures more accurate WSI retrieval results. FISH proposes a complex codebook-based vEB tree retrieval, which first utilizes the vEB tree to find roughly similar patches and then uses Hamming distance to find the best matches. Thus, its

Algorithm 1 Database Construction for WSI Retrieval

$D_s \leftarrow 16$ ▷ Downsample for segmentation
$MPP \leftarrow 1.0$ ▷ Magnification for patching
$S_p \leftarrow 512$ ▷ Size of patch
$K_1 \leftarrow 9$ ▷ First clustering number
$R \leftarrow 0.2$ ▷ Second clustering ratio
$\boldsymbol{G} \leftarrow \{\}$ ▷ Set of all selected features
for $\boldsymbol{I} \in$ All WSIs **do**
  procedure MosaicGeneration($\boldsymbol{I}, D_s, MPP, S_p, K_1, R$)
    $\boldsymbol{f} = \{\}$
    $\boldsymbol{S} \leftarrow segment(\boldsymbol{I}, D_s)$ ▷ Foreground segment for WSI
    $\boldsymbol{p} \leftarrow patching(\boldsymbol{S}, MPP, S_p)$ ▷ Obtain all patches
    $\boldsymbol{f}_{all} \leftarrow model(\boldsymbol{p})$ ▷ Obtain all features
    $\boldsymbol{F}_i \leftarrow$ FeatureKMeans($\boldsymbol{f}_{all}, K_1$) ▷ Feature clustering
    for $i \in K_1$ **do** $where, i = 1, 2, ..., K_1$
      $\boldsymbol{f}_{rep} \leftarrow$ SpatialKMeans($\boldsymbol{F}_i, R$) ▷ Coordinate clustering
      $\boldsymbol{f} \leftarrow \boldsymbol{f} \cup \boldsymbol{f}_{rep}$
    end for
    return $\boldsymbol{f}$ ▷ Return representative patch features
  end procedure
  $\boldsymbol{G} \leftarrow \boldsymbol{G} \cup \boldsymbol{f}$
end for
return $\boldsymbol{G}$ ▷ Return mosaic database for WSI retrieval

Algorithm 2 WSI Query Process

$\boldsymbol{w}_y \leftarrow w_{y_1}, ..., w_{y_m}, ..., w_{y_U}$ ▷ Weight of each diagnosis in the database
$WSI = \{\boldsymbol{P}_1, \boldsymbol{P}_2, ..., \boldsymbol{P}_i ..., \boldsymbol{P}_k\}$ ▷ Given a query WSI with k patches
$Bag = \{\mathbb{B}_1, \mathbb{B}_2, ..., \mathbb{B}_i ..., \mathbb{B}_k\}$ ▷ A bag contains a query patch and its retrieved patches
$\mathbb{B}_i = \{\boldsymbol{b}_i^1, \boldsymbol{b}_i^2, \boldsymbol{b}_i^j ..., \boldsymbol{b}_i^t\}$ ▷ Each query patch retrieves t patches
procedure Calculate entropy for each bag
  for $\mathbb{B}_i \in Bag$ **do** ▷ A bag $\mathbb{B}_i$ has u_i associated WSI diagnosis
    $\mathbb{D} = CosineSimilarity(\boldsymbol{P}_i, \mathbb{B}_i)$ ▷ Calculate cosine similarity, where $\mathbb{D} = \{d^1, d^2, ..., d^j ..., d^t\}$
    $p_m = Probability(\boldsymbol{w}_y, \mathbb{D}, \mathbb{B}_i)$ ▷ Probability calculated for the m^{th} diagnosis occurrence within a bag
    $Ent_i = -\sum_{m=1}^{u_i} p_m \cdot \log(p_m)$ ▷ Entropy within a bag
  end for
  $Bag' = \{\mathbb{B}_1, \mathbb{B}_2, \mathbb{B}_i ..., \mathbb{B}_k\}$ ▷ Reorder bag by entropy
end procedure
procedure Remove bags with low quality
  $\eta = \frac{1}{k}\sum_{i=1}^{k} AveTop\{\mathbb{B}_i\}$ ▷ Means of cosine similarity scores in top-5
  $Bag'' = \{\mathbb{B}_1, \mathbb{B}_2, \mathbb{B}_i ..., \mathbb{B}_{k'}\}$ ▷ Remove bags with small η
end procedure
for $\mathbb{B}_i \in Bag$ **do** ▷ Vote for each diagnosis within a bag
  $\mathbb{B}_i = \{\boldsymbol{b}_i^1, \boldsymbol{b}_i^2, \boldsymbol{b}_i^j ..., \boldsymbol{b}_i^5\}$ ▷ Obtain the top-5 samples in each bag
  $\boldsymbol{W}_i \leftarrow \mathbb{B}_i$ ▷ Majority vote to obtain associated WSI for each bag
end for
WSIRet $= \{\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_i ..., \boldsymbol{W}_{k''}\}$ ▷ Find similar WSIs
return WSIRet[$1 : k$] ▷ Return top-k similar WSIs

retrieval results depend heavily on the accuracy of these roughly similar patches. In addition, many parameters are used in the tree (e.g., threshold of Hamming distance and the number of times for addition and deduction, etc.), and careful parameter setting may be required when applying them to new data. Unlike FISH, our method directly uses cosine-similarity-based nearest neighbor searching to find the most similar patches, which has no additional handcrafted parameters and can be directly applied to any new data.

The overall procedure of our WSI retrieval has been described in Fig. 1B, which contains a two-step operation. In the first step, as shown in Fig. 1B, WSIs are first cropped into patches and these patches are encoded into content-relevant representations (patch features) using our customized CCL-based feature extractor. For each WSI, CCL-feature-based and spatial-coordinate-based clustering algorithms are used to generate a small number of distinctive patches (called a mosaic) to represent the full WSI. An effective mosaic generation can greatly reduce the storage and calculation burden. In the second step, when querying a WSI, our method first extracts feature representations for patches within the WSI, and then performs mosaic generation to obtain a representative mosaic. Next, these obtained patches are respectively adopted as the new query images to conduct patch-level retrieval using a nearest neighbor searching method. Finally, based on these retrieved patches along with their meta-information, our ranking and aggregation algorithm is used to find their associated similar WSIs. The meta-information includes the WSI name and its original diagnosis, patch locations within its associated WSI, and the similarity score between each patch and its corresponding query patch. Benefiting from the meta-information, our RetCCL can highlight the diagnosis-relevant regions within WSIs to help provide interpretability for pathologists. The following section will introduce the two steps in detail.

Database Construction for WSI Retrieval. A large database is required to be constructed to perform feature matching. The detailed process is shown in Fig. 1B and Algorithm 1. A WSI is pre-processed by threshold technique to extract foreground tissue regions (Campanella et al., 2019) and then cropped into small patches based on a sliding window technique. These patches are then fed into our CCL model to obtain their corresponding feature vectors, denoted as $\boldsymbol{f}_{all}$. Next, a mosaic generation method is conducted within a WSI based on a double-clustering method. In the first K-means clustering, K_1 distinctive classes within the WSI are obtained, which are represented by $\{\boldsymbol{F}_1, \boldsymbol{F}_2, \dots, \boldsymbol{F}_i, \dots, \boldsymbol{F}_{K_1}\}$. Then, each cluster is further re-clustered into K_2 sub-classes using their spatial coordinates as features, where $K_2 = round(R \cdot n)$, R is a ratio parameter and is set as 20%, and n is the number of patches within each cluster $\boldsymbol{F}_i$. Last, the patches in these final clustering centroids are adopted as the representation of the WSI.

WSI Query Process. After building the WSI database, the subsequent WSI retrieval can be regarded as a patch-level nearest neighbor matching, ranking, and aggregation process, which is illustrated in Fig. 1B and Algorithm 2. As shown in Fig. 1B, given a query WSI, feature encoding and mosaic generation are first performed to obtain the feature embedding of the query WSI. Consequently, the query WSI can be represented as a mosaic with k patches, such as $WSI = \{\boldsymbol{P}_1, \boldsymbol{P}_2, \dots, \boldsymbol{P}_i ..., \boldsymbol{P}_k\}$, where $\boldsymbol{P}_i$ denotes feature vector of the ith patch, and k denotes the total number of patches within the WSI. Next, each patch will be adopted as a query image to generate the corresponding retrieval results that are stored in k bags $Bags = \{\mathbb{B}_1, \mathbb{B}_2, \dots, \mathbb{B}_i ..., \mathbb{B}_k\}$, where the ith bag $\mathbb{B}_i = \{\boldsymbol{b}_i^1, \boldsymbol{b}_i^2, \boldsymbol{b}_i^j ..., \boldsymbol{b}_i^t\}$ contains t retrieved patches along with their cosine similarity scores calculated with the query patch $\boldsymbol{P}_i$. In other words, cosine similarity scores $\{d^1, \dots, d^j, \dots, d^t\}$ can be computed between $\boldsymbol{P}_i$ and each sample in $\mathbb{B}_i$. Note that t varies for different bags.

