Medical Image Analysis 81 (2022) 102559

Contents lists available at ScienceDirect

Medical Image Analysis

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

Transformer-based unsupervised contrastive learning for histopathological image classification

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

[a] College of Biomedical Engineering, Sichuan University, Chengdu 610065, China
[b] College of Computer Science, Sichuan University, Chengdu 610065, China
[c] Tencent AI Lab, Shenzhen 518057, China

ARTICLE INFO

Keywords:
Histopathology
Transformer
Self-supervised learning
Feature extraction

ABSTRACT

A large-scale and well-annotated dataset is a key factor for the success of deep learning in medical image analysis. However, assembling such large annotations is very challenging, especially for histopathological images with unique characteristics (e.g., gigapixel image size, multiple cancer types, and wide staining variations). To alleviate this issue, self-supervised learning (SSL) could be a promising solution that relies only on unlabeled data to generate informative representations and generalizes well to various downstream tasks even with limited annotations. In this work, we propose a novel SSL strategy called semantically-relevant contrastive learning (SRCL), which compares relevance between instances to mine more positive pairs. Compared to the two views from an instance in traditional contrastive learning, our SRCL aligns multiple positive instances with similar visual concepts, which increases the diversity of positives and then results in more informative representations. We employ a hybrid model (*CTransPath*) as the backbone, which is designed by integrating a convolutional neural network (CNN) and a multi-scale Swin Transformer architecture. The *CTransPath* is pretrained on massively unlabeled histopathological images that could serve as a collaborative local–global feature extractor to learn universal feature representations more suitable for tasks in the histopathology image domain. The effectiveness of our SRCL-pretrained *CTransPath* is investigated on five types of downstream tasks (patch retrieval, patch classification, weakly-supervised whole-slide image classification, mitosis detection, and colorectal adenocarcinoma gland segmentation), covering nine public datasets. The results show that our SRCL-based visual representations not only achieve state-of-the-art performance in each dataset, but are also more robust and transferable than other SSL methods and ImageNet pretraining (both supervised and self-supervised methods). Our code and pretrained model are available at https://github.com/Xiyue-Wang/TransPath.

1. Introduction

Benefiting from a massive amount of labeled data, deep learning has achieved remarkable success in the field of medical image analysis, even outperforming humans (Yu et al., 2018; Liu et al., 2020; Zhang et al., 2019). However, manual annotation is an expensive and time-consuming task, which leads to limited elaborate annotations available in the medical image community. For histopathological whole-slide images (WSIs), such curated annotations are even more scarce due to their unique challenges (e.g., gigapixel image size, enormous heterogeneity, multiple cancer types, and wide staining variations). WSIs could cover complex biologically relevant structures ranging from cellular-level (e.g., subcellular vesicle and nuclear granule) to tissue-level (e.g., endothelia, epithelia, muscle, vessel, and gland) (Rashid et al., 2022; Javed et al., 2020). The gigapixel image size of WSIs creates an extremely large search space for labeling and the heterogeneous tissue distribution within WSI makes it difficult to localize target lesion regions that usually constitute a small portion of the entire WSI. Moreover, the multiple cancer types lead to variant tissue styles that further increase the annotation challenge, and the wide staining variations further increase color divergence. Thus, there is an urgent requirement to develop an effective feature extractor from unlabeled histopathological images to alleviate the burden of heavy annotation, which has the potential to promote the development of digital pathology and aid pathologists for fast and precise diagnoses.

To reduce the annotation dependency of histopathological images, transfer learning from large-scale labeled natural images (e.g.,

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

https://doi.org/10.1016/j.media.2022.102559
Received 21 February 2022; Received in revised form 24 June 2022; Accepted 25 July 2022
Available online 30 July 2022

ImageNet (Russakovsky et al., 2015)) may be an alternative approach, which has been proven to be an effective training strategy that can improve classification, regression and segmentation performance with limited annotations (Mormont et al., 2020; Talo, 2019; Srinidhi et al., 2021; Lu et al., 2021). However, domain differences between natural images and histopathological images are tremendous, ranging from low-level texture features and high-level semantic features. For example, objects and faces are very different from cells and tissues, resulting in limited performance gains. A preferable approach to tackle this domain discrepancy is to pretrain or train from scratch on domain-relevant data, which is limited by the annotation-lacking problem mentioned earlier. To address this, self-supervised pretraining without the requirement of manual labels is a possible option, which learns the visual representation based on supervised signals generated by the data itself.

The tremendous successes of self-supervised learning (SSL) techniques in the computer vision community have promoted the development of SSL in histopathological image analysis. There have been some published works that apply SSL techniques to boost the performance of classification, regression, and segmentation of histopathological images (Koohbanani et al., 2021; Srinidhi et al., 2022; Sahasrabudhe et al., 2020; Yang et al., 2021; Patil et al., 2021; Li et al., 2021a; Xie et al., 2020; Li et al., 2021b; Ciga et al., 2022; Huang et al., 2021; Abbet et al., 2020; Li et al., 2021c). These approaches process histopathological images by simply applying existing contrastive learning (CL)-based SSL frameworks (e.g., SimCLR and MoCo) or tailoring some histopathology-oriented SSL tasks on a convolutional neural network (CNN)-specific backbone. These studies confirm the importance of SSL in the field of histopathology. However, there are still three aspects that could be further improved. First, the contrastive pairs defined in CL are extremely biased for histopathological images. CL assigns two augmented views from the same instance as one positive pair, which limits the variability and diversity of positive samples. When applied to histopathological images, a large number of semantically correlated pairs will be misidentified as negative samples, such as patches with similar cells/tissues within/across WSIs. Thus, a histopathology-oriented CL approach should be considered to improve the quality of positive views. Second, only CNN structures are applied. CNN has a good capacity to learn low-level texture content features (local features), which is a crucial determinant in classification tasks. The learning of global context features is often limited by the receptive field of CNN. The cropped histopathological image patches are usually large enough to capture both cell-level structures (e.g., cellular microenvironment) and tissue-level contexts (e.g., tumor microenvironment). Thus, both local and global features are beneficial for pathological image analysis and should be extracted. Third, the data currently used for SSL training are relatively homogeneous and their number is rather limited. Even though Ciga et al. (2022) claimed that 58 datasets had been utilized in their SSL pretraining process, the final amount of patches was around 400 thousand. The small amount of training data makes it difficult to cover the diversity of histopathological images, especially for pre-training using unlabeled data. In summary, there is a lack of a universal SSL algorithm for feature extraction based on large-scale and diverse datasets in the histopathology field.

To address the above-mentioned limitations, we develop a new SSL method to better capture histopathology-oriented features by constructing a semantically-relevant contrastive learning (SRCL) framework and a hybrid CNN-transformer backbone. Motivated by the presence of a large amount of similar cell or tissue patches, our proposed SRCL framework aims to find more semantically matched positives for each instance in the latent space. In traditional CL, the visual diversity of positives depends heavily on the design of data augmentation algorithms. Our SRCL improves this by selecting more similar positives from different instances, which introduces more visual diversity than traditional positive samples, resulting in more informative semantic representations. Our hybrid backbone (called *CTransPath*) captures both local fine structure and global context for histopathological image analysis, which also guarantees more stability in the Transformer training process. CNN extracts local features by convolutional computation and Transformer captures global dependencies through the interaction among CNN-generated tokens. The combination of CNN and Transformer networks further facilitates the construction of our powerful and universal feature extractor. In our self-supervised pretraining procedure, the used database is the largest publicly available in the histopathology scenario, comprising the cancer genome atlas (TCGA[1]) and pathology AI platform (PAIP[2]) datasets and including around 15 million patches cropped from over 30 thousand WSIs (approximately 87T). Both TCGA and PAIP cover multiple organs and cancer types (over 25 anatomic sites and 32 cancer subtypes in total), which ensures sample diversity and helps train a universal feature extractor. In addition, to validate the effectiveness of our SSL algorithm, we fully evaluate our pretrained model on five different downstream tasks, including patch retrieval, supervised patch classification, weakly-supervised WSI classification, mitosis detection, and colorectal adenocarcinoma gland segmentation.

Our main contributions are summarized below:

- To the best of our knowledge, this is the first Transformer-based unsupervised feature extractor carried out on the largest public histopathological image datasets.
- We propose an SRCL approach, which introduces instance variations by selecting more correct and diverse positive samples, helping more informative feature representations.
- We construct a hybrid architecture (*CTransPath*) for histopathological image classification. It replaces the patch partition of Swin Transformer with a simple CNN, which enables more stable network training and also helps build a powerful feature extractor with fine local structure and global context.
- Benefiting from the above design, our model shows state-of-the-art performance in five different downstream tasks (covering nine public histopathological datasets), which also shows a more robust and transferable performance than other SSL methods and ImageNet pretraining (either supervised or self-supervised). The proposed *CTransPath* could serve as a general-purpose feature extractor for various histopathological applications. Our code and pretrained model have been released online to facilitate reproductive research.

This work is an extended version of our previous conference paper (Wang et al., 2021b). We have made three major modifications to further improve the universality and robustness of our proposed SSL-based feature extractor for histopathological image analysis. First, based on the unique characteristics of histopathological images, we improve the traditional CL paradigm by considering more diverse positive pairs, including augmented views from the current instance and semantically relevant instances selected from an independent memory bank. Second, we change our previous backbone using a more powerful Transformer framework (Swin Transformer), which enhances multi-scale feature learning while maintaining a small number of parameters. Third, we use all image patches instead of 100 randomly selected patches from each WSI in our previous version. More pretraining data could provide more sample diversity and help train a robust feature extractor. In addition, to validate the effectiveness of our proposed histopathology-oriented feature extractor, we reconstruct five types of downstream experiments (across nine datasets), including patch retrieval, patch classification, weakly-supervised WSI classification, mitosis detection, and colorectal adenocarcinoma gland segmentation. These experiments contain a thorough ablation study, comparisons to state-of-the-art SSL methods, comparisons to other best-performing methods evaluated on these test data, and interpretability analysis to visualize the learned feature representations.

