SRA: A Novel Method to Improve Feature Embedding in Self-supervised Learning for Histopathological Images

Hamid Manoochehri^{*1,2} Bodong Zhang^{*†1,2} Beatrice S. Knudsen³ Tolga Tasdizen^{1,2}

¹Department of Electrical and Computer Engineering, University of Utah, Salt Lake City, UT, USA

²Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT, USA

³Department of Pathology, University of Utah, Salt Lake City, UT, USA

 $\verb+hamid.manoochehriQutah.edu, \verb+bodong.zhangQutah.edu, bdong.zhangQutah.edu, bdong.zhangQ$

beatrice.knudsen@pathology.utah.edu, tolga.tasdizen@utah.edu

Abstract

Self-supervised learning has become a cornerstone in various areas, particularly histopathological image analysis. Image augmentation plays a crucial role in self-supervised learning, as it generates variations in image samples. However, traditional image augmentation techniques often overlook the unique characteristics of histopathological images. In this paper, we propose a new histopathology-specific image augmentation method called stain reconstruction augmentation (SRA). We integrate our SRA with MoCo v3, a leading model in self-supervised contrastive learning, along with our additional contrastive loss terms, and call the new model SRA-MoCo v3. We demonstrate that our SRA-MoCo v3 always outperforms the standard MoCo v3 across various downstream tasks and achieves comparable or superior performance to other foundation models pre-trained on significantly larger histopathology datasets.

1 Introduction

Deep learning serves as an invaluable tool for medical diagnosis, including cancer detection and grading. Developing robust models for these purposes is critically important and has gained widespread popularity. Possessing such accurate classification models allows medical doctors to expedite cancer diagnoses and make appropriate prognostic decisions.

Self-supervised learning (SSL) [20, 15] has become a mainstream technique in deep learning. One of its key advantages is that it eliminates the need for labeled data during pre-training by generating labels directly from the data itself. Another important advantage of selfsupervised learning is its capacity to enable transfer learning through pre-trained models. By leveraging largescale datasets and extensive computational resources for pre-training, self-supervised learning models capture rich and generalizable feature representations. Once trained, these models can be shared and deployed by researchers or practitioners who lack access to large datasets or significant computational power. Contrastive learning [19] has become particularly successful in self-supervised pretraining, as it aims to maximize the similarity between different augmentations of the same data points (positive pairs) and minimize the similarity between views of different data points (negative pairs). This encourages the model to learn robust and discriminative feature representations, which can perform well across a variety of downstream tasks.

Data augmentation is crucial for self-supervised contrastive learning, as it introduces variability for each data point and helps the model generalize better to unseen data [32]. Standard image augmentation techniques apply general color, morphological, or geometric transformations, or a combination of these, such as rotation, cropping, resizing, and color jittering. However, these methods do not account for the specific characteristics of histopathological images [31].

Hematoxylin & Eosin (H&E) stained histopathological image is the dominant image type in histopathology. Hematoxylin stains the cell nuclei a deep blue or purple, while Eosin stains the cytoplasm and extracellular matrix pink, allowing for clear differentiation between different cellular and tissue components. The variations of stains across images are shown in fig. 1.

In this paper, we propose a novel histopathologyspecific image augmentation method called stain reconstruction augmentation (SRA) and apply it to H&E stained histopathological images. We first performed stain separation [27] to decompose RGB images into Hematoxylin stain channel images and Eosin stain channel images in Optical Density (OD) space. Instead of multiplying a preset random scalar to each channel for

^{*}The first two authors contributed equally to this work.

[†]Corresponding author. Email: bodong.zhang@utah.edu



Figure 1: Patch examples from various datasets.

augmentation [29, 30, 7], where the augmented images are statistically affected by the intensities of original stain channel images, we first normalize each single stain channel image by dividing it by its max intensity, then multiplying the random scalar we predefined to augment each stain channel image. By this way, the intensities of augmented stain channel images are not affected by the max intensity of original stain channel images. Additionally, when we reconstruct augmented images from OD space back to RGB space, we define a probability for excluding one stain channel and only allow the other channel to remain to provide stronger augmentations. We integrate our SRA into MoCo v3 [11], one of the state-of-the-art self-supervised contrastive learning models, as well as our additional contrastive loss terms, to form our new model SRA-MoCo v3.