Based on these obtained bags and inspired by FISH (Chen et al., 2021), we propose a simple but effective ranking and aggregation algorithm to select the most promising bags. Entropy with the ability of uncertainty measure is used to calculate the uncertainty of each bag. According to the definition of entropy, if these patches within a bag have diverse distribution, a high entropy is generated, indicating a high uncertainty of the bag and its corresponding query patch, and vice versa. For example, if these retrieved patches within a bag show the same diagnosis (obtained from WSI-level meta-information), the entropy of the bag will be 0. Note that the patches in our pre-built database carry their associated meta-information (e.g., WSI name, original diagnosis at WSI level, patch location within the WSI, etc.). Thus, these retrieved patches can be considered to have pseudo-labels obtained from their associated WSI-level diagnosis information. Based on the above descriptions, our entropy is calculated by

$$Ent_i = -\sum_{m=1}^{u_i} p_m \cdot \log(p_m) \tag{8}$$

where Ent_i denotes the entropy of the ith bag $\mathbb{B}_i$. u_i represents the total number of diagnosis types within the bag $\mathbb{B}_i$, which will be the number of anatomical sites or the number of cancer subtypes in one specific cancer. p_m denotes the probability of the mth diagnosis type occurring in a bag, which is calculated as

$$p_m = \frac{\sum_{j=1}^{t} \delta(y_j, m) \cdot w_{y_j} \cdot \left(d^j + 1\right)/2}{\sum_{j=1}^{t} w_{y_j} \cdot \left(d^j + 1\right)/2} \tag{9}$$

where $y_j \in \{1, \ldots, m, \ldots, u_i\}$ represents the diagnosis information (pseudo-label) saved in the jth retrieved sample within the bag. $\delta()$ is a discriminant function that determines whether its two inputs are equal. In this scene, $\delta(y_j, m)$ judges whether the current sample belongs to the mth diagnosis type, which will output 1 if it is and 0 otherwise. w_{y_j} is the occurrence frequency (normalized probability) of the jth diagnosis type (pseudo class) in the database, which can be obtained once the WSI retrieval database is constructed. As mentioned above, d^j represents the cosine similarity between the query patch and its corresponding retrieved the jth patch within the current bag. We use $(d^j + 1)/2$ to guarantee that the range is between 0 and 1. To better understand this equation, we can regard the $w_{y_j} \cdot \left(d^j + 1\right)/2$ as a weighted score v_j to represent the jth sample within the bag. The denominator represents the sum of the scores of all samples, and the numerator represents the sum of scores for samples with only the mth diagnosis type. Based on these calculations, we can reorder these bags in descending order based on their entropy to obtain new bags $Bag' = \{\mathbb{B}_1, \mathbb{B}_2, \mathbb{B}_i ..., \mathbb{B}_k\}$.

Since patches in each bag are sorted in descending order of cosine similarity, our second ranking algorithm depends on the means of cosine similarity stored in the top-5 most similar patches, which can be described by

$$\eta = \frac{1}{k} \sum_{i=1}^{k} AveTop\{\mathbb{B}_i\} \tag{10}$$

where $\mathbb{B}_i$ and k represent the ith bag and the total number of bags, respectively. $AveTop$ denotes the average of the top-5 cosine similarity scores within the bag. η is adopted as the criterion to remove bags whose average cosine similarity scores are smaller than the η. Then a new bag sequence $Bag'' = \{\mathbb{B}_1, \mathbb{B}_2, \mathbb{B}_i ..., \mathbb{B}_{k'}\}$ can be generated. We adopt patches in the top-5 bags as the most similar retrieval results, which are projected as the most similar WSIs using the meta-information carried in each patch.

3.3. Patch retrieval method

The patch retrieval can be regarded as a special case of WSI retrieval where the number of patches within each WSI is set as one. Thus, there is no need for any complex ranking and aggregation algorithm. Similar to the WSI retrieval, there are also two steps for patch retrieval: database construction and patch query process. In the first step, the patch-level searching database is built based on diagnosed patches, the feature vectors of which are extracted using the proposed CCL method. Specifically, the finally adopted features are extracted from the output of the last pooling layer of the CCL-pretrained encoder (ResNet50). In the second step, the cosine similarity scores between an input query patch and all patches in the database are calculated and ranked from high to low. Then, the top-k similar patches are directly taken as the returns of the search engine.

4. Experimental results and discussions

This section first introduces five datasets utilized for our CCL-based pretraining, histopathological image retrieval procedures, and downstream classification. Then, the experimental setups in the training process and evaluation metrics for the image retrieval and downstream classification are described in detail. The remaining parts cover a series of validation experiments presented in terms of patch-level retrieval, WSI-level retrieval, and downstream classification. Patch-level retrieval with strong dependencies of the feature representation is conducted to validate the effectiveness of our CCL-based features. Our patch retrieval experiment includes ablation studies and cancer subtype retrieval experiments compared with other SSL-based patch retrieval methods. The WSI retrieval experiment contains anatomical site retrieval, cancer subtype retrieval, and interpretability analysis. The classification is simply conducted to verify the generalizability of our feature representations.

4.1. Datasets

The datasets utilized in this work include the cancer genome atlas dataset (TCGA), pathology AI platform (PAIP), UniToPatho, and DiagSet-A.2. In the backbone network training procedure, unlabeled patches extracted from TCGA and PAIP are used for our unsupervised pretraining. In the testing process, WSIs from the TCGA with WSI-level annotations are used to evaluate the WSI retrieval performance and the patches from UniToPatho and TissueNet datasets are used for patch retrieval assessment. It is noted that our WSI retrieval process adopts a similar number of WSI as the Yottixel (Kalra et al., 2020a,b) and FISH (Chen et al., 2021) approaches to keep a fair comparison. Our downstream classification experiment is conducted on the DiagSet-A.2 dataset. The details of each dataset are introduced below.

TCGA. TCGA[2] is a pan-cancer WSI dataset provided by the National Institutes of Health (NIH). It contains more than 30,000 WSIs (normal and cancerous) acquired from around 11,000 patients, which covers 25 anatomic sites with 32 cancer subtypes and is prepared with both frozen and diagnostic (formalin-fixed paraffin-embedded, FFPE) slides. Annotations for the anatomic sites and cancer subtypes are provided at the WSI level. After removing WSIs without magnification information, we finally collected a total of 29,763 WSIs from 10,953 patients for our SSL pretraining. In the preprocessing procedure, we first use the Otsu threshold method to obtain a binary foreground mask for each WSI. Then, all the WSIs are cropped into patches with the size of $1,024 \times 1,024$ pixels at 20×, which results in a total of 14,325,848 patches for SSL pretraining.

PAIP. PAIP[3] (Kim et al., 2021) contains 2457 WSIs, covering six cancer types (liver, renal, colorectal, prostatic, pancreatic, and cholangio cancers). It is noted that PAIP includes detailed local region annotations, but this work does not use any label information in the SSL pretraining process. Following the same patch cropping strategy as the TCGA dataset mentioned above, we obtained a total of 1,254,414 unlabeled patches with the size of $1,024 \times 1,024$ pixels for SSL pretraining.

UniToPatho. UniToPatho[4] (Barbano et al., 2021) is a patch-level dataset released to classify each colorectal polyp image into one of four types: normal tissue (NORM), hyperplastic polyp (HP), tubular adenoma (TA), and tubulo-villous adenoma (TVA). There are 8699 patches with a size of $1,812 \times 1,812$ pixels (0.4415 microns-per-pixel at 20 ×) cropped from 292 WSIs by its provider.

TissueNet. TissueNet[5] is built to classify epithelial lesions of the uterine cervix at the patch level, including benign, low malignant potential tissue, high malignant potential tissue, and invasive cancer tissue. As introduced by its provider, there are 5926 labeled patches with a fixed size of 300 × 300 micrometers cropped from 1016 WSIs (around four patches for each WSI). The size of 300 × 300 micrometers corresponds to around 1200 × 1200 pixels.