[1] https://portal.gdc.cancer.gov/
[2] http://www.wisepaip.org/paip/

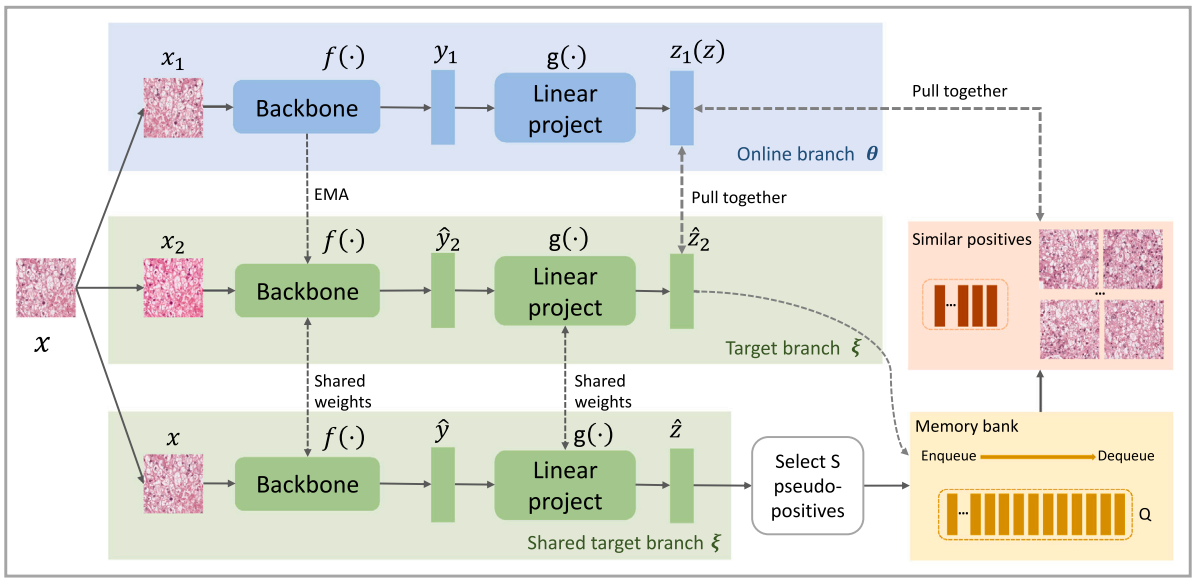

Fig. 1. An overview of our proposed SRCL approach for histopathological image applications. It is an improved framework based on MoCo v3 (Chen et al., 2021). The negative samples are stored in each mini-batch and the positives are from two paths: (i) two data augmentations of the current input image and (ii) top S semantically-relevant images identified by comparing the current input feature with samples in the memory bank. Based on the above design, a semantically-relevant contrastive loss is proposed to guide the network training.

2. Related works

This section reviews the literature about self-supervised learning (SSL) in the computer vision and histopathological image fields, respectively.

2.1. Self-supervised learning

SSL can be regarded as a form of unsupervised learning due to the absence of manual annotation, which aims to construct a rich visual representation using the supervision formulated by the data itself. The learned representation could be further used to improve performance in various downstream tasks. SSL approaches have presented remarkable success in the field of computer vision, which can be divided into two categories: handcrafted-pretext-based and CL-based schemes.

The pretext tasks (auxiliary prediction tasks) are designed to make full use of the information contained in the image pixels. By solving the pretext task, the network can extract general visual representations. These pretext tasks can be summarized in three categories: (i) global image prediction for rotation angles (Gidaris et al., 2018) or image coloring (Zhang et al., 2016), (ii) small patch prediction, such as jigsaw puzzle solving (Noroozi and Favaro, 2016), and (iii) image context prediction, such as predicting the relative position of sub-regions within an image (Doersch et al., 2015). However, these pretext tasks encourage models to learn covariant feature representations rather than invariant ones (Misra and Maaten, 2020), which leads to limited generalization ability.

More recently, CL-based SSL has emerged as a promising alternative method, which has shown excellent performance (even better than the supervised method) on natural image scenes (Chen et al., 2020; He et al., 2020). CL extracts augmentation-invariant and instance-discriminating features by pulling similar (positive) samples and repelling dissimilar (negative) ones. The positive pair is defined as two random data augmentations from the same image, while negative pairs are data augmentations from different images in the current batch or memory bank. It is clear that the definitions of positive and negative samples are potentially wrong since instances in different pairs may contain highly relevant semantics. To alleviate this problem, some recent studies have dedicated a better selection of positive and negative samples. For example, to better select positives, one nearest neighbor in the space or feature level is alternatively adopted as a positive sample (Yèche et al., 2021; Pantazis et al., 2021; Dwibedi et al., 2021). To better select negatives, SwAV used online clustering to divide the feature space into several distinctive prototypes (Caron et al., 2020).

2.2. Self-supervised learning in digital pathology

With the rapid development of the SSL technique, it has some applications in digital pathology image analysis, which can also be categorized into pretext-based, CL-based methods, and their combinations. These pretext tasks are designed according to the characteristics of histopathological images, including magnification prediction (Koohbanani et al., 2021; Srinidhi et al., 2022; Sahasrabudhe et al., 2020), hematoxylin channel prediction (Srinidhi et al., 2022), cross-stain prediction (Yang et al., 2021), color reconstruction (Patil et al., 2021; Li et al., 2021a), and neighborhood image related transformations, such as scale-wise triplet learning and count ranking (Xie et al., 2020). Although these pretext tasks take into account the unique characteristics of histopathological images, the pretrained model will focus on features involved in a specific task. As a result, these pretext-based approaches have difficulty in obtaining universal features in the histopathological images, reducing their generalization power. These CL-based methods directly apply current frameworks (e.g., SimCLR and MoCo) to the histopathological images without considering histopathology characteristics (Li et al., 2021b; Ciga et al., 2022; Huang et al., 2021). To involve histopathology-specific knowledge, some hybrid methods are proposed to combine the advantage of instance discrimination in CL and histopathology-oriented pretext tasks. Yang et al. (2021) utilized a two-stage SSL training method that incorporates a cross-stain prediction and a CL pretraining. However, it only considers color variances in histopathological images and its two-stage SSL training requires more computational resources. To mine more accurate positive samples, Abbet et al. (2020) and Li et al. (2021c) further considered spatial and semantic proximity. There may be conflicting positives or negatives to confuse the networking training. For instance, spatially neighbored samples may not be adjacent in the feature space, such as the boundary between normal and cancerous tissues. Moreover, these mentioned studies also lack extensive evaluation on large and diverse datasets.

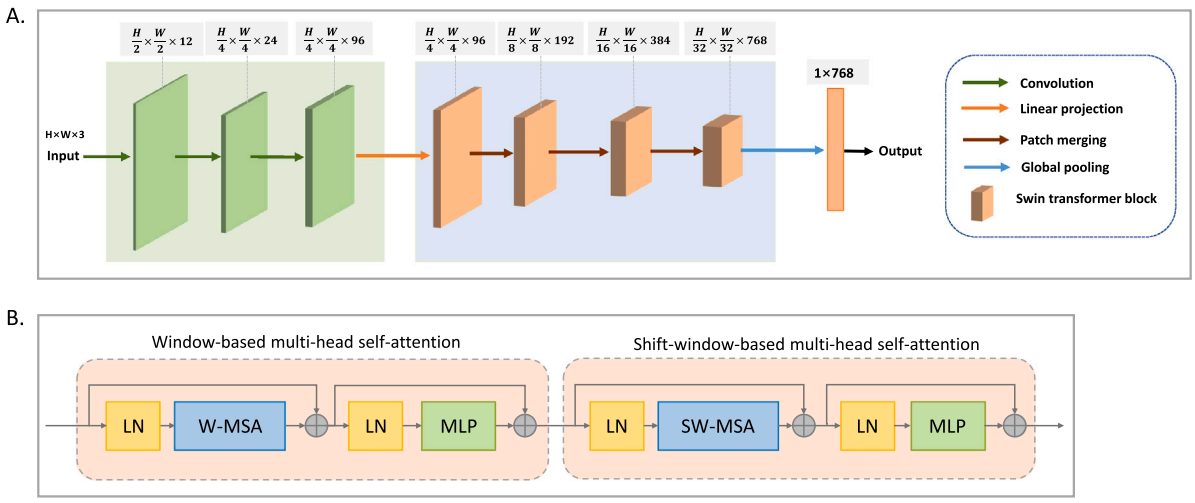

Fig. 2. The structure of our hybrid CNN-transformer backbone (*CTransPath*). (A). The backbone network employs the Swin Transformer framework (Liu et al., 2021), where the patch partition is replaced by a CNN structure. The CNN part is designed with three sequential convolutional layers of kernel sizes 3 × 3, 3 × 3, and 1 × 1. Similar to the popular ResNet structure, Swin Transformer is designed to generate hierarchical feature representation using four sequential stages. At each stage, several repeating Swin Transformer blocks are connected. (B). Illustration of a Swin transformer block, which contains a window-based multi-head self-attention (W-MSA) layer and a shift-window-based multi-head self-attention (SW-MSA) layer.

3. Methods

This section presents an overview of our proposed SSL algorithm based on a semantically-relevant contrastive learning (SRCL) and a hybrid backbone (*CTransPath*), which is shown in Fig. 1. As shown in Fig. 1, the current input patch and its two data augmentations can be regarded as three different views of an input, which are first encoded into corresponding feature vectors using three network branches. Unlike conventional CL (e.g., MoCo v3 (Chen et al., 2021)) which has a pair of positives from the same instance, our positive samples cover an augmented view of the current input instance and additional pseudo-positives selected from a very large memory bank, which guarantees the diversity of positives and thus the more discriminative feature representations. In the pseudo-positive mining process, given a query vector from the current input, S semantically relevant patches are retrieved from the memory bank based on the cosine similarity metric, which are then adopted as additional pseudo-positives for the SRCL calculation. It is noted that the memory bank is independent from the current mini-batch since the memory bank only contains samples from previous mini-batches. In the proposed framework, the hybrid backbone adopts the Swin Transformer for its multi-scale feature extraction capacity, whereas the patch partitioning part is replaced with a CNN-based nonlinear mapper to improve the stability of network training and facilitate better local feature extraction. The integration of CNN and Swin Transformer enables better local and global feature extraction.

3.1. Problem formulation

Let $D^u = \{\mathbf{x}_i^u\}_{i=1}^N$ denote the unlabeled dataset used for SSL pretraining, where $\mathbf{x}_i \in \mathbb{R}^{h\times w\times 3}$ is a small patch cropped from WSI and N represents the total number of images (patches). The purpose of SSL is to generate pseudo labels based on the data itself to drive network training. The CL-based SSL has exhibited competitive performance, which is thus adopted as our main framework. The CL-based SSL method performs two data augmentations on two network branches for each sample, generating $D^q = \{\mathbf{x}_i^q\}_{i=1}^N$ and $D^k = \{\mathbf{x}_j^k\}_{j=1}^N$, respectively. Then, the two data augmentations from the same input are regarded as positive pairs $D^{pos} = \{\mathbf{x}_i^q, \mathbf{x}_j^k\}_{\|i=j\|}$ while data augmentations from different images are used to form negative pairs $D^{neg} = \{\mathbf{x}_i^q, \mathbf{x}_j^k\}_{\|i\neq j\|}$. Two shared backbone neural networks ($f(\cdot)$ and $f'(\cdot)$) on two separate branches are employed to extract feature representations from the augmented samples. The contrastive loss is designed to pull together positive representations and push away negative ones.