The main contributions of our paper are as below:

- We propose a novel histopathological image augmentation method called stain reconstruction augmentation (SRA) and integrate it into MoCo v3 for better self-supervised contrastive learning on digital histopathological images.
- Original MoCo v3's loss function only calculates contrastive loss between feature sets obtained from different augmentations by different encoders (momentum and query encoders). To further stabilize the learned features after SRA, we introduce additional loss terms calculating the contrastive loss between features obtained from different augmentations of the images by the same encoder.
- We conducted experiments on multiple publicly available datasets. We demonstrate that our proposed SRA-MoCo v3 always outperforms MoCo v3 on various downstream tasks. We also achieved comparable or superior performance to other foundation models pre-trained on significantly larger histopathology datasets.
- We made the code publicly available at github.com/hamidmanoochehri/Paper_SRA

2 Related works

2.1 Pathology-specific augmentation

In the context of pathology-specific augmentations, various methods have been proposed to address domainspecific challenges and variations in histopathological images. Shen et al. [28] introduced RandStainNA, which generates random template slides for color normalization and augmentation in HSV, LAB, and HED color spaces to tackle variations in staining and colors across different slides and datasets. Additionally, Gullapally et al. [16] addressed inter-laboratory and scanner variability through Scanner Transform (ST) and Stain Vector Augmentation (SVA), enhancing out-of-distribution performance on tasks such as tissue segmentation.

A fundamental operation for many pathology-specific augmentations is stain separation, which isolates singlechannel images in Optical Density (OD) space based on the Beer-Lambert law [22, 3]. Augmentations can then be applied independently to each stain channel. In [34], perturbations are applied to the stain separation matrix to deal with the errors in separation matrix calculation. In [29, 30], each channel is randomly scaled and biased within a narrow range before converting back to RGB space. However, the maximum possible intensity after augmentation is still influenced by the original image's maximum intensity. [7] also utilizes this method, along with random stain matrix interpolation, to handle domain variations across datasets by incorporating information from both source and target data.

2.2 Self-supervised learning and contrastive loss

Recent advancements in self-supervised learning [35, 14] have introduced novel frameworks for learning robust and accurate features across various datasets. Barlow Twins [35] encourages two augmented views of the same input to produce similar but decorrelated representations by minimizing the cross-correlation between them. DINO [5, 25, 12] employs a student-teacher framework within a vision transformer architecture, without requiring labeled data. In DINO, the student encoder attempts to mimic the teacher encoder, which is updated based on an exponential moving average (EMA). Unlike the student encoder, which processes both global and local views of the images, the teacher encoder only receives global views. Building on the DINO framework, PathDino [2] combines lightweight transformers with a novel 360° rotation augmentation (HistoRotate), achieving robust performance across 12 diverse pathology datasets.

Contrastive learning is one of the most widely used and fundamental approaches in self-supervised learning pipelines. For instance, the SimCLR framework [9, 8] utilizes NT-Xent loss on strongly augmented views of images, aiming to minimize the distance between different



Figure 2: The pipeline of our SRA-MoCo v3. We integrate our Stain Reconstruction Augmentation (SRA) as well as additional contrastive loss terms (CL3 and CL4) into MoCo v3.

views of the same image while maximizing the distance between views of different images. In contrast, SwAV [4] employs a cluster-based contrastive learning approach rather than a pairwise one, using a swapped prediction mechanism to encourage the features of the same cluster to be as invariant as possible. The MoCo framework [17, 10, 11] takes a distinct approach to contrastive learning called momentum contrast. In this framework, there are two encoders: a momentum encoder and a query encoder. The momentum encoder is updated as the exponential moving average of the query encoder and maintains a consistent dictionary of negative samples, enabling contrastive learning with a large and diverse set of negatives across both current and previous batches. Other popular contrastive learning methods include iBOT [39], RePre [33] and RECON [26].

3 Methods

3.1 An overview of SRA-MoCo v3

fig. 2 shows the overall workflow of SRA-MoCo v3. First, all pixels within the tissue regions of an H&E whole slide image (WSI) are collected to analyze the max intensity (strength) of each stain in the current WSI, where the intensity is measured on Optical Density (OD) space after stain separation process. To perform stain reconstruction augmentation, we predefine an absolute range for the target strength of each stain and map the real strength of each stain to a random value in this target range. Unlike the approach in [29, 30, 7], which only slightly adjusts the strength of each stain within a relative range by multiplying a random factor between 0.95 and 1.05, our SRA directly defines a much broader absolute range for target strength of each stain channel. For instance, if the target range is set between 0.5 and 2, and the original strength of a particular stain in a WSI is 2, this indicates a deeply stained image. Augmenting this stain channel only makes the new maximum intensity fall between 0.5 and 2, without surpassing the original strength. While the traditional augmentation makes the new maximum intensity fall between 1.9 and 2.1. Our method allows for more extensive and stronger augmentations while ensuring the strength remains within an appropriate range. Moreover, inspired by multi-modal contrastive learning [6, 36], we define specific probabilities for excluding one stain channel, allowing only the other stain channel to remain after augmentation. Examples of images resulting from our stain reconstruction augmentations, including single stain channel images, can be seen in fig. 3.