DiagSet-A.2. DiagSet-A.2[6] is a patch-level dataset for prostate cancer classification (Koziarski et al., 2021). The classification experiment

[2] https://portal.gdc.cancer.gov/
[3] http://wisepaip.org/paip
[4] https://ieee-dataport.org/open-access/unitopatho
[5] https://www.drivendata.org/competitions/67/competition-cervical-biopsy/page/254/
[6] https://ai-econsilio.diag.pl/

aims to classify each image into one of four classes: Normal, Gleason score 3, Gleason score 4, and Gleason score 5. These images have a pixel size of 224 × 224 under a magnification of 5×. The dataset provider splits the dataset into a training set (48,782 patches from 346 WSIs), a validation set (8977 patches from 42 WSIs), and a test set (10,836 patches from 42 WSIs).

4.2. Experimental setups

In our WSI retrieval architecture, only the CCL model requires network training. To keep a fair comparison with other SSL methods, we adopt ResNet50 (He et al., 2016) as the backbone model. For the construction of multiple sub-memories, the number of cluster centers is set as 25 ($Q = 25$) according to the ablation experiment. The length of each queue (sub-memory bank) is set as 2048 to make the size to be comparable with MoCo v2 (Chen et al., 2020a; He et al., 2020). $m \in [0, 1)$ is a momentum coefficient as defined in the MoCo v2 paper. Following the original MoCo v2 method, we set m to 0.999 and temperature τ to 0.07 in $\mathcal{L}_{\text{W-InfoNCE}}$. In $\mathcal{L}_{\text{G-InfoNCE}}$, S is set to 30 based on the ablation experiment. The λ in our final loss is set as 0.25 for the $\mathcal{L}_{\text{G-InfoNCE}}$.

A cosine learning rate scheduler (Loshchilov and Hutter, 2016) with an initial learning rate of 0.1 is used for model training. The batch size is set as 2048 during training. Following MoCo v2, an SGD optimizer with momentum 0.9 and weight decay of $1e-4$ is hired to update the model. Similar to SimCLR (Chen et al., 2020b), our data augmentation strategies include Gaussian blur, color jittering, horizontal/vertical flipping, and random crop and resize.

We also train other SSL methods for comparison, including SimCLR v1[7] (Chen et al., 2020b), SwAV[8] (Caron et al., 2020), and MoCo v2[9] (Chen et al., 2020a; He et al., 2020). All the models are trained with their corresponding open-source implementations and default configurations, only the augmentation methods and training data are changed to be the same as ours. All these compared SSL methods use ResNet50 as their backbone model, same as in our method. We initialize the ResNet50 backbone using its ImageNet pretrained weights before self-supervised training, as also suggested by previous studies (Yan et al., 2022; Azizi et al., 2021). For a fair comparison, we do not include SimCLR v2 and MoCo v3 because they use different backbone models. SimCLR v2 aims to improve the feature learning ability by using a much larger ResNet model (i.e., ResNet-152 (3×+SK)) as the backbone. MoCo v3 is specifically designed using a Transformer backbone.

A downstream classification experiment is conducted to further validate the robustness of our feature learning method. It is implemented using a linear evaluation, that is, a fully connected layer is attached to the output of the frozen CCL-based backbone model for classifier training. The mini-batch size is set to 48. Adam is adopted as the optimizer with an initial learning rate of 0.0003.

Our SSL pretraining experiments are implemented on the PyTorch platform and 32 Nvidia V100 32G GPU cards. Each model is trained for 100 epochs and takes roughly 300 h to converge due to a large amount of training data. To the best of our knowledge, this work is the first feature representation with such large-scale histopathological images.

4.3. Evaluation metrics

Accuracy (Acc) and F1 score are used to evaluate our downstream classification experiments. $Acc@k$ (top-k accuracy) and $mMV@k$ (majority vote at the top k search results) are used as evaluation metrics for our image retrieval task. $Acc@k$ is widely used for histopathological image retrieval (Hegde et al., 2019; Kalra et al., 2020a,b). If any of the retrieved similar images have the correct label as the query image, $Acc@k$ will think that the search task at this time is successful. Compared with $Acc@k$, $mMV@k$ is a more strict metric since $mMV@k$ thinks that if most of the retrieved results are correct compared with the query sample, this search will be considered successful. The two metrics are calculated by

$$Acc@k = \frac{1}{N}\sum_{i}^{N}\xi(\boldsymbol{a}_i, TOP(ans_i[:k])) \tag{11}$$

$$mMV@k = \frac{1}{N}\sum_{i}^{N}\delta(\boldsymbol{a}_i, MV(ans_i[:k])) \tag{12}$$

Where N is the number of query patches/WSIs and $\boldsymbol{a}_i$ is the label of the ith query patch/WSI. $TOP(ans_i[:k])$ returns top-k retrieved results. ξ is used to compare the labels of top-k results with the query annotation, which outputs 1 if any of the top-k results is matched with the query, and 0 otherwise. For example, if $TOP(ans_i[:k]) \in \boldsymbol{a}_i$, $\xi() = 1$. $MV(ans_i[:k])$ returns the predicted label by the majority vote among the top-k retrieved results. $\delta()$ is a discriminant function to judge equality as mentioned above. For instance, if $MV(ans_i[:k]) = \boldsymbol{a}_i$, $\delta() = 1$.

To maintain a fair comparison with state-of-the-art WSI retrieval methods (Chen et al., 2021; Kalra et al., 2020a), macro-average and weighted-average for the $mMV@k$ are also used in our WSI retrieval experiments. All the image retrieval validation experiments are conducted using the leave-one-patient-out strategy to avoid information leakage due to the occasional existence of multiple WSIs from the same patient.

4.4. Results of patch-level retrieval

Our patch-level retrieval experiment is conducted to evaluate the feature extraction ability of our proposed CCL-based model. As mentioned before, TissueNet and UniToPatho datasets with subtype annotations are used in the patch-level retrieval experiment. Our patch retrieval is conducted to find similar cancer subtypes. It is noted that, during the retrieval process, patches from the same WSI as the query are removed from the database to avoid data leakage problems. In this section, the performance of patch retrieval is evaluated based on $Acc@1$, $Acc@3$, $Acc@5$, and $mMV@5$ metrics. The following part will first show the ablation experiments to verify the validity of key components in our feature extraction. Then, the patch retrieval experiments based on our customized CCL-based model are compared with those based on other SSL methods.

4.4.1. Results of ablation experiments

Effect of network components. The ablation experiments are conducted to validate the effectiveness of the core components in our proposed CCL-based model. Based on the baseline model (ImageNet pretrained ResNet50) (He et al., 2016), the key innovations added in our CCL architecture include large-scale histopathological images-based pretraining (*MoCo v2*), group-level InfoNCE loss (*MoCo v2+Gro.*), clustering-based memory bank construction (*MoCo v2+Mem.*), and their combination (*MoCo v2+Gro.+Mem.*). It is noted that, except for the baseline model (ImageNet pretrained ResNet50), others are all trained on our large-scale histopathological image database (TCGA and PAIP). The detailed ablation results are summarized in Table 1.

As shown in Table 1, it is seen that the proposed components bring very large improvements (around +24% and +9% in terms of $mMV@5$ on the TissueNet and UniToPatho datasets, respectively) compared with the ImageNet pretrained ResNet50, which verifies the effectiveness of our CCL-based feature learning method. Specifically, adding the group-level InfoNCE loss into the MoCo v2 framework has a consistent improvement of around +2% across the two datasets in terms of the $Acc@1$ and $mMV@5$ metrics. The construction of multiple sub-memory banks also brings high improvements compared to the original MoCo v2

[7] https://github.com/google-research/simclr
[8] https://github.com/facebookresearch/swav
[9] https://github.com/facebookresearch/moco

Table 1
Ablation results on TissueNet and UniToPatho datasets.