3.2. Semantically-relevant contrastive learning

Self-supervised pretraining aims to learn a transferable representation of raw data without requiring manual supervision. Traditional CL-based SSL approaches construct supervised signals by regarding two augmented views from the same image as a positive pair and those from different images as negative pairs (Chen et al., 2020; He et al., 2020). For histopathological images, there are a large number of similar patches (i.e., patches with similar cellular and tissue compositions) both within and across WSIs, which are defined as semantically relevant samples. Thus, the positive pairs should be counted more instead of fixed one pair in the traditional CL setting. Motivated by this observation, we aim to modify the traditional CL strategy by selecting more semantically relevant positive pairs using cosine similarity metric. These positive pairs no longer come from the same instance, which greatly increases the diversity of positive samples.

As illustrated in Fig. 1, there are three parallel paths: online, target, and shared target branches for encoding three different views of the input. These branches all use the proposed *CTransPath* architecture as the backbone model. Similar to MoCo (He et al., 2020), there is a memory bank that is constructed by enqueueing the features from the target branch during training, which is updated at the end of each iteration. We train the online branch with parameter θ and update the target branch with parameter ξ by $\xi \leftarrow m\xi + (1-m)\theta$. The target and the shared target branches share the same network structure and parameters ξ, which are represented using the same color as shown in Fig. 1.

Three different views of an input sample are passed into the three branches, respectively, including two augmented versions of the input and the original non-altered input itself. As shown in Fig. 1, similar to MoCo v3 (Chen et al., 2021), the feature vector obtained in the online branch serves as an anchor, which is used to pull positives closer and push negatives away during the CL computation. The feature vector computed in the target branch is used to refresh the memory bank as training proceeds. The feature vector generated in the shared target branch acts as a query to retrieve semantically-similar samples from the memory bank. Note that the roles of the online and target branches can be interchangeable. Using the original view as a query to select more positive samples helps generate cross-view variations and guarantees

more reliability and stability compared with only considering two augmented views in the online and target branches (Caron et al., 2020; Wang et al., 2021a).

Given a random histopathological image patch $\mathbf{x}$, it generates two augmentations ($\mathbf{x}_1$ and $\mathbf{x}_2$). When $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively pass through our *CTransPath* (formulated as $f(\cdot)$) in the online and target networks, corresponding feature representations as generated $\mathbf{y}_1 = f^{\theta}(\mathbf{x}_1)$, $\hat{\mathbf{y}}_2 = f^{\xi}(\mathbf{x}_2)$. Then, a linear projection head $g(\cdot)$ is adopted to transform these representations into another latent space, i.e., $\mathbf{z}_1 = g^{\theta}(\mathbf{y}_1)$ in online network, and $\hat{\mathbf{z}}_2 = g^{\xi}(\hat{\mathbf{y}}_2)$ in target network. Symmetrically, the swapped prediction separately feeds $\mathbf{x}_1$ and $\mathbf{x}_2$ into the target and online networks, obtaining $\mathbf{y}_2 = f^{\theta}(\mathbf{x}_2)$, $\hat{\mathbf{y}}_1 = f^{\xi}(\mathbf{x}_1)$, $\mathbf{z}_2 = g^{\theta}(\mathbf{y}_2)$, $\hat{\mathbf{z}}_1 = g^{\xi}(\hat{\mathbf{y}}_1)$.

For contrastive learning, we choose the feature vector in the online branch as an anchor $\mathbf{z}$ as shown in Fig. 1, which is used to construct positive and negative pairs. In conventional CL method, $\mathbf{z}$ has one positive sample $\hat{\mathbf{z}}_2$. To obtain more positive samples, we aim to find samples that are visually similar to $\mathbf{z}$. For this purpose, cosine similarities D between $\mathbf{z}$ and each feature vector $\mathbf{c}_i$ ($i = 1, \ldots, Q$) stored in the memory bank with a length of Q are calculated

$$D\left(\mathbf{z}, \mathbf{c}_i\right) = \frac{\mathbf{z} \cdot \mathbf{c}_i}{\|\mathbf{z}\| \|\mathbf{c}_i\|}, i = 1, \ldots, Q. \tag{1}$$

Then, the obtained $D\left(\mathbf{z}, \mathbf{c}_i\right)$ is sorted in descending order. The top S samples with the highest cosine similarity are taken as the new positives for anchor $\mathbf{z}$. Then, combining the original positive sample $\hat{\mathbf{z}}_2$ in conventional CL, the total number of positive pairs for the anchor $\mathbf{z}$ becomes S+1. Our final semantically-relevant contrastive loss L_{SRCL} is optimized by maximizing the similarity between all S+1 positive samples:

$$\mathcal{L}_2(\mathbf{z}, \mathbf{z}^+, \mathbf{z}^-) = -\log \frac{\sum_{i=1}^{S+1} \exp\left(\mathbf{z}_i^+ \cdot \mathbf{z}/\tau\right)}{\sum_{i=1}^{S+1} \exp\left(\mathbf{z}_i^+ \cdot \mathbf{z}/\tau\right) + \sum_{j=1}^{N} \exp\left(\mathbf{z}_j^- \cdot \mathbf{z}/\tau\right)} \tag{2}$$

$$\mathcal{L}_{SRCL} = \frac{1}{2}\mathcal{L}_2(\mathbf{z}_1, \hat{\mathbf{z}}_2, \mathbf{z}^-) + \frac{1}{2}\mathcal{L}_2(\hat{\mathbf{z}}_2, \mathbf{z}_1, \mathbf{z}^-) \tag{3}$$

where $\mathbf{z}$ represents an anchor sample (e.g., $\mathbf{z}_1$ in the online branch as shown in Fig. 1). $\mathbf{z}^+$ and $\mathbf{z}^-$ denote the positive and negative features of the anchor feature. S+1 and N represent the number of positive and negative pairs, respectively.

3.3. Backbone construction

The proposed hybrid network backbone *CTransPath* fully utilizes the local feature mining ability of CNN and the global interaction ability of Transformer, which is shown in Fig. 2. Previous studies (Xiao et al., 2021; Chen et al., 2021) have indicated that the Transformer architecture is much harder to optimize compared with CNNs, mainly due to the patch projection implemented through large-kernel large-stride convolution operations. To alleviate this problem and as motivated by Xiao et al. (Xiao et al., 2021) and Liu et al. (Liu et al., 2021), we adopt the Swin Transformer as the backbone model to take advantage of its ability of multi-scale feature extraction and computation efficiency, but replace the patch partition part with a CNN module to help mining local features and ensure more stable training. The CNN module is designed with three consecutive convolutional layers with kernel sizes of 3 × 3, 3 × 3, and 1 × 1. In our backbone model, an input image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$ is first passed through the CNN to generate a local feature map F with the size of $\frac{H}{4} \times \frac{W}{4} \times C$, which is then taken as the input to the Swin Transformer network. The Swin Transformer network computes self-attentions for local windows instead of on the whole input image as performed by the traditional Transformer method. We assume that the local window size is $M \times M$. The feature map $\boldsymbol{F}$ can be divided into $\frac{H}{4M} \times \frac{W}{4M}$ non-overlapping windows. We use $\boldsymbol{I} \in \mathbb{R}^{M \times M \times C}$ to represent the feature map of each local window, which is then used to calculate window-based self-attention (W-SA) as follows:

$$\begin{aligned} &\text{Linear projection}: \boldsymbol{Q} \leftarrow \boldsymbol{W}_q \boldsymbol{I}, \boldsymbol{K} \leftarrow \boldsymbol{W}_k \boldsymbol{I}, \boldsymbol{V} \leftarrow \boldsymbol{W}_v \boldsymbol{I} \\ &\text{W-SA}: \boldsymbol{F}_{W-SA} = \text{Softmax}\left(\boldsymbol{Q}\boldsymbol{K}^{\top}/\sqrt{d} + \boldsymbol{B}\right)\boldsymbol{V} \end{aligned} \tag{4}$$

where the input $\boldsymbol{I}$ is linearly projected into three subspaces with weights $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, and $\boldsymbol{W}_v$ to obtain $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$. In the self-attention computation process, the interaction between $\boldsymbol{K}$ and $\boldsymbol{Q}$ is computed by the dot product. Then, the weight is scaled and projected into space $\boldsymbol{V}$ to obtain a W-SA based feature embedding $\boldsymbol{F}_{W-SA}$. The self-attention operation is performed multiple times in parallel and these results are concatenated to form the multi-head window-based self-attention features $\boldsymbol{F}_{W-MSA}$.

The regular partition-based W-SA considers all pixels of the local window. Thus, it cannot capture context information across local windows. Swin Transformer overcomes this problem by adding another shift window operation to obtain shift-window-based self-attention (SW-SA), which displaces the original local window partitions by $(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor)$ pixels and then recomputes another set of window-based attentions (Liu et al., 2021). In particular, a Swin Transformer block with two layers can be calculated as follows.

$$\begin{aligned} \hat{\mathbf{z}}^l &= \text{W}-\text{MSA}\left(\text{LN}\left(\mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1} \\ \mathbf{z}^l &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^l\right)\right) + \hat{\mathbf{z}}^l \\ \hat{\mathbf{z}}^{l+1} &= \text{SW}-\text{MSA}\left(\text{LN}\left(\mathbf{z}^l\right)\right) + \mathbf{z}^l \\ \mathbf{z}^{l+1} &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^{l+1}\right)\right) + \hat{\mathbf{z}}^{l+1} \end{aligned} \tag{5}$$

These are the two layers of a Swin Transformer block, which calculate W-MSA and SW-MSA, respectively. In the first layer, the output $\mathbf{z}^{l-1}$ of the $(l-1)$th layer is adopted as the input to the lth layer, which passes through the layer normalization (LN) layer, and then the W-SA operation of (4) is performed. After that, the window-based attention weight is imposed on the input feature embedding $\mathbf{z}^{l-1}$ by residual connection to form the intermediate features $\hat{\mathbf{z}}^l$. Next, an LN, a multilayer perceptron (MLP), and a residual connection is performed sequentially to obtain the output features $\mathbf{z}^l$ of the lth layer. In the second layer, the structure of SW-MSA is similar to that of the W-MSA layer, except that the contexts in each window are different.

4. Experimental results and discussions

This section first introduces the datasets utilized and detailed experimental setups for SSL pretraining and downstream experiments. Next, we describe in detail the evaluation metrics used for the downstream tasks. Finally, we conduct five types of downstream experiments to validate the universal applicability of the proposed SSL feature learning method, including patch retrieval, patch classification, weakly-supervised WSI classification, mitosis detection, and colorectal adenocarcinoma gland segmentation. These experiments include ablation study, comparisons with state-of-the-art methods on these downstream datasets, and comparisons with different network pretraining methods (e.g., both supervised and self-supervised ImageNet-based pretraining, and several other SSL methods).