Figure 3: Examples of augmentations by SRA with different target strengths of H channel and E channel.

After stain reconstruction augmentation, additional general image augmentation methods are applied to introduce further variations. MoCo v3 is used as the backbone for histopathology image representation learning. In MoCo v3, the contrastive loss is computed between queries from the query encoder and keys from the momentum encoder using different augmentations. However, there are no loss terms that specifically focus on con-



Figure 4: Demonstration of stain reconstruction augmentation (SRA). Single stain images are shown in both RGB space and OD space. The augmentations are performed on each stain channel independently. There is a probability of p that only single channel is adopted.

trastive learning between different augmentations from the same encoder, such as the contrastive loss between identical queries with different augmentations. Given the substantial variations introduced by stain reconstruction augmentation, we further explored the addition of contrastive loss terms that focus solely on augmentations. More details are provided in the following subsections.

3.2 Stain reconstruction augmentation

As shown in fig. 4, for an H&E image, we first perform stain separation using the algorithm from [23] to obtain single-stain images. For each whole slide image (WSI), the RGB pixel values are mapped into Optical Density (OD) space (OD_R, OD_G, OD_B) according to the Beer-Lambert law [3, 22], where higher OD values indicate stronger stains. All pixels are mapped to the same OD space. Based on the distribution of these pixels in OD space, three unit vectors, V_H , V_E , and $V_{Residual}$, are derived, which allow for the decomposition of OD values as shown below:

$$(OD_R, OD_G, OD_B) = \alpha V_H + \beta V_E + \gamma V_{Residual}$$
(1)

In each slide, we calculate the values of α and β for each tissue pixel, where α represents the proportion of Hematoxylin stain and β represents the proportion of Eosin stain for each pixel. For each WSI, we define H_{max} (maximum intensity of Hematoxylin stain) as the 99th percentile of all α values in the tissue pixels of the slide and similarly define E_{max} (maximum intensity of Eosin stain) as the 99th percentile of all β values. This slidelevel operation is performed because the tissues in the same slide are stained and stored under identical conditions. Additionally, due to the large number of pixels in a slide, slide-level operation provides stable results.

In the first step of implementing our stain reconstruction augmentation, we predefine global target ranges for new H_{max} and E_{max} after augmentation. For each training image, following the stain separation process, we independently and randomly select coefficients $coef_H$ and $coef_E$ from within the target ranges. We then multiply α by $coef_H/H_{max}$ and β by $coef_E/E_{max}$ to randomly adjust the stain strength. Furthermore, we introduce a hyperparameter p, which defines the probability of randomly setting either $coef_H$ or $coef_E$ to zero during stain reconstruction augmentation, thereby creating additional variations. Finally, after all processes are complete, we reconstruct the images back into RGB space from OD space, based on the new proportions of Hematoxylin and Eosin stains. The code for stain reconstruction augmentation can be found in Supplementary material.

3.3 Contrastive learning loss

We adopted MoCo v3 [11] as the platform for our contrastive learning framework. MoCo v3 consists of two encoders: a momentum encoder and a query encoder. The model aims to match an encoded query q from the query encoder to a dictionary of encoded keys from the momentum encoder. The contrastive loss term between queries and keys can be written as:

$$CL(k,q) = -\log\left(\frac{\exp\left(q \cdot k^{+}/\tau\right)}{\exp\left(q \cdot k^{+}/\tau\right) + \sum_{k^{-}}\exp\left(q \cdot k^{-}/\tau\right)}\right)$$
(2)

where τ is temperature hyperparameter. As shown in fig. 2, each input image is augmented twice, and the augmented images are then passed through both the query encoder and the momentum encoder. Let f_M^{Aug1} represent the output features from the momentum encoder after the first augmentation, f_M^{Aug2} the output after the second augmentation, f_B^{Aug1} the output from the query encoder after the first augmentation, and f_B^{Aug2} the output after the second augmentation. The original contrastive loss can be written as:



Figure 5: Patch examples from different classes and different datasets.