	TissueNet				UniToPatho			
	$Acc@1$	$Acc@3$	$Acc@5$	$mMV@5$	$Acc@1$	$Acc@3$	$Acc@5$	$mMV@5$
ImageNet	50.35	77.62	87.68	46.15	58.17	82.89	89.45	59.01
MoCo v2	64.74	86.27	92.77	65.57	63.36	83.38	89.57	64.86
MoCo v2+Gro.	66.20	87.07	93.10	67.56	65.49	83.95	90.04	66.63
MoCo v2+Mem.	66.64	87.21	93.12	68.78	65.87	84.10	90.08	67.19
MoCo v2+Gro.+Mem. (Ours)	**67.09**	**87.81**	**93.40**	**70.01**	**66.55**	**84.32**	**90.31**	**68.35**

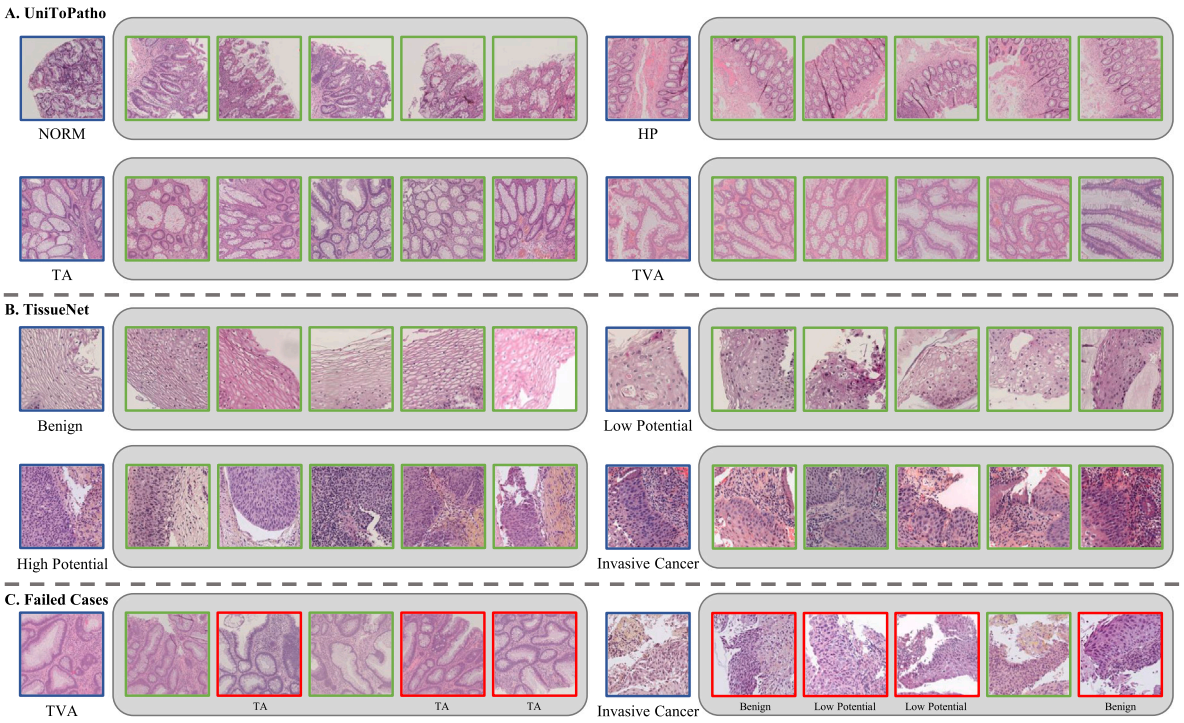

Fig. 2. Visualization of patch-level retrieval results on UniToPatho (A) and TissueNet (B) datasets. Two failed cases from the two datasets are also shown in (C). The blue, green, and red boxes indicate query patches, correct returns, and failed returns, respectively. In the UniToPatho dataset, NORM (normal tissue), HP (hyperplastic polyp), TA (tubular adenoma), and TVA (tubulo-villous adenoma) are respectively shown using one example. In the TissueNet dataset, benign, low malignant potential tissue, high malignant potential tissue, and invasive cancer tissue are respectively shown using one example.

Table 2
Effect of different number of Q values on retrieval accuracy using the TissueNet and UniToPatho datasets.

Q	TissueNet		UniToPatho	
	$Acc@1$	$mMV@5$	$Acc@1$	$mMV@5$
15	66.37	68.78	66.07	67.60
20	66.86	69.25	66.18	67.83
25	**67.09**	**70.01**	**66.55**	**68.35**
30	66.99	69.63	66.28	67.69
35	66.82	69.02	65.98	67.49

Table 3
Effect of different number of S values on retrieval accuracy using the TissueNet and UniToPatho datasets.

S	TissueNet		UniToPatho	
	$Acc@1$	$mMV@5$	$Acc@1$	$mMV@5$
20	66.33	68.29	65.56	67.10
25	66.76	69.12	65.91	67.33
30	**67.09**	**70.01**	**66.55**	**68.35**
40	66.96	69.39	66.01	67.61

Table 4
Effect of different number of MLP heads on retrieval accuracy using the TissueNet and UniToPatho datasets.

	TissueNet		UniToPatho	
	$Acc@1$	$mMV@5$	$Acc@1$	$mMV@5$
One head	65.25	66.94	64.09	65.69
Two heads	**67.09**	**70.01**	**66.55**	**68.35**

Table 5
Effect of different settings of w on retrieval accuracy using the TissueNet and UniToPatho datasets.

w	TissueNet		UniToPatho	
	$Acc@1$	$mMV@5$	$Acc@1$	$mMV@5$
0.1	66.71	69.61	66.13	67.69
0.2	**67.09**	**70.01**	**66.55**	**68.35**
0.5	66.52	68.97	65.79	67.15
1	66.20	67.56	65.49	66.63
Soft	66.23	67.85	65.44	66.81
Hard (w=0.2)	**67.09**	**70.01**	**66.55**	**68.35**

framework (around +3% both on TissueNet and UniToPatho datasets in $mMV@5$). Our final solution makes use of all the above innovations, which provides the best performance as shown in the last row of Table 1.

Although the weighting strategy used in $\mathcal{L}_{\text{W-InfoNCE}}$ is effective, it is no longer required for $\mathcal{L}_{\text{G-InfoNCE}}$. The reason can be explained as follows. $\mathcal{L}_{\text{W-InfoNCE}}$ and $\mathcal{L}_{\text{G-InfoNCE}}$ are used together to achieve more comprehensive contrastive learning, whereas the negative samples of

Table 6
Patch-level retrieval results by comparing our CCL with other SSL-based feature extractors.

	TissueNet				UniToPatho			
	$Acc@1$	$Acc@3$	$Acc@5$	$mMV@5$	$Acc@1$	$Acc@3$	$Acc@5$	$mMV@5$
SimCLR v1 (Chen et al., 2020b)	62.60	85.53	92.85	65.04	61.12	83.25	89.50	62.08
MoCo v2 (Chen et al., 2020a)	64.74	86.27	92.77	65.57	63.36	83.38	89.54	64.86
SwAV (Caron et al., 2020)	65.39	86.06	92.54	66.67	64.18	83.45	89.78	64.98
Ours	**67.09**	**87.81**	**93.40**	**70.01**	**66.55**	**84.32**	**90.31**	**68.35**

$\mathcal{L}_{\text{W-InfoNCE}}$ and $\mathcal{L}_{\text{G-InfoNCE}}$ come from the offline memory bank and the current mini-batch, respectively. $\mathcal{L}_{\text{W-InfoNCE}}$ uses weights less than 1 to reduce the effect of too unbalanced positive and negative numbers. On the contrary, in the computation of $\mathcal{L}_{\text{G-InfoNCE}}$, the ratio of positives to negatives is already greatly increased compared to the traditional contrastive loss. This is achieved by first clustering samples into clusters, and only the cluster centroids are used as positive and negative samples to the anchor sample. Thus, the weighting scheme in $\mathcal{L}_{\text{W-InfoNCE}}$ is no longer needed for $\mathcal{L}_{\text{G-InfoNCE}}$.

Effect of different number of clustering centers (Q and S). To find potential false negative samples, we divide the memory bank into Q sub-memory banks using a K-means algorithm. Similarly, to rebalance the proportion of positive and negative samples, we add an auxiliary branch to divide the samples within a batch into S classes. Two ablation experiments are conducted to investigate the effect of different values of Q and S as shown in Tables 2 and 3. It is seen that although our model is robust to the changes of Q and S, the performance is slightly higher when $Q = 25$ and $S = 30$. Therefore, we empirically set the value of Q to 25 and S to 30.