4.1. Datasets

We collected the largest histopathological image data as publicly available for our self-supervised pretraining, containing around 15 million unlabeled patches cropped from WSIs in TCGA and PAIP. After the pretraining process, we evaluate its feature learning ability on five types of downstream tasks covering nine datasets: patch retrieval (on UniToPatho and TissueNet), patch classification (on NCT-CRC-HE and Colorectal cancer), weakly-supervised WSI classification (on Camelyon16, TCGA-NSCLC, and TCGA-RCC), mitosis detection (on MIDOG), and colorectal adenocarcinoma gland segmentation (on CRAG). All these datasets are introduced below.

TCGA. TCGA[3] is a public large-scale multi-modal dataset, which contains genome, epigenome, transcriptome, and image data. This work

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

only considers the image data (frozen and formalin-fixed paraffin-embedded (FFPE) slides), which includes a total of 30,072 WSIs covering over 25 anatomic sites and over 32 cancer subtypes. For each WSI, a primary diagnosis is provided for the entire WSI but no detailed annotations. 309 WSIs are removed due to the lack of magnification information. In total, we collect 29,763 WSIs at $20\times$ from 10,953 patients. After excluding regions without tissues, we crop these WSIs into non-overlapping patches with the size of $1,024 \times 1,024$ pixels. Finally, we generate a TCGA pretraining dataset with 14,325,848 unlabeled histopathological patches.

PAIP. PAIP[4] (Kim et al., 2021) provides 2457 WSIs collected from three centers (Seoul National University Hospital, Seoul National University Bundang Hospital, and SMG-SNU Boramae Medical Center), which cover six cancer types, including 571 WSIs from liver, 400 WSIs from renal, 900 WSIs from colorectal, 400 WSIs from prostatic, 166 WSIs from pancreatic, and 20 WSIs from cholangio cancers. Although region-of-interest (ROI) annotations are provided in this dataset, we do not use any labeling information for SSL pretraining. Following the similar image extraction strategy as the TCGA dataset, we produce a PAIP pretraining dataset with 1,254,414 unlabeled histopathological patches.

UniToPatho. UniToPatho[5] (Barbano et al., 2021) is a well-annotated patch-level dataset released for colorectal polyp classification: normal tissue (NORM), hyperplastic polyp (HP), tubular adenoma (TA), and tubulo-villous adenoma (TVA). This dataset contains 8,699 patches ($1,812 \times 1,812$ pixels) cropped from 292 WSIs ($20\times$) for the four type tissue classification task.

TissueNet. TissueNet[6] is released in the *TissueNet: Detect Lesions in Cervical Biopsies* challenge, which aims to classify epithelial lesions of the uterine cervix into four classes: benign (class 0), low malignant potential lesion (class 1), high malignant potential lesion (class 2), and invasive cancer (class 3). As presented in this challenge, TissueNet contains 1,016 WSIs and 5,926 locally labeled patches (300×300 micrometers) within these WSIs. The size of 300×300 micrometers is equivalent to approximately $1,200 \times 1,200$ pixels. Only these patches are used for the patch retrieval experiments.

NCT-CRC-HE. NCT-CRC-HE[7] is provided to identify nine tissues, including eight colorectal cancer tissues and one normal tissue (Kather et al.). The training set contains a total of 100,000 images (extracted from 86 WSIs) with a size of 224×224 pixels. An independent set of 7,180 images are used for testing.

Colorectal cancer (CRC). CRC[8] (Kather et al., 2016) is proposed for colorectal classification task. It is composed of 5,000 patches with the size of 150×150 pixels (74×74 microns) and covers eight different tissue types (625 patches for each type), including epithelium, simple stroma, complex stroma, lymphoid follicles, debris, mucosal glands, adipose and background ROIs with no tissue.

Camelyon16. Camelyon16[9] is released in the *Camelyon16* challenge (Bejnordi et al., 2017) for two types of breast cancer classification: benign tissue and metastatic breast cancer, which contains a total of 399 WSIs at $40\times$ (270 WSIs for training and 129 WSIs for testing). Although this dataset provides exhaustive pixel-level annotations, we only utilize global WSI-level annotations for the weakly-supervised classification task.

TCGA-NSCLC. TCGA-NSCLC is collected from the TCGA dataset for two types of lung cancer classification: lung squamous cell carcinoma (TCGA-LUSC) and lung adenocarcinoma (TCGA-LUAD), which consists of a total of 993 FFPE WSIs (507 WSIs with LUAD and 486 WSIs with LUSC).

TCGA-RCC. TCGA-RCC is a subset of TCGA for the classification of three subtypes of kidney tumor: kidney chromophobe renal cell carcinoma (TCGA-KICH), kidney renal clear cell carcinoma (TCGA-KIRC), and kidney renal papillary cell carcinoma (TCGA-KIRP). There are a total of 884 FFPE WSIs, including 111 KICH WSIs, 489 KIRC WSIs, and 284 KIRP WSIs.

MIDOG. MIDOG[10] is released in the MICCAI MIDOG 2021 challenge for the mitosis detection (Aubreville et al., 2022). The publicly available training set contains 150 WSIs with a size of $8,000 \times 8,000$ pixels, which are cropped into 79,399 patches (6,699 with mitosis) with a size of 256×256 pixels. In our experiments, these 150 WSIs are divided into training, validation, and test sets with a ratio of 7:1:2.

CRAG. CRAG[11] is proposed for colorectal adenocarcinoma gland (CRAG) segmentation, which contains 213 images with a size of around $1,512 \times 1,516$ pixels (Awan et al., 2017; Graham et al., 2019). Following official settings (Graham et al., 2019), these images are randomly split into 173 training images and 40 testing images. Then, 20% of these training images are picked out for parameter validation.

4.2. Experimental setups

In the pretraining stage, we use our proposed SRCL-based framework to train the *CTransPath* model with a mini-batch of 1,024. Histopathology-oriented data augmentation strategies are adopted (Tellez et al., 2019), including random cropping, Gaussian blur, and hue and saturation shifting in the HSV color space. Following MoCo v3 (Chen et al., 2021), τ in the contrastive loss is set as 0.2. AdamW (Loshchilov and Hutter, 2018) is adopted as the optimizer with an initial learning rate of 0.00015. The learning rate is updated using a cosine decay schedule with a long warmup of 40 epochs. The number of new positive pairs S is set as four and the number of epochs for traditional CL training (a short warmup) is set as five, which will be explained in the following ablation experiments. Our method is implemented using the PyTorch package and the SRCL model pretraining takes around 250 h to converge using 48 Nvidia V100 GPUs. It is noted that the number of iterations is set to 100 epochs to ensure convergence for the pre-training of both our SRCL model and all other SSL models compared. After self-supervised pretraining, the pretrained backbones can be fine-tuned or used directly for various downstream tasks.

The downstream experiments are divided into five main categories: patch retrieval, patch classification, WSI classification, mitosis detection, and colorectal adenocarcinoma gland segmentation. (1) The patch retrieval does not require any further fine-tuning, which can be regarded as an inference procedure. (2) The patch classification is evaluated using standard linear probing, which is implemented by training a supervised linear classifier (a fully connected layer) on top of the frozen *CTransPath*. To train the linear classifier, Adam is used as the optimizer with a batch size of 96. The initial learning rate is set as 0.0003. The data augmentations include random horizontal, vertical, and 90-degree flipping, and random scaling. (3) For the WSI classification task, the pretrained *CTransPath* model is also frozen. Then, following CLAM (Lu et al., 2021), we adopt Adam as the optimizer with an initial learning rate of 0.0002 and a weight decay of 0.00001. The mini-batch size is set to 1 (WSI/bag). (4) The mitosis detection task is solved using the Faster RCNN method (Ren et al., 2015) with our pretrained *CTransPath* as the encoder. The batch size is set to 64 and Adam is employed as optimizer with an initial learning rate of 0.0003. The learning rate reduces by a factor of 10 at the 10th and 20th epochs. It takes around a total of 30 training epochs to converge. The utilized data augmentation strategies include random cropping, random horizontal/vertical flipping, random

[4] http://wisepaip.org/paip
[5] https://ieee-dataport.org/open-access/unitopatho
[6] https://www.drivendata.org/competitions/67/competition-cervical-biopsy/page/254/
[7] https://zenodo.org/record/1214456#.YVrmANpBwRk
[8] https://zenodo.org/record/53169#.YRfeKYgzbmE
[9] https://camelyon16.grand-challenge.org/
[10] https://midog2021.grand-challenge.org/
[11] https://warwick.ac.uk/fac/sci/dcs/research/tia/data/mildnet

Table 1
Ablation study. Sup. denotes the supervised pretraining process. ImageTrans and HistoTrans denote ImageNet-pretrained Swin Transformer and histopathology-pretrained Swin Transformer, respectively. HistoTrans+CNN means our *CTransPath* backbone. SN denotes spatial-neighbor-based contrastive learning method.

Methods	TissueNet				UniToPatho			
	ACC@1	ACC@3	ACC@5	mMV@5	ACC@1	ACC@3	ACC@5	mMV@5
ImageTrans (Sup.)	0.5324	0.7892	0.8799	0.5035	0.5334	0.7899	0.8708	0.5463
ImageTrans (SSL)	0.5618	0.8171	0.9047	0.5565	0.5749	0.8103	0.8803	0.5799
HistoTrans+CL	0.6051	0.8395	0.9220	0.5910	0.5935	0.8194	0.8891	0.6011
HistoTrans+CNN+CL	0.6239	0.8405	0.9109	0.6247	0.6183	0.8234	0.8837	0.6294
HistoTrans+CNN+SN	0.6304	0.8488	0.9158	0.6363	0.6201	0.8305	0.8928	0.6332
HistoTrans+CNN+SRCL (ours)	**0.6505**	**0.8606**	**0.9261**	**0.6617**	**0.6329**	**0.8370**	**0.8966**	**0.6417**

Table 2
Effect of different values of S for the number of pseudo-positives on the patch retrieval accuracy.

		$S=0$	$S=1$	$S=2$	$S=4$	$S=6$	$S=8$
TissueNet	ACC@1	0.6239	0.6333	0.6409	**0.6505**	0.6498	0.6417
	mMV@5	0.6247	0.6338	0.6527	**0.6617**	0.6610	0.6537
UniToPatho	ACC@1	0.6183	0.6194	0.6211	**0.6329**	0.6271	0.6223
	mMV@5	0.6294	0.6340	0.6370	**0.6417**	0.6370	0.6311

Table 3
Effect of different number of epochs for warmup on the patch retrieval accuracy.