Effect of different number of MLP heads. To investigate the effect of different configurations of MLP heads, we conduct an ablation study to compare the performance differences between using one shared head and two independent heads (g_1 and g_2). The results are shown in Table 4. It is seen that the approach with two MLP heads offers better performance. The reason for this can be explained as follows. The MLP heads g_1 and g_2 are used to project features into two different spaces to compute $\mathcal{L}_{\text{G-InfoNCE}}$ and $\mathcal{L}_{\text{W-InfoNCE}}$, respectively. $\mathcal{L}_{\text{W-InfoNCE}}$ repels against other instances, while $\mathcal{L}_{\text{G-InfoNCE}}$ attracts instances within each cluster. To alleviate the possible conflict between the two loss functions, two separate branches can be used, which helps train the feature extractor to be more discriminative across different instances whereas keeping similar samples within each cluster together in the feature space (Wang et al., 2021; Zheng et al., 2021).

Effect of different settings of w. In $\mathcal{L}_{\text{W-InfoNCE}}$, we introduce a weight w to reduce the contribution of potential false negative samples. To investigate the sensitivity of our algorithm to different values of w (0.1, 0.2, 0.5, and 1), we perform an ablation study as shown in Table 5. It is seen that a slightly small w leads to better performance, with the best performance obtained when w equals 0.2. The reason is that a slightly smaller value of w suppresses the adverse effects of potentially false negatives more. In addition, we also investigate the effectiveness of a soft setting for w values. That is, for negative samples belonging to the jth cluster, their w values are kept the same as the corresponding cosine similarity score Sim_j. As mentioned earlier, the Sim_j is calculated between the clustering center $\mathbf{c}_j$ and the anchor feature $\mathbf{g}_k$. As can be seen from the last two rows of Table 5, the soft weighting is less effective compared to the hard setting (constant weight of 0.2). This may be due to the fact that the instance discrimination task in contrastive learning needs more explicit pseudo-labels to supervise network training.

4.4.2. Comparison between our CCL and other SSL-based feature extractors

To verify the effectiveness of our CCL-based feature extractor, our method is compared with other state-of-the-art SSL methods, including SimCLR v1 (Chen et al., 2020b), SwAV (Caron et al., 2020), and MoCo v2 (Chen et al., 2020a; He et al., 2020). These SSL methods typically involve two parallel branches during training to conduct contrastive learning. To guarantee a sufficient number of negative samples, SimCLR v1 keeps a large batch size and MoCo v2 uses a separated queue as a memory bank to store negative samples. The SwAV performs a cluster assignment prediction to avoid the direct pairwise feature comparisons and generate more effective negative samples. This comparison experiment is performed based on our patch retrieval method but using different feature extractors. The detailed comparison results are provided in Table 6.

As shown in Table 6, it is demonstrated that the patch retrieval based on our feature extractor consistently shows better performance than those based on other SSL-based feature extractors in terms of all the four metrics on the two patch-level datasets. Specifically, in terms of the $mMV@5$ metric, our method has an improvement of +3.34% and +3.37% over the previous highest method (SwAV). We also present visualized results to provide interpretability analysis as shown in Fig. 2. We show the top-5 retrieved patches across all the subtypes on the UniToPatho (NORM, HP, TA, and TVA) and TissueNet datasets (benign, low potential, high potential, and invasive cancer). It is seen that our system has the ability to find the patches with similar semantic features and return the correct class labels of these patches. For these examples in Fig. 2B, even returned images from the same sub-category show diverse appearances in terms of color and texture, our search system can still return correct results. For the two bad cases shown in Fig. 2C, these erroneously retrieved patches show a similar appearance to the query patches, mainly in terms of morphological features. For example, given a TVA patch as a query, these retrieved error patches are diagnosed as TV. The reason for this is mainly due to the high morphological similarity between the two polyp subtypes. TV and TVA have different villous features but have similar tubular gland structures (Ishii et al., 2011). Thus, distinguishing between the two polyps is a very challenging task even for experienced pathologists (Barbano et al., 2021). Previous studies have also shown the discordance in the diagnosis of colorectal polyp subtypes by pathologists (Turner et al., 2013; Wei et al., 2020).

4.5. Results of WSI retrieval

The WSI-level retrieval experiments are conducted from two aspects: (1) searching for anatomic sites and (2) searching for cancer subtypes based on the same human site. To fairly compare with previous Yottixel (Kalra et al., 2020a,b) and FISH (Chen et al., 2021) approaches, our WSI retrieval experiments are evaluated on the same mechanism as the two methods, including the similar WSI database (TCGA) and the same evaluation metrics. Since the frozen and FFPE WSIs in TCGA show quite different appearances, mixing them will influence retrieval performance (Cooper et al., 2018). Thus, keeping consistent with the two methods (Chen et al., 2021; Kalra et al., 2020a,b), we separate the two kinds of WSIs to respectively perform WSI retrieval tasks.

4.5.1. Results of anatomic site retrieval

Although anatomic sites of tissue sections are usually known in the slide preparation process, anatomic site retrieval experiments can be conducted to validate the searching performance of our algorithm. Our anatomic site retrieval experiment is designed to find WSIs with the same anatomic sites as the query WSI in the TCGA database (frozen or

Table 7
Results of anatomical site retrieval experiment on frozen WSIs in terms of $mMV@10$.

Anatomic Sites	#WSI	#Patient	$mMV@10$	
			Yottixel	Ours
Brain	1818	1091	83.86	**90.21**
Endocrine	796	769	35.37	**80.44**
Gastrointestinal	1984	1234	62.86	**81.77**
Gynecologic	2284	1502	**68.86**	50.74
Hematopoiesis	182	170	45.85	**66.49**
Melanocytic	542	536	37.20	**47.42**
Liver/PB	669	610	35.35	**76.23**
Pulmonary	1658	1093	59.30	**79.99**
Urinary	2035	1323	64.59	**79.35**
Prostate/Testis	759	639	68.07	**84.28**
Breast	1520	1080	66.35	**91.35**
Mesenchymal	263	260	11.19	**74.13**
Head and Neck	727	471	26.24	**79.64**
Macro-average	–	–	51.16	**75.54**
Weighted-average	–	–	60.45	**75.50**

Table 8
Results of anatomical site retrieval experiment on FFPE WSIs in terms of $mMV@10$.

Anatomic Sites	#WSI	#Patient	$mMV@10$		
			Yottixel	FISH	Ours
Brain	1699	878	91.37	**95.80**	93.41
Endocrine	942	737	**73.93**	70.00	69.64
Gastrointestinal	1205	1148	65.12	56.10	**83.80**
Gynecologic	1074	933	63.71	69.40	**76.82**
Hematopoiesis	224	165	52.03	79.40	**80.36**
Melanocytic	554	512	37.20	48.60	**53.97**
Liver/PB	628	586	63.75	72.50	**89.97**
Pulmonary	1137	1028	75.83	71.60	**81.60**
Urinary	1394	1280	66.01	54.20	**69.80**
Prostate/Testis	703	552	80.31	84.40	**86.49**
Breast	1160	1045	70.87	75.80	**93.71**
Mesenchymal	599	254	50.70	61.70	**91.65**
Head and Neck	472	450	49.14	51.40	**77.97**
Macro-average	–	–	64.61	68.53	**80.71**
Weighted-average	–	–	69.05	70.17	**81.62**

Table 9
Results of cancer subtype retrieval experiment on frozen WSIs in terms of $mMV@5$.

WSI Type	#WSI	#Patient	$mMV@5$		WSI Type	#WSI	#Patient	$mMV@5$	
			Yottixel	Ours				Yottixel	Ours
Pulmonary					**Liver/PB**				
LUAD	822	505	68.23	**78.10**	*CHOL*	51	51	35.29	**39.22**
LUSC	751	486	78.25	**90.28**	*LIHC*	398	375	94.36	**94.97**
MESO	87	87	27.71	**83.91**	*PAAD*	218	184	**91.66**	90.83
Urinary					**Gynecologic**				
BLCA	429	410	92.85	**98.37**	*UCEC*	711	542	**90.07**	81.86
KIRC	1088	536	**97.81**	93.75	*CESC*	309	302	64.42	**78.32**
KICH	146	90	78.26	**91.78**	*UCS*	57	57	10.20	**68.42**
KIRP	375	281	62.12	**88.80**	*OV*	1203	589	**99.07**	93.43
Gastrointestinal					**Endocrine**				
COAD	855	459	**63.73**	55.55	*ACC*	91	91	45.67	**81.32**
ESCA	173	172	25.90	**67.05**	*PCPG*	180	175	85.63	**88.89**
READ	330	171	14.32	**37.27**	*THCA*	538	502	97.08	**98.33**
STAD	632	432	71.10	**73.42**					
Melanocytic					**Prostate/Testis**				
UVM	69	69	46.37	**88.41**	*TGCT*	155	149	86.45	**98.06**
SKCM	470	467	**98.70**	94.68	*PRAD*	605	489	**98.33**	98.18
Brain					**Hematopoiesis**				
GBM	1097	577	**94.19**	85.78	*DLBC*	59	46	**91.22**	77.97
LGG	715	509	82.58	**86.30**	*THYM*	124	124	**97.58**	96.77

The macro-average $mMV@5$ of Yottixel and our method are 72.04 and 82.76, respectively.

FFPE). To be consistent with FISH and Yottixel methods (Chen et al., 2021; Kalra et al., 2020a,b), our anatomic site retrieval database covers 13 anatomic sites and includes 27,028 WSIs (11,791 FFPE and 15,237 frozen WSIs). Tables 7 and 8 show anatomical site retrieval results and the number of WSIs and patients in each anatomical site on the frozen and FFPE WSIs, respectively. In addition, to investigate the WSI retrieval performance with different feature extraction methods, we replace the ImageNet features and color histogram features with our CCL-based histopathology features for comparison. The detailed results are shown in Table 1 in the Appendix.

As seen in Tables 7 and 8, the columns Yottixel and FISH are copied directly from their publications. Our method achieves 75.54%

Table 10
Results of cancer subtype retrieval experiment on FFPE WSIs in terms of $mMV@5$.

WSI Type	#WSI	#Patient	$mMV@5$			WSI Type	#WSI	#Patient	$mMV@5$		
			Yottixel	FISH	Ours				Yottixel	FISH	Ours
Pulmonary						**Liver/PB**					
LUAD	538	475	70.96	79.81	**84.01**	*CHOL*	38	38	43.58	46.15	**55.26**
LUSC	512	478	81.70	71.68	**84.18**	*LIHC*	381	365	93.65	90.30	**96.06**
MESO	87	75	8.13	55.81	**72.41**	*PAAD*	203	183	91.04	89.47	**96.55**
Urinary						**Gynecologic**					
BLCA	457	385	95.81	93.22	**98.03**	*UCEC*	595	506	**92.22**	84.28	84.87
KIRC	519	513	91.66	92.29	**93.06**	*CESC*	285	268	62.45	78.78	**86.67**
KICH	109	109	75.92	90.10	**95.41**	*UCS*	87	53	42.22	71.26	**72.41**
KIRP	297	273	67.22	66.33	**90.91**	*OV*	107	106	66.98	**83.18**	70.09
Gastrointestinal						**Endocrine**					
COAD	469	447	**76.14**	48.30	69.72	*ACC*	226	56	93.83	**96.04**	94.69
ESCA	158	156	59.87	79.75	**82.28**	*PCPG*	195	175	88.77	**91.84**	82.99
READ	169	161	10.19	**44.94**	25.44	*THCA*	521	506	97.66	98.07	**99.04**
STAD	409	384	74.23	74.23	**76.53**						
Melanocytic						**Prostate/Testis**					
UVM	80	80	83.75	70.00	**97.50**	*TGCT*	254	149	**99.21**	97.64	97.64
SKCM	474	432	99.57	**99.58**	97.89	*PRAD*	449	403	98.43	98.44	**98.66**
Brain						**Hematopoiesis**					
GBM	857	387	**91.88**	87.75	81.43	*DLBC*	44	44	58.13	**88.37**	72.73
LGG	842	491	89.77	**97.02**	83.73	*THYM*	180	121	98.87	93.89	**98.89**

The macro-average $mMV@5$ of Yottixel, FISH, and our method are 75.99, 81.33, and 84.11, respectively.

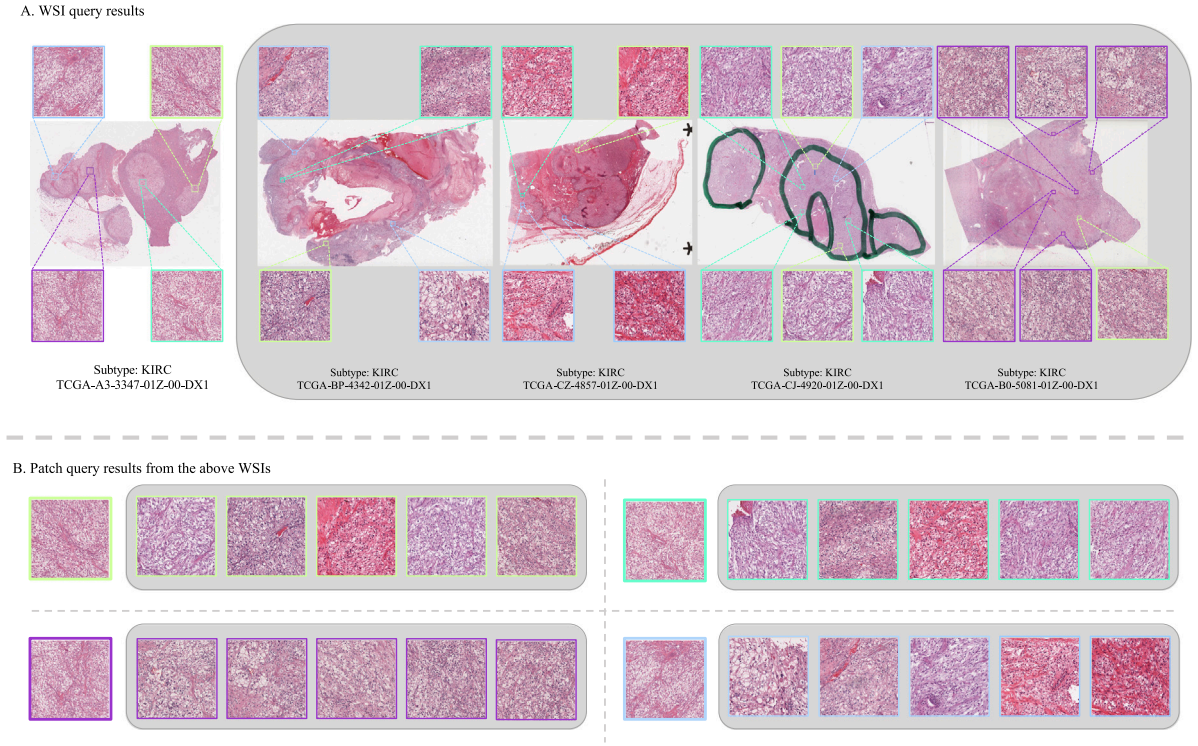

Fig. 3. Interpretability analysis for the WSI-level query results (good case), taking a query slide with the cancer subtype of KIRC as an example. Images with the same color box are considered to have similar semantics. (A) shows an example of WSI query results. The patches with color boxes in the query WSI are generated by the mosaic generation method, which are subsequently used for patch-level retrieval as shown in (B). In (B), each query patch finally returns five similar patches. These returned patches are then projected into their associated WSIs, which are visualized in (A). It is noted that all these highlighted patches are verified as KIRC-positive by two pathologists from Shanghai Jiao Tong University School of Medicine and Sun Yat-sen University Cancer Center.

(frozen) and 80.71% (FFPE) macro-average $mMV@10$, which greatly outperforms the FISH (68.53% on FFPE slides) and Yottixel (51.16% on frozen slides and 64.61% on FFPE slides). In other words, our method improves the overall performance by more than 15% on frozen WSIs and about 10% on FFPE WSIs compared with the existing two WSI retrieval methods. For some specific human sites, such as Gastrointestinal, Breast, Head, and Neck, our method exceeds the existing methods by nearly 20% as shown in Tables 7 and 8. We also notice that the performance across anatomic sites is strongly related to the number of WSIs in the database. The low searching performance usually occurs on sites with a small number of WSIs in the database as expected. For example, as shown in Tables 7 and 8, Melanocytic (frozen and FFPE) shows consistent low performance across all the three methods.

Table 11
Linear evaluation results on DiagSet-A.2 dataset with different sizes of training data. All these methods adopt ResNet50 as the backbone model. The ImageNet means ImageNet pretrained features in a supervised manner. Note that a supervised baseline using 100% of the training data is also implemented, which produces an ACC of 0.8482 and a F1 score of 0.8462.