Epoch		0	2	5	10
TissueNet	ACC@1	0.6304	0.6417	0.6505	0.6493
	mMV@5	0.6358	0.6525	0.6617	0.6606
UniToPatho	ACC@1	0.6209	0.6286	0.6329	0.6309
	mMV@5	0.6332	0.6363	0.6417	0.6407

scaling, and random color jitter. Focal loss is used as the objective function. (5) The colorectal adenocarcinoma gland segmentation task is implemented using the U-Net framework (Ronneberger et al., 2015) with our pretrained *CTransPath* as the encoder. The training loss is a combination of Dice and cross-entropy losses. The other parameter settings are kept the same as that for the mitosis detection task.

4.3. Evaluation metrics

Image classification, retrieval, detection, and segmentation experiments are conducted to evaluate the effectiveness of our SSL-pretrained feature extractor. For the classification task, accuracy (ACC), area under the curve (AUC) score, and F1 score are used for performance evaluation. For image retrieval, querying an image will return a series of similar images. Based on these returns, $ACC@k$ (top-k accuracy) and $mMV@k$ (majority vote at the top k search returns) are calculated as evaluation metrics. For a query image, $ACC@k$ will be 1 if any one of the top-k returns has the same label as the query image, otherwise it will be 0. Compared to $ACC@k$, $mMV@k$ is a stricter metric since $mMV@k$ will be 1 only if the majority of these retrieved images have the same label as the query image. For the detection and segmentation tasks, F1 and Dice scores are respectively used following the convention of the original works (Aubreville et al., 2022; Graham et al., 2019).

4.4. Results of patch retrieval

This subsection conducts patch retrieval to validate the robustness and transferability of our pretrained histopathology-oriented features. The patch retrieval process contains two steps: i) feature extraction for searching database and ii) similarity measurement across these features. The first step can be implemented by passing every sample of the searching database to the pretrained model to generate the sample features. In the second step, the feature of a query image is compared with all features in the searching database based on the leave-one-patient-out validation strategy. These retrieved images can be sorted in descending order of similarity scores to calculate the $ACC@k$ and $mMV@k$ metrics. From the above introduction, it can be seen that the patch retrieval results can directly reflect the feature learning ability of our *CTransPath* method. Two patch-level datasets (TissueNet and UniToPatho) with subtype annotations are employed for the image retrieval experiment, which aims to retrieve images with the same cancer subtypes. In the patch retrieval experiments, we first conduct an ablation study to validate the effectiveness of key components in the design of backbone and CL strategy as shown in Table 1, Table 2, and Table 3. Then, we compare our proposed SRCL method with other SSL baselines as shown in Table 4.

4.4.1. Ablation study

Our ablation study first investigates the effects of three key components of our SSL algorithm: in-domain SSL pretraining strategy, hybrid CNN and Transformer encoder, and semantically-relevant CL algorithm, the results of which are summarized in Table 1. Then, the effects of different S values and different numbers of epochs for traditional CL training (a short warmup) are shown in Table 2 and Table 3, respectively. In Table 1, ImageTrans and HistoTrans denote ImageNet-pretrained Swin Transformer and Histopathology-pretrained Swin Transformer, respectively. HistoTrans+CNN means using the proposed *CTransPath* as the backbone. ImageNet-pretrained weights of Swin Transformer are obtained directly from previous studies (Liu et al., 2021; Xie et al., 2021).

Benefit of in-domain SSL pretraining: As shown in the first three rows of Table 1, with traditional CL loss and Swin Transformer backbone, replacing the ImageNet training data with large-scale histopathology datasets brings an improvement of about +4% in terms of both $ACC@1$ and $mMV@5$ when tested on the TissueNet data. This indicates that the self-supervised histopathology-oriented feature extractor could significantly boost the retrieval performance compared with the self-supervised ImageNet-pretrained feature extractor. It can also be seen from Table 1 that the self-supervised ImageNet pretraining provides higher accuracy than the supervised ImageNet pretraining for this histopathology image retrieval task, which is consistent with results obtained by previous studies (Hosseinzadeh Taher et al., 2021, 2022).

Benefit of hybrid CNN and Transformer encoder: To alleviate the weak local feature extraction problem of Transformer, we replace the patch partition of Transformer using a CNN module, which yields consistent performance gains in all four metrics on both datasets (e.g., +2% for $ACC@1$ and +3% for $mMV@5$ on TissueNet).

Benefit of semantically-relevant positives: To alleviate biased definition of positive and negative samples in traditional CL, we modify the contrastive loss by mining more positives in a large memory bank. Our memory bank is only used for searching several similar samples as positives, which differs from the previous methods (e.g., MoCo (He et al., 2020)) that regard all samples in the memory bank as negatives. We implement this by adding an SRCL loss function, which achieves an obvious performance improvement compared to the traditional contrastive loss (e.g., +3% and +4% in terms of $ACC@1$ and $mMV@5$ on the TissueNet). Furthermore, we also compare our SRCL with previous spatial-neighbor-based contrastive learning strategy that adopts any two spatially adjacent patches as positives (Abbet et al.,

Table 4
Results of patch retrieval and comparison with other state-of-the-art SSL frameworks. Note that the implementation of all other SSL methods is based on their publicly available code but the backbone model and the training data are switched to be the same as ours.

Methods	TissueNet				UniToPatho			
	ACC@1	ACC@3	ACC@5	mMV@5	ACC@1	ACC@3	ACC@5	mMV@5
SimCLR (Chen et al., 2020)	0.6019	0.8297	0.9036	0.6021	0.6070	0.8149	0.8819	0.6170
MoBY (Xie et al., 2021)	0.6131	0.8360	0.9077	0.6077	0.6110	0.8197	0.8873	0.6173
DINO (Caron et al., 2021)	0.6169	0.8481	0.9183	0.6183	0.6149	0.8224	0.8909	0.6222
MoCo v3 (Chen et al., 2021)	0.6239	0.8405	0.9109	0.6247	0.6183	0.8234	0.8837	0.6294
SRCL (ours)	**0.6505**	**0.8606**	**0.9261**	**0.6617**	**0.6329**	**0.8370**	**0.8966**	**0.6417**

Table 5
Linear evaluation results on NCT-CRC-HE dataset with different sizes of training data. ImageTrans (Sup.) and ImageTrans (SSL) refer to models pre-trained using the ImageNet data in a supervised and self-supervised manner, respectively. All other compared SSL frameworks are pretrained using our training data. A supervised baseline using 100% of the training data achieves an F1 score of 0.9295 and an ACC of 0.9458.

Methods	Backbone	Percentage of training data									
		0.5%		1%		10%		50%		100%	
		F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
ImageTrans (Sup.)	Swin	0.4770	0.5703	0.5584	0.6157	0.7780	0.8139	0.8206	0.8585	0.8323	0.8705
ImageTrans (SSL)	Transformer	0.6997	0.7348	0.7816	0.8213	0.8715	0.9035	0.8867	0.9171	0.8903	0.9216
SimCLR	ResNet50	0.8062	0.8503	0.8331	0.8649	0.8613	0.9195	0.8971	0.9269	0.9025	0.9315
BYOL		0.8234	0.8636	0.8649	0.8965	0.8876	0.9245	0.9050	0.9324	0.9144	0.9413
SimSiam		0.8348	0.8730	0.8660	0.9054	0.8903	0.9286	0.9085	0.9387	0.9144	0.9457
MoCo v2		0.8435	0.8875	0.8702	0.9130	0.9050	0.9362	0.9156	0.9467	0.9213	0.9514
SimCLR	CTransPath	0.8204	0.8581	0.8439	0.8716	0.8740	0.9297	0.9043	0.9316	0.9081	0.9349
MoBY		0.8317	0.8699	0.8578	0.8965	0.8919	0.9357	0.9144	0.9442	0.9156	0.9465
DINO		0.8355	0.8799	0.8671	0.9128	0.8915	0.9369	0.9135	0.9438	0.9198	0.9502
MoCo v3		0.8682	0.8978	0.8739	0.9228	0.9046	0.9415	0.9208	0.9516	0.9254	0.9548
SRCL (ours)		**0.8988**	**0.9266**	**0.9334**	**0.9539**	**0.9420**	**0.9635**	**0.9474**	**0.9648**	**0.9482**	**0.9652**

2020), which can be seen in the last two rows of Table 1. To conduct a fair comparison, we adopt four adjacent patches as new positives similar to our SRCL. As seen in Table 1, treating spatial neighbors as positives offers higher accuracy than treating these augmented views from the same instance as positives, but less effective than treating semantically-related patches as positives. The reason may be that neighboring patches represent only local similarity within WSIs, while our method can find globally similar patches across WSIs and guarantees more sample diversity.

Effect of different S values: We also conduct an ablation experiment on the TissueNet and UniToPatho datasets to explore the influence of different S values on the patch retrieval task, the results of which are shown in Table 2. Although the performance of our method is relatively stable for S values varying from 2 to 8, it is seen that $S = 4$ is the optimal setting and larger S shows slightly degraded performance. The reason can be explained from two aspects. First, these selected potential positives are pseudo-positives since the patches have no labels. Second, the sizes of both the memory bank and mini-batch are fixed, which are two independent containers for the positive mining and contrastive loss calculation, respectively. Thus, a very large S may introduce some hard or false positives, posing a challenge for discriminative feature learning.

Effect of different numbers of epochs for warmup: Intuitively, the results of SRCL-based positive sample selection may be unreliable in the early training stage because the feature is not well learned yet. Thus, we employ traditional CL loss (e.g., loss in MoCo v3 (Chen et al., 2021) to warm up the model training in the first several epochs. Therefore, this ablation experiment is conducted to investigate how many epochs are suitable for this warmup strategy. The detailed results are shown in Table 3. It is seen that a small number of epochs can bring a better performance gain compared with the situation without warmup. We empirically set the number of epochs for warmup as five.

4.4.2. Comparison with other SSL methods

We compare our SRCL approach with other SSL strategies in Table 4, including SimCLR (Chen et al., 2020), MoBY (Xie et al., 2021), DINO (Caron et al., 2021), and MoCo v3 (Chen et al., 2021). In their official implementations, all the methods use Transformer as the backbone except SimCLR. SimCLR constructs two symmetrical branches with shared weights to perform CL. MoCo v3 maintains a similar structure as SimCLR but differs in the momentum encoder. DINO maintains a similar teacher–student structure as MoCo but uses a cross-entropy loss to directly predict the output of the teacher network, which avoids model collapsing through centering and sharpening of the teacher branch. MoBY inherits two asymmetric encoders as BYOL (Grill et al., 2020), while retaining the memory bank used in MoCo to store negative samples for the calculation of contrastive loss. Since here we aim to compare the performance differences caused by different SSL strategies, we directly use the published code of these methods to conduct SSL pre-training but switch the backbone model and the training data to be the same as our SRCL method.