Methods	Percentage of training data									
	2%		5%		10%		20%		50%	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
ImageNet	0.7686	0.7057	0.7764	0.7512	0.7842	0.7600	0.7885	0.7655	0.7969	0.7749
SimCLR v1	0.7862	0.7370	0.7962	0.7725	0.8176	0.7903	0.8233	0.8041	0.8266	0.8129
MoCo v2	0.7955	0.7419	0.8066	0.7929	0.8248	0.8119	0.8298	0.8124	0.8368	0.8198
SwAV	0.7970	0.7506	0.8211	0.8056	0.8354	0.8211	0.8408	0.8236	0.8471	0.8308
CCL (ours)	**0.8086**	**0.7617**	**0.8381**	**0.8136**	**0.8461**	**0.8301**	**0.8536**	**0.8467**	**0.8563**	**0.8469**

As seen in Table 1 in the Appendix, our CCL-based feature learning method works well with other WSI retrieval methods as well, and produces better accuracy than the original feature extraction methods in these methods. In particular, we rerun the Yottixel and FISH methods but replace the two features (ImageNet features and color histogram features) used in both the Yottixel and FISH methods with our CCL-based features, which produces CCL+Yottixel and CCL+FISH respectively. Compared with the original Yottixel and FISH methods, using the CCL-based features improves the macro-average $mMV@10$ by around +6% and +8%, respectively, further demonstrating the effectiveness of our histopathology-specific feature learning. Also, it is seen that the proposed RetCCL method achieves the highest performance, outperforming the CCL+Yottixel and CCL+FISH methods in terms of macro-average $mMV@10$ by around +10% and +3%, respectively. These results clearly show the advantages of the proposed WSI retrieval algorithm.

4.5.2. Results of cancer subtype retrieval

Recognition for cancer subtypes requires many years of experience and rich expertise, which is very challenging especially for trainees. The cancer subtype retrieval system assumes a query WSI from a known anatomical site and retrieves WSIs with similar cancer subtypes from the dataset collected from the same anatomical site as the query. Among these 13 anatomic sites mentioned above, 10 of them are composed of more than one cancer subtype. Following the FISH and Yottixel methods (Chen et al., 2021; Kalra et al., 2020a,b), we conduct cancer subtype retrieval experiments over these 10 anatomic sites. The detailed cancer subtypes in each anatomic site and their retrieval results are shown in Tables 9 and 10 for frozen and FFPE WSIs, respectively.

As shown in Tables 9 and 10, our method achieves an improvement of +10% than the Yottixel and +3% than the FISH in terms of the macro-average $mMV@5$ in the FFPE WSIs. In the frozen WSI dataset, compared with the Yottixel, our method has achieved a considerable improvement of +40% on some specific subtypes, such as MESO in Pulmonary site, UCS in Gynecologic site, and ESCA in Gastrointestinal site. In the FFPE WSIs, compared with the FISH method, our method has achieved an improvement of around +16% on MESO in Pulmonary site, +25% on KIRP in Urinary site, and +27% on UVM in Melanocytic site, respectively. At some cancer subtypes, such as DLBC and THYM in Hematopoiesis site, our performance is lower than the two methods, which may be influenced by the limited number of WSIs in the corresponding cancer types.

4.5.3. Interpretability analysis

Computer-aided diagnosis needs to provide pathologists with interpretable results to promote its clinical applications. Highlighting evidence that supports the specific results produced by the models is a common strategy. Based on the cancer subtype retrieval experiment, this work provides visualized analysis during WSI retrieval by pointing out high-related patches within each retrieved WSI.

For the good case shown in Fig. 3, our mosaic generation technique can generate a set of key patches (represented using different colors) from the query WSI with KIRC (kidney renal clear cell carcinoma). Then, as shown in Fig. 3B, these patches are adopted as new query images to search for their corresponding similar patches from the database. These found patches along with their meta-information are projected into their associated WSIs as shown in Fig. 3A. Note that all these highlighted patches are also verified as KIRC-positive by two pathologists from Shanghai Jiao Tong University School of Medicine and Sun Yat-sen University Cancer Center. More visualization of the retrieval results is provided in the Appendix. For the bad case as shown in Fig. 1D in the Appendix, the query WSI is diagnosed with COAD (colon adenocarcinoma), which retrieves some failure cases with READ (rectum adenocarcinoma). The reason is largely due to the highly similar glandular in both COAD and READ, that is, the extracted features may not be distinctive enough to produce correct predictions between the two similar subtypes. In summary, this visualization process can help interpret the searching results for pathologists, assisting them to better understand what dominates the returns for one specific query WSI.

4.6. Results of downstream classification task

Our SSL pre-trained feature extractor can be used as a universal feature representation for histopathological image analysis tasks. To further verify the generalizability of the features, this subsection conducts a downstream four-class classification task on the DiagSet-A.2 dataset with different proportions of training data. Also, we compare the performance of our method with other state-of-the-art SSL methods. The detailed results are shown in Table 11.

In Table 11, the results of these state-of-the-art SSL methods are obtained by pretraining on the same data as ours using their released codes. As shown in Table 11, our method achieves the highest performance, which is about 1% higher than the SwAV method under all training data settings. This indicates that our CCL-based feature embedding can be transferred well to the histopathological image classification task. Furthermore, it is seen that our method based on frozen histopathology features uses only 20% of the training data to achieve the performance of the Imagenet-based features using 100% of the training data, which further confirms the effectiveness of our self-supervised feature learning method.

5. Conclusion

This work proposes a histopathological image retrieval algorithm, which is applicable for both WSI-level and patch-level retrieval and can provide visually interpretable results for pathologists. Since a rich and descriptive feature is the key success factor in the image retrieval task, our work pays more attention to the design of the feature extractor. We developed a CCL-based backbone model, which is trained by integrating the multiple sub-memory banks and group-level discrimination together to reduce the number of potential false-negative samples under the assumption of traditional contrastive learning. In the WSI retrieval procedure, we perform a patch-by-patch retrieval rather than a whole WSI level retrieval to obtain interpretable search results. These

retrieved patches are ranked, curated, and aggregated to obtain their associated similar WSIs. As demonstrated in our experimental results, our algorithm has outperformed current WSI retrieval methods by a large margin in both the anatomical site and cancer subtype search settings. Compared with other SSL methods in the patch-level image retrieval experiments, our CCL-based features also demonstrate superior performance. Given the high performance in the histopathological image retrieval task, the proposed feature extractor has the potential to be a universal pretrained model for various histopathological image applications. Future work is warranted to apply our feature representation on more histopathological image analysis tasks to further validate its robustness and generalizability. To assess the clinical benefits of our WSI retrieval algorithm, future work can also validate it on a larger, diverse, and multi-center database.

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

Link to our data is available in the manuscript.

Acknowledgment

This research was in part funded by the National Natural Science Foundation of China (No. 61571314), Science & technology department of Sichuan Province (No. 2020YFG0081), and the Innovative Youth Projects of Ocean Remote Sensing Engineering Technology Research Center of State Oceanic Administration of China (No. 2015001). We also thank Dr. Jietian Jin from the Sun Yat-sen University Cancer Center and Dr. Siteng Chen from the Shanghai Jiao Tong University School of Medicine for their help in validating the results in WSI retrieval.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2022.102645.

References

Akakin, H.C., Gurcan, M.N., 2012. Content-based microscopic image retrieval system for multi-image queries. IEEE Trans. Inf. Technol. Biomed. 16 (4), 758–769.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al., 2021. Big self-supervised models advance medical image classification. In: ICCV. pp. 3478–3488.

Barbano, C.A., Perlo, D., Tartaglione, E., Fiandrotti, A., Bertero, L., Cassoni, P., Grangetto, M., 2021. UniToPatho, a labeled histopathological dataset for colorectal polyps classification and adenoma dysplasia grading. arXiv preprint arXiv:2101.09991.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., 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., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS. Vol. 33. pp. 9912–9924.

Chen, X., Fan, H., Girshick, R.B., He, K., 2020a. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297.

Chen, X., He, K., 2020. Exploring simple siamese representation learning. In: CVPR. pp. 15750–15758.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020b. A simple framework for contrastive learning of visual representations. In: ICML. Vol. 1. pp. 1597–1607.

Chen, C., Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Schaumberg, A.J., Mahmood, F., 2021. Fast and scalable image search for histology. arXiv preprint arXiv:2107.13587.

Cooper, L.A.D., Demicco, E.G., Saltz, J.H., Powell, R.T., Rao, A., Lazar, A.J., 2018. PanCancer insights from the cancer genome atlas: the pathologist's perspective. J. Pathol. 244 (5), 512–524.

Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A., Courtiol, P., 2020. Self-supervision closes the gap between weak and strong supervision in histology. arXiv preprint arXiv:2012.03583.

Doersch, C., Gupta, A., Efros, A.A., 2015. Unsupervised visual representation learning by context prediction. In: ICCV. pp. 1422–1430.

Evans, A.J., Bauer, T.W., Bui, M.M., Cornish, T.C., Duncan, H., Glassy, E.F., Hipp, J., McGee, R.S., Murphy, D., Myers, C., et al., 2018. US food and drug administration approval of whole slide imaging for primary diagnosis: a key milestone is reached and new questions are raised. Arch. Pathol. Lab. Med. 142 (11), 1383–1387.

Foster, A., Pukdee, R., Rainforth, T., 2021. Improving transformation invariance in contrastive representation learning. In: ICLR. pp. 1–7.

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.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020. Bootstrap your own latent: A new approach to self-supervised learning. In: NeurIPS. Vol. 33. pp. 21271–21284.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: CVPR. pp. 770–778.

Hegde, N., Hipp, J.D., Liu, Y., Emmert-Buck, M., Reif, E., Smilkov, D., Terry, M., Cai, C.J., Amin, M.B., Mermel, C.H., Nelson, P.Q., Peng, L.H., Corrado, G.S., Stumpe, M.C., 2019. Similar image search for histopathology: SMILY. Npj Digit. Med. 2 (1), 1–9.

Ishii, T., Notohara, K., Umapathy, A., Mallitt, K.-A., Chikuba, H., Moritani, Y., Tanaka, N., Rosty, C., Matsubara, N., Jass, J., et al., 2011. Tubular adenomas with minor villous changes show molecular features characteristic of tubulovillous adenomas. Am. J. Surg. Pathol. 35 (2), 212–220.

Kalra, S., Tizhoosh, H.R., Choi, C., Shah, S., Diamandis, P., Campbell, C.J.V., Pantanowitz, L., 2020a. Yottixel - An image search engine for large archives of histopathology whole slide images. Med. Image Anal. 65, 1–12.

Kalra, S., Tizhoosh, H.R., Shah, S., Choi, C., Damaskinos, S., Safarpoor, A., Shafiei, S., Babaie, M., Diamandis, P., Campbell, C.J.V., Pantanowitz, L., 2020b. Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. Npj Digit. Med. 3 (1), 1–15.

Kim, Y.J., Jang, H., Lee, K., Park, S., Min, S.-G., Hong, C., Park, J.H., Lee, K., Kim, J., Hong, W., et al., 2021. PAIP 2019: Liver cancer segmentation challenge. Med. Image Anal. 67, 1–11.

Koohbanani, N.A., Unnikrishnan, B., Khurram, S.A., Krishnaswamy, P., Rajpoot, N.M., 2021. Self-path: Self-supervision for classification of pathology images with limited annotations. IEEE Trans. Med. Imaging 40 (99), 2845–2856.

Koziarski, M., Cyganek, B., Olborski, B., Antosz, Z., Żydak, M., Kwolek, B., Wasowicz, P., Bukała, A., Swadźba, J., Sitkowski, P., 2021. DiagSet: a dataset for prostate cancer histopathological image classification. arXiv preprint arXiv:2105.04014.

Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A., 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans. Med. Imaging 36 (7), 1550–1560.

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: CVPR. pp. 14318–14328.

Li, Z., Zhang, X., Müller, H., Zhang, S., 2018. Large-scale retrieval for medical image analytics: A comprehensive review. Med. Image Anal. 43, 66–84.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030.

Loshchilov, I., Hutter, F., 2016. SGDR: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983.

Lu, M.Y., Chen, R.J., Wang, J., Dillon, D., Mahmood, F., 2019. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. arXiv preprint arXiv:1910.10825.

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.

Ma, Y., Jiang, Z., Zhang, H., Xie, F., Zheng, Y., Shi, H., Zhao, Y., 2016. Breast histopathological image retrieval based on latent dirichlet allocation. IEEE J. Biomed. Health Inf. 21 (4), 1114–1123.

Ma, Y., Jiang, Z., Zhang, H., Xie, F., Zheng, Y., Shi, H., Zhao, Y., Shi, J., 2018. Generating region proposals for histopathological whole slide image retrieval. Comput. Methods Programs Biomed. 159, 1–10.

Misra, I., Maaten, L.v.d., 2020. Self-supervised learning of pretext-invariant representations. In: CVPR. pp. 6707–6717.

Mukhopadhyay, S., Feldman, M.D., Abels, E., Ashfaq, R., Beltaifa, S., Cacciabeve, N.G., Cathro, H.P., Cheng, L., Cooper, K., Dickey, G.E., et al., 2018. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study). Am. J. Surg. Pathol. 42, 39–52.

Noroozi, M., Favaro, P., 2016. Unsupervised learning of visual representations by solving Jigsaw puzzles. In: ECCV. pp. 69–84.

Shao, W., Wang, T., Sun, L., Dong, T., Han, Z., Huang, Z., Zhang, J., Zhang, D., Huang, K., 2020. Multi-task multi-modal learning for joint diagnosis and prognosis of human cancers. Med. Image Anal. 65, 1–10.

Shi, X., Xing, F., Xu, K., Xie, Y., Su, H., Yang, L., 2017. Supervised graph hashing for histopathology image retrieval and classification. Med. Image Anal. 42, 117–128.

Srinidhi, C.L., Kim, S.W., Chen, F.-D., Martel, A.L., 2021. Self-supervised driven consistency training for annotation efficient histopathology image analysis. arXiv preprint arXiv:2102.03897.

Tizhoosh, H.R., Diamandis, P., Campbell, C.J., Safarpour, A., Kalra, S., Maleki, D., Riasatian, A., Babaie, M., 2021. Searching images for consensus: Can AI remove observer variability in pathology? Am. J. Pathol. 191, 1702–1708.

Tsai, Y.-H.H., Ma, M.Q., Yang, M., Zhao, H., Morency, L.-P., Salakhutdinov, R., 2021. Self-supervised representation learning with relative predictive coding. In: ICLR. pp. 1–7.

Turner, J.K., Williams, G.T., Morgan, M., Wright, M., Dolwani, S., 2013. Interobserver agreement in the reporting of colorectal polyp pathology among bowel cancer screening pathologists in Wales. Histopathology 62 (6), 916–924.

Van Den Oord, A., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Van Den Oord, A., Vinyals, O., Kavukcuogluothers, K., 2017. Neural discrete representation learning. In: NeurIPS. Vol. 30. pp. 6306–6315.

Wang, X., Liu, Z., Yu, S.X., 2021. Unsupervised feature learning by cross-level instance-group discrimination. In: CVPR. pp. 12586–12595.

Wei, J.W., Suriawinata, A.A., Vaickus, L.J., Ren, B., Liu, X., Lisovsky, M., Tomita, N., Abdollahi, B., Kim, A.S., Snover, D.C., et al., 2020. Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. JAMA Netw. Open 3 (4), e203398.

Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y., Hu, H., 2021. Self-supervised learning with swin transformers. arXiv preprint arXiv:2105.04553.

Xue, Y., Ye, J., Zhou, Q., Long, L.R., Antani, S., Xue, Z., Cornwell, C., Zaino, R., Cheng, K.C., Huang, X., 2021. Selective synthetic augmentation with HistoGAN for improved histopathology image classification. Med. Image Anal. 67, 101816.

Yan, K., Cai, J., Jin, D., Miao, S., Guo, D., Harrison, A.P., Tang, Y., Xiao, J., Lu, J., Lu, L., 2022. SAM: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. IEEE Trans. Med. Imaging http://dx.doi.org/10.1109/TMI.2022.3169003.

Zhang, R., Isola, P., Efros, A.A., 2016. Colorful image colorization. In: ECCV. Springer, pp. 649–666.

Zhang, X., Liu, W., Dundar, M., Badve, S., Zhang, S., 2014. Towards large-scale histopathological image analysis: Hashing-based image retrieval. IEEE Trans. Med. Imaging 34 (2), 496–506.

Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Ma, Y., Shi, H., Zhao, Y., 2017. Size-scalable content-based histopathological image retrieval from database that consists of WSIs. IEEE J. Biomed. Health Inf. 22 (4), 1278–1287.

Zheng, M., Wang, F., You, S., Qian, C., Zhang, C., Wang, X., Xu, C., 2021. Weakly supervised contrastive learning. In: ICCV. pp. 10042–10051.