As shown in Table 4, our method achieves the best retrieval performance. SimCLR with our *CTransPath* backbone produces the lowest performance. The reason may be that it lacks a momentum encoder, which has been demonstrated as a key factor for performance improvement in SSL training (Tao et al., 2021). Compared with SimCLR, MoBY and DINO provide similar performance gains, which may be due to the fact that both methods have a similar asymmetric structure like BYOL (Grill et al., 2020). However, MoBY is slightly lower than DINO, which may be due to the memory bank used for negative sample storage in MoBY. It has been demonstrated that a memory bank may cause diminishing gain if the batch is sufficiently large (Chen et al., 2021). MoCo v3 keeps the momentum encoder but abandons the memory bank, which produces the best performance only second to ours as shown in Table 4. For instance, it exceeds SimCLR by about +2% on the TissueNet and +1% on the UniToPatho in terms of $ACC@1$ and $mMV@5$. Compared against MoCo v3, our method offers an improvement of around +3% for $ACC@1$ and +4% for $mMV@5$ on the TissueNet dataset and +1.5% of $ACC@1$ and +1% of $mMV@5$ on UnitoPatho dataset. This performance improvement is mainly attributed to our semantic-relevance strategy.

4.5. Results of patch classification

Our SSL pretrained feature extractor can serve as a universal representation learning method for histopathological images. To further

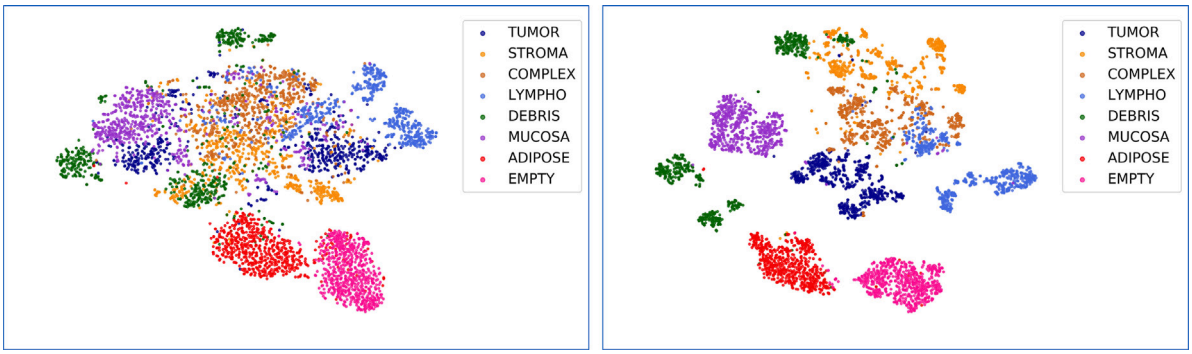

Fig. 3. Visualization (t-SNE) of the classification performance of all images in the CRC dataset based on the features generated from KimiaNet (left) and SRCL-pretrained *CTransPath* (right).

Table 6
Results of downstream classification tasks performed on CRC dataset (SVM classification)

Methods	ACC
Combined texture descriptors (Kather et al., 2016)	87.40
Ensemble of DNNs (Ghosh et al., 2021)	92.83
Fine-tuned VGG-19 (Faust et al., 2018)	93.58
KimiaNet (Riasatian et al., 2021)	96.80
Ensemble of CNNs (Nanni et al., 2021)	97.60
Ours	**98.20**

Table 7
Results of weakly-supervised classification on three public datasets.

	CAMELYON16		TCGA-NSCLC		TCGA-RCC	
	ACC	AUC	ACC	AUC	ACC	AUC
Kernel ATT (Rymarczyk et al., 2021)	0.773	0.804	0.841	0.921	0.856	0.945
C2C (Sharma et al., 2021)	0.809	0.841	0.849	0.921	0.909	0.972
MIL-RNN (Campanella et al., 2019)	0.819	0.856	0.856	0.931	0.914	0.974
AbMIL (Ilse et al., 2018)	0.820	0.857	0.838	0.920	0.902	0.980
CLAM-MB (Lu et al., 2021)	0.835	0.854	0.863	0.938	0.925	0.988
CLAM-SB (Lu et al., 2021)	0.837	0.873	0.859	0.938	0.921	0.987
TransMIL (Shao et al., 2021)	0.884	0.931	0.884	0.960	0.947	0.988
DSMIL (Li et al., 2021b)	0.899	0.917	**0.929**	0.958	–	–
CLAM-SB + Ours	**0.922**	**0.942**	0.912	**0.973**	**0.967**	**0.991**

verify the generalizability of the features, this subsection studies downstream classification tasks on two publicly available datasets: NCT-CRC-HE and CRC. On the NCT-CRC-HE dataset (Table 5), following the common linear evaluation protocol (Chen et al., 2020), we conduct experiments on training data of different sizes to investigate the classification performance under limited labeling settings. Also, we compare the performance of our method with state-of-the-art SSL methods. On the CRC dataset (Table 6), to conduct a fair comparison with KimiaNet (Riasatian et al., 2021), an SVM-evaluation is conducted to compare the results of different methods.

In Table 5, linear evaluation is performed by freezing the pretrained *CTransPath* backbone and training a fully connected layer for classification. We explore the performance variation of our SSL-pretrained features on different proportions of downstream training data, especially with limited annotations. For data splitting, we keep the same test data as the official setup, but randomly select 0.5%, 1%, 10%, 50%, and 100% of training data for comparison. We compare the performance produced by our SRCL-based histopathology-specific pretraining with that of supervised ImageNet-pretraining, self-supervised ImageNet-pretraining, and histopathology-specific pretraining by other SSL frameworks (CNN-based and Transformer-based backbones). These CNN-based SSL frameworks include SimCLR (Chen et al., 2020), BYOL (Grill et al., 2020), SimSiam (Chen and He, 2021), and MoCo v2 (Chen et al., 2020) and Transformer-based SSL frameworks include SimCLR (Chen et al., 2020), MoBY (Xie et al., 2021), DINO (Caron et al., 2021), and MoCo v3 (Chen et al., 2021). ImageTrans (Sup.) and ImageTrans (SSL) adopt the Swin Transformer as the backbone, which are consistent with those in Table 1. It is noted that these three kinds of backbones have a similar computational complexity (Liu et al., 2021). As shown in Table 5, it can be seen that our SRCL-pretrained model achieves consistently higher performance compared to all other methods. Specifically, our method exceeds the fully supervised baseline (backbone training with ImageNet initialization and 100% labeled data) with only 1% of the labeled data. Moreover, our performance gains are even more significant when the labeled training data is limited. For instance, when the annotation rate of the training data increases from 0.5% to 100%, the performance gap between our SRCL-pretrained model and SSL-based ImageNet-pretrained one reduces from 20% to 6% in terms of F1 score. Furthermore, to conduct a direct comparison with our previous *TransPath* method (Wang et al., 2021b), we freeze the pretrained *TransPath* model to perform the classification on the NCT-CRC-HE dataset with 100% training data. It produces an F1 score of 0.9008 and an ACC of 0.9405, which are inferior to the current method. These results demonstrate that the features learned by our extended version (SRCL-pretrained *CTransPath*) have better discriminative power.

In Table 6, we freeze our backbone and train an SVM to perform the classification task. To perform a fair comparison with KimiaNet (Riasatian et al., 2021), we run a 10-fold cross-validation to evaluate the classification performance. The results of these state-of-the-art methods are copied from their respective publications. As shown in Table 6, our method achieves the highest performance, which exceeds the previous best-reported method (model ensemble) by +0.6%. Specifically, KimiaNet is similar to us in that it applies a histopathology-oriented feature extractor pretrained on the TCGA using weak annotations. Our SRCL-pretrained *CTransPath* exceeds KimiaNet by 1.4% in terms of ACC. Also, a t-SNE visualization is conducted to visually compare the discriminative power between features generated by KimiaNet and our SRCL-pretrained *CTransPath*. As shown in Fig. 3, it is seen that our method can better push away inter-class samples and pull together intra-class ones, which proves the strong feature discrimination ability of our method and demonstrates that our SSL-based feature embedding transfers better to downstream classification tasks.

4.6. Results of weakly-supervised WSI classification

To further verify the discriminative capacity of our proposed self-supervised representation learning, we conduct a weakly-supervised classification experiment on three WSI-level datasets: CAMELYON16, TCGA-NSCLC, and TCGA-RCC. Meanwhile, we compare the weakly-supervised classification results based on SRCL-pretrained features with current state-of-the-art methods, which are detailed in Table 7. The weakly-supervised classification problem at the WSI level is defined as giving only global annotations (slide level) without details of internal regions. The current weakly-supervised algorithms developed

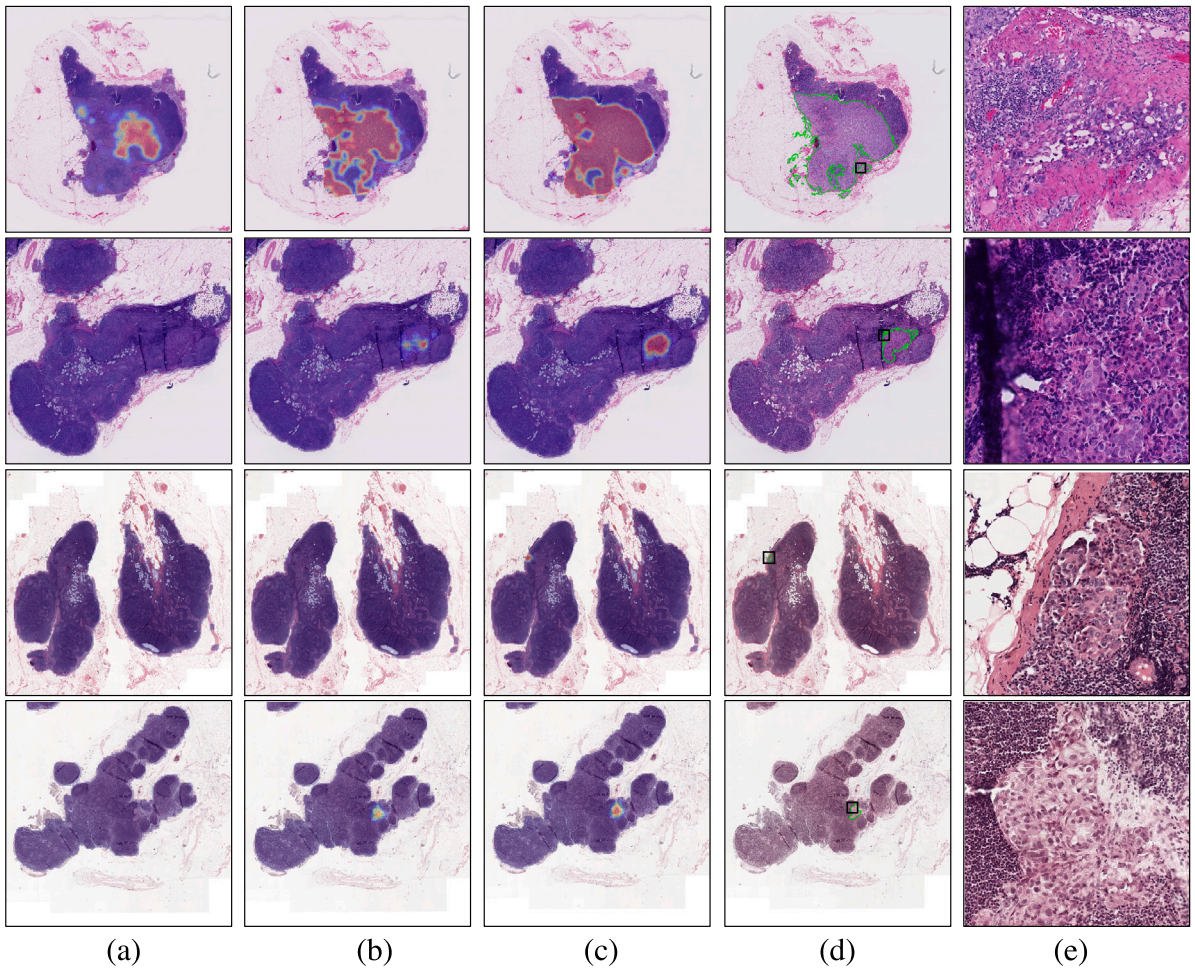

Fig. 4. Interpretability and visualization for some good cases. Subfigures (a)–(c) show the original WSIs overlaid with the automatically computed attention heatmaps. Warmer colors of the attention map indicate higher estimated probabilities of being tumorous tissue. (a) CLAM-SB (ImageNet pretraining in a supervised manner) (Lu et al., 2021). (b) DSMIL (histopathology pretraining in a self-supervised manner) (Li et al., 2021b). (c) Our SRCL method. (d) Ground truth. In (d), these green lines represent the ground truth of the cancer metastasis, while dark rectangles indicate local ROIs highlighting the boundaries between metastatic and normal tissues, as shown in (e). These four WSIs come from the CAMELYON16 test dataset.

for WSI classification can be divided into two steps: (i) feature extraction for patches cropped from WSIs and (ii) feature aggregation for these patch features. These feature extraction methods include ImageNet-pretraining, end-to-end training, and SSL-pretraining. These aggregation algorithms contain attention-based pooling and RNN-based (or Transformer-based) feature fusion. In our weakly-supervised classification method, the feature extractor adopts our SRCL-pretrained *CTransPath*, while the feature aggregator directly utilizes that of CLAM-SB (attention-based pooling) (Lu et al., 2021). In our implementation, the data split process is kept consistent with the proposer of the CAMELYON16 dataset (Bejnordi et al., 2017). For TCGA-NSCLC and TCGA-RCC datasets, we use 5-fold cross-validation to organize the experiment. It is noted that the results of TransMIL and DSMIL are copied directly from their publications and the remaining methods are implemented using their released codes.

As shown in Table 7, our weakly-supervised classification results consistently outperform all other methods. Especially on the CAMELYON16 dataset, our algorithm obtains an ACC of 0.922 and an AUC of 0.942, which has an improvement of around +9% and +7% compared to the CLAM-SB method. A similar phenomenon can be seen on the TCGA-NSCLC and TCGA-RCC datasets. The results further validate the powerful feature learning of our SRCL-pretrained *CTransPath*. By comparing with the previously best-performing DSMIL on the CAMELYON16 dataset, our results achieve an improvement of around +3% in both ACC and AUC. However, on the TCGA-NSCLC dataset, our ACC metric is slightly lower than that of DSMIL. The reason may be that our results are averaged over multiple folds, while DSMIL was tested only once. DSMIL is similar to ours in that it also uses an SSL approach on histopathological images to extract features and a similar aggregation scheme based on attention pooling. Therefore, our results can also indirectly indicate that our feature extractor is better than that of DSMIL. Another obvious phenomenon is that the results tested on CAMELYON16 generally have lower accuracy than those tested on TCGA-NSCLC and TCGA-RCC, which is caused by the different percentage of tumor regions in each slide (e.g., $< 10\%$ in CAMELYON16 and $> 80\%$ in TCGA-NSCLC and TCGA-RCC).

We also conduct interpretability and visualization analysis to explain the mechanism behind the weakly-supervised classification. Fig. 4 and Fig. 5 show some examples of good and bad cases, respectively. WSIs from the CAMELYON16 test set are adopted for demonstration due to the availability of detailed region annotations (Fig. 4(d) and Fig. 5(d)). It is noted that these annotations are only used for visual comparison of different results. The visualization results compare the attention heatmaps (Lu et al., 2021) corresponding to the weakly supervised classification with different model pretraining, including the proposed SRCL method (Fig. 4(c) and Fig. 5(c)), ImageNet pretraining (Fig. 4(a) and Fig. 5(a)), and previous state-of-the-art DSMIL method (Fig. 4(b) and Fig. 5(b)). To be consistent with the previous experiments, the supervised ImageNet pretraining adopts Swin Transformer as the feature extractor. Both our method and the supervised ImageNet

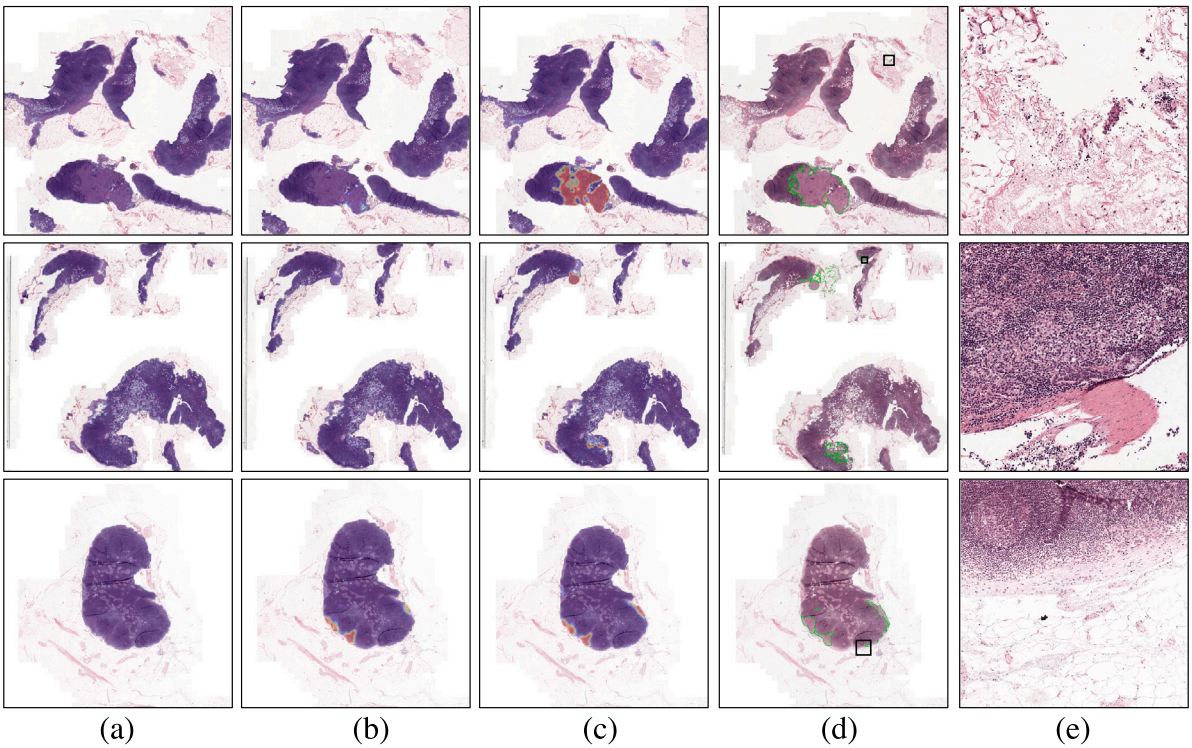

Fig. 5. Interpretability and visualization for some bad cases. Subfigures (a)–(c) show the original WSIs overlaid with the automatically computed attention heatmaps. Warmer colors of the attention map indicate higher estimated probabilities of being tumorous tissue. (a) CLAM-SB (ImageNet pretraining in a supervised manner) (Lu et al., 2021). (b) DSMIL (histopathology pretraining in a self-supervised manner) (Li et al., 2021b). (c) Our SRCL method. (d) Ground truth. In (d), these green lines represent the ground truth of the cancer metastasis, while dark rectangles indicate local ROIs highlighting the boundaries between metastatic and normal tissues, as shown in (e). These three WSIs come from the CAMELYON16 test dataset.

pretraining one utilize the same feature aggregating scheme as the CLAM-SB method. The DSMIL is implemented based on their released code and network weights (Li et al., 2021b). As shown in Fig. 4(a-c) and Fig. 5(a-c), these attention-based heatmaps are generated according to the importance of each sub-region in the classification procedure. As shown in Fig. 4, our weakly-supervised classification method produces very accurately localized tumor heatmaps, which matches the ground truth very well. Especially for these tiny cancerous regions as shown in the last two rows, our method can still correctly capture the lesion regions, which is very challenging even for experienced pathologists. However, our method may fail to detect some hard cases with small or isolated tumor micrometastasis as shown in Fig. 5. These hard cases can also be easily missed by pathologists if only observing the H&E stained slides but without the assistance of extra immunohistochemical staining (Bejnordi et al., 2017; Weaver, 2010). In summary, the visualized results further demonstrate that our features have the potential to delineate tumor boundaries in combination with weakly-supervised aggregation methods, even with only WSI-level annotations.

4.7. Results of downstream detection and segmentation tasks

To further verify the generalizability of our feature extractor (SRCL-pretrained *CTransPath*), we construct two experiments for mitosis detection and colorectal adenocarcinoma gland segmentation by full network fine-tuning, as shown in Table 8. Faster RCNN (Ren et al., 2015) and U-Net (Ronneberger et al., 2015) are employed as the detection and segmentation frameworks, respectively. In the current implementation, their encoders are initialized by the pretrained *CTransPath* model and decoders are initialized randomly. And then, the detection and segmentation networks are retrained based on full supervision. Our method is also compared with ImageNet-pretrained Swin Transformer in both a fully supervised manner and a self-supervised setting, and histopathology-pretrained *CTransPath* using different SSL strategies. As shown in Table 8, our method outperforms previous best-performing SSL strategies by around +1% in both tasks. Also, by comparing against ImageNet-based network pretraining (out-of-domain data), our method demonstrates that the in-domain pretraining has the ability to learn better feature representations.

Table 8
Results of downstream mitosis detection and colorectal adenocarcinoma gland segmentation tasks via full network fine-tuning. The ImageTrans adopts Swin Transformer as the encoder and the four compared SSL frameworks employ our *CTransPath* as the encoder.

Model	Mitosis detection (F1)	CRAG segmentation (Dice)
ImageTrans (Sup.)	0.6842	0.8743
ImageTrans (SSL)	0.6958	0.8824
SimCLR	0.7078	0.8962
MoBY	0.7110	0.9010
DINO	0.7083	0.8996
MoCo v3	0.7204	0.9050
Ours	**0.7332**	**0.9156**

5. Conclusion

We propose a customized SSL architecture for various histopathological image analysis, which contains a hybrid CNN-transformer backbone (*CTransPath*) and a semantically-relevant contrastive learning (SRCL) strategy. Our *CTransPath* makes use of both local and global receptive fields to extract discriminative and rich features. Motivated by the biased assumption in traditional CL, our SRCL aims to select more semantically relevant positives to increase the sample diversity in the instance discrimination process. In addition to data augmentations from the same instance, our positive samples also contain augmented data from semantically similar samples in the feature space, aiming to extract features with better discriminative capacity. Our

SRCL pretrained *CTransPath* on large-scale histopathological images has the potential to benefit various downstream tasks by transfer learning or direct feature extraction. Experimental evaluations in five different downstream tasks, covering nine public datasets, demonstrate the effectiveness of our pretraining model. As demonstrated in the experiments of patch retrieval, patch classification, mitosis detection, and colorectal adenocarcinoma gland segmentation, our SRCL strategy can remarkably improve the performance compared to other SSL methods. In a limited annotation setting, our method can exceed the performance of a supervised baseline (with 100% training data and ImageNet initialization) using only 1% of the training data. Moreover, our five downstream applications also indicate that our SRCL-pretrained feature learning (unsupervised) outperforms ImageNet-pretrained one (supervised/self-supervised) by a large margin. These results validate that our proposed feature extractor has the potential to be a universal model for various histopathological image applications. In the future, more attempts can be made to reduce the computational complexity and memory consumption during SSL pretraining to achieve similar results.

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.

Acknowledgments

This research was in part funded by the National Natural Science Foundation of China (No. 61571314), Science & technology department of Sichuan Province, China (No. 2020YFG0081), and the Innovative Youth Projects of Ocean Remote Sensing Engineering Technology Research Center of State Oceanic Administration of China (No. 2015001).

References

Abbet, C., Zlobec, I., Bozorgtabar, B., Thiran, J.P., 2020. Divide-and-rule: self-supervised learning for survival analysis in colorectal cancer. In: MICCAI. Springer, pp. 480–489.

Aubreville, M., Stathonikos, N., Bertram, C.A., Klopleisch, R., ter Hoeve, N., Ciompi, F., Wilm, F., Marzahl, C., Donovan, T.A., Maier, A., et al., 2022. Mitosis domain generalization in histopathology images–The MIDOG challenge. arXiv preprint arXiv:2204.03742.

Awan, R., Sirinukunwattana, K., Epstein, D., Jefferyes, S., Qidwai, U., Aftab, Z., Mujeeb, I., Snead, D., Rajpoot, N., 2017. Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. Sci. Rep. 7 (1), 1–12.

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.

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

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature 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. pp. 9912–9924.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660.

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

Chen, X., He, K., 2021. Exploring simple siamese representation learning. In: CVPR. pp. 15750–15758.

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

Chen, X., Xie, S., He, K., 2021. An empirical study of training self-supervised vision transformers. In: ICCV. pp. 9640–9649.

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

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

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A., 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In: ICCV. pp. 9588–9597.

Faust, K., Xie, Q., Han, D., Goyle, K., Volynskaya, Z.I., Djuric, U., Diamandis, P., 2018. Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction. BMC Bioinformatics 19 (1), 1–15.

Ghosh, S., Bandyopadhyay, A., Sahay, S., Ghosh, R., Kundu, I., Santosh, K., 2021. Colorectal histology tumor detection using ensemble deep neural network. Eng. Appl. Artif. Intell. 100, 1–11.

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

Graham, S., Chen, H., Gamper, J., Dou, Q., Heng, P.A., Snead, D., Tsang, Y.W., Rajpoot, N., 2019. MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images. Med. Image Anal. 52, 199–211.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-A new approach to self-supervised learning. 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.

Hosseinzadeh Taher, M.R., Haghighi, F., Feng, R., Gotway, M.B., Liang, J., 2021. A systematic benchmarking analysis of transfer learning for medical image analysis. In: DART FAIR. Springer, pp. 3–13.

Hosseinzadeh Taher, M.R., Haghighi, F., Gotway, M.B., Liang, J., 2022. CAiD: Context-aware instance discrimination for self-supervised learning in medical imaging. arXiv preprint arXiv:2204.07344.

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

Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning. In: ICML. PMLR, pp. 2127–2136.

Javed, S., Mahmood, A., Fraz, M.M., Koohbanani, N.A., Benes, K., Tsang, Y.W., Hewitt, K., Epstein, D., Snead, D., Rajpoot, N., 2020. Cellular community detection for tissue phenotyping in colorectal cancer histology images. Med. Image Anal. 63, 101696.

Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., Jansen, L., Reyes-Aldasoro, C.C., Zörnig, I., Jäger, D., Brenner, H., Chang-Claude, J., Hoffmeister, M., Halama, N., Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLOS Med. 16.

Kather, J.N., Weis, C.A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A., Zöllner, F.G., 2016. Multi-class texture analysis in colorectal cancer histology. Sci. Rep. 6 (1), 1–11.

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., 2021. Self-path: self-supervision for classification of pathology images with limited annotations. IEEE Trans. Med. Imaging 40, 2845–2856.

Li, B., Keikhosravi, A., Loeffler, A.G., Eliceiri, K.W., 2021a. Single image super-resolution for whole slide image using convolutional neural networks and self-supervised color normalization. Med. Image Anal. 68, 101938.

Li, B., Li, Y., Eliceiri, K.W., 2021b. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: CVPR. pp. 14318–14328.

Li, J., Lin, T., Xu, Y., 2021c. SSLP: Spatial guided self-supervised learning on pathological images. In: MICCAI. Springer, pp. 3–12.

Liu, Y., Jain, A., Eng, C., Way, D.H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., et al., 2020. A deep learning system for differential diagnosis of skin diseases. Nature Med. 26 (6), 900–908.

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. In: ICCV. pp. 1–7.

Loshchilov, I., Hutter, F., 2018. Decoupled weight decay regularization. In: ICLR.

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. Nature Biomed. Eng. 5 (6), 555–570.

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

Mormont, R., Geurts, P., Marée, R., 2020. Multi-task pre-training of deep neural networks for digital pathology. IEEE J. Biomed. Health Inform. 25 (2), 412–421.

Nanni, L., Ghidoni, S., Brahnam, S., 2021. Ensemble of convolutional neural networks for bioimage classification. 17, (1), pp. 19–35,

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

Pantazis, O., Brostow, G.J., Jones, K.E., Mac Aodha, O., 2021. Focus on the positives: Self-supervised learning for biodiversity monitoring. In: ICCV. pp. 10583–10592.

Patil, A., Talha, M., Bhatia, A., Kurian, N.C., Mangale, S., Patel, S., Sethi, A., 2021. Fast, self supervised, fully convolutional color normalization of H&E stained images. In: ISBI. IEEE, pp. 1563–1567.

Rashid, R., Chen, Y.A., Hoffer, J., Muhlich, J.L., Lin, J.R., Krueger, R., Pfister, H., Mitchell, R., Santagata, S., Sorger, P.K., 2022. Narrative online guides for the interpretation of digital-pathology images and tissue-atlas data. Nature Biomed. Eng. 6 (5), 515–526.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In: NeurIPS, Vol. 28.

Riasatian, A., Babaie, M., Maleki, D., Kalra, S., Valipour, M., Hemati, S., Zaveri, M., Safarpoor, A., Shafiei, S., Afshari, M., Rasoolijaberi, M., Sikaroudi, M., Adnan, M., Shah, S., Choi, C., Damaskinos, S., Campbell, C.J.V., Diamandis, P., Pantanowitz, L., Kashani, H., Ghodsi, A., Tizhoosh, H.R., 2021. Fine-tuning and training of densenet for histopathology image representation using TCGA diagnostic slides. Med. Image Anal. 70, 1–11.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. Springer, pp. 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.

Rymarczyk, D., Borowa, A., Tabor, J., Zielinski, B., 2021. Kernel self-attention for weakly-supervised image classification using deep multiple instance learning. In: WACV. pp. 1721–1730.

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

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al., 2021. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In: NeurIPS, Vol. 34.

Sharma, Y., Shrivastava, A., Ehsan, L., Moskaluk, C.A., Syed, S., Brown, D., 2021. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In: MIDL. PMLR, pp. 682–698.

Srinidhi, C.L., Ciga, O., Martel, A.L., 2021. Deep neural network models for computational histopathology: A survey. Med. Image Anal. 67, 101813.

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

Talo, M., 2019. Automated classification of histopathology images using transfer learning. Artif. Intell. Med. 101, 101743.

Tao, C., Wang, H., Zhu, X., Dong, J., Song, S., Huang, G., Dai, J., 2021. Exploring the equivalence of siamese self-supervised learning via a unified gradient framework. arXiv preprint arXiv:2112.05141.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., Van Der Laak, J., 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med. Image Anal. 58, 101544.

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

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., Yang, W., Han, X., 2021b. TransPath: Transformer-based self-supervised learning for histopathological image classification. In: MICCAI. Springer, pp. 186–195.

Weaver, D.L., 2010. Pathology evaluation of sentinel lymph nodes in breast cancer: protocol recommendations and rationale. Modern Pathol. 23 (2), S26–S32.

Xiao, T., Dollar, P., Singh, M., Mintun, E., Darrell, T., Girshick, R., 2021. Early convolutions help transformers see better. In: NeurIPS, Vol. 34.

Xie, X., Chen, J., Li, Y., Shen, L., Ma, K., Zheng, Y., 2020. Instance-aware self-supervised learning for nuclei segmentation. In: MICCAI. Springer, pp. 341–350.

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.

Yang, P., Hong, Z., Yin, X., Zhu, C., Jiang, R., 2021. Self-supervised visual representation learning for histopathological images. In: MICCAI. Springer, pp. 47–57.

Yèche, H., Dresdner, G., Locatello, F., Hüser, M., Rätsch, G., 2021. Neighborhood contrastive learning applied to online patient monitoring. In: ICML. PMLR, pp. 11964–11974.

Yu, K.H., Beam, A.L., Kohane, I.S., 2018. Artificial intelligence in healthcare. Nature Biomed. Eng. 2 (10), 719–731.

Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., et al., 2019. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat. Mach. Intell. 1 (5), 236–245.

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