Set	Normal/ Benign	Cancer	Necrosis	Total	
	1,37	3,684 Unlab	oeled	1 0 40 005	
Train	84,578	$180,\!471$	7,932	1,646,665	
Val.	19,638	79,382	1,301	100,321	
Test	15,323	62,565	$6,\!168$	84,056	

Table 1: Summary of the number of patches for each category in each set on TCGA KIRC dataset. The training set includes both labeled and unlabeled patches.

$$CL_{ori} = CL_1(f_M^{Aug1}, f_B^{Aug2}) + CL_2(f_M^{Aug2}, f_B^{Aug1})$$
 (3)

In the standard contrastive loss pairs, both the augmentations and the encoders differ. To make the contrastive learning process more sensitive to augmentations, especially given the introduction of strong stain augmentation, we introduce the following additional loss terms:

$$CL_{aug} = CL_3(f_B^{Aug1}, f_B^{Aug2}) + CL_4(f_M^{Aug1}, f_M^{Aug2})$$
 (4)

Thus, the overall contrastive learning loss becomes $CL_{ori} + CL_{aug}$. We will evaluate the benefits of incorporating these extra loss terms through ablation studies in the Experiments section.

4 Experiments

4.1 Datasets

We conducted all our experiments on three datasets: The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma (TCGA KIRC) dataset [24], the Utah KIRC dataset [37], and the Camelyon 16 dataset [13], all of which consist of H&E-stained images. The first two datasets contain kidney WSIs, while the Camelyon 16 dataset consists of breast WSIs. The labels for the TCGA

Set	Normal/ Benign	Low-risk Cancer	High-risk Cancer	Necrosis	Total
- ·		171,113 U	Unlabeled		200.201
Train	28,497	2,044	2,522	4,115	208,291
Val.	$5,\!472$	416	334	$2,\!495$	8,117
Test	7,263	598	389	924	$9,\!174$

Table 2: Summary of the number of patches for each category in each set on Utah KIRC dataset. The training set includes both labeled and unlabeled patches.

KIRC and Utah KIRC images were obtained from [37], and the labels for the Camelyon 16 dataset were sourced from [1]. Examples of patches from the three datasets are shown in fig. 5.

From the TCGA KIRC dataset, we collected 420 WSIs in total, with 300 slides used for training, 60 slides for validation, and 60 slides for testing. The dataset provides 1,646,665 tissue patches of size 400x400 at 20X resolution from the 300 training slides for self-supervised contrastive learning, as well as labeled patches for 3-class classification tasks. More details can be found in table 1.

In the Utah KIRC dataset, there are 49 slides from different patients, with 32 slides for training, 10 slides for validation, and 7 slides for testing. The 32 training slides provide 208,291 tissue patches of size 400x400 at 10X resolution for self-supervised contrastive learning. This dataset also includes labeled patches for 4-class classification tasks. For more details, please refer to table 2.

Camelyon 16 is another public dataset consisting of breast cancer slides for multi-instance learning. Instead of patch-level labels, it provides only slide-level labels for binary classification between Normal and Tumor. The training set contains 160 Normal slides and 111 Tumor slides, while the test set includes 80 Normal slides and 50 Tumor slides. We randomly selected 10% of the training slides to create a validation set. We followed the standard procedure as other papers [38] to crop patches in



Figure 6: Distributions of strengths of Hematoxylin stain and Eosin stain in Optical Density (OD) space on TCGA training set and Utah training set.

20X resolution.

4.2 Experiment settings

In our experiments, we began by pre-training the encoders using self-supervised contrastive feature representation learning, followed by testing the pre-trained encoders on downstream tasks. The first scenario involves performing both pre-training and downstream classification tasks on the same dataset. The second scenario involves pre-training the encoders on one dataset and evaluating them on downstream tasks from a different dataset. We implemented the first scenario on the TCGA KIRC and Utah KIRC datasets separately. For the second scenario, we pre-trained the encoders on the TCGA KIRC dataset, then evaluated them on the Utah KIRC and Camelyon 16 separately. Since the patches in the Utah dataset have a resolution of 10X, compared to 20X in the TCGA KIRC dataset, we also cropped patches at 10X resolution from the TCGA KIRC training slides to ensure resolution consistency between pre-training and downstream tasks.

For stain reconstruction augmentation, we first predefined a range for the target strengths of the Hematoxylin and Eosin stains. We calculated the distributions of H_{max} and E_{max} across all training slides in both the TCGA KIRC and Utah KIRC datasets. As shown in fig. 6, the stain strengths (H_{max} and E_{max}) varies across different slides in both datasets. In our experiments, we tested two sets of target ranges. The first set marginally covers the distribution of stain intensities in TCGA KIRC, with a target range of [0.5, 2.0] for new H_{max} and [0.2, 2.0] for new E_{max} . The second set has wider target ranges to introduce stronger augmentations, with both new H_{max} and new E_{max} set to [0.1, 2.5]. We selected this range because further widening would cause saturation when reconstructing images from Optical Density (OD) space back to RGB space.

After applying stain reconstruction augmentation, we followed the standard MoCo v3 training procedure, including additional general augmentations, with ResNet50 [18] as backbone. The encoders consistently received 224x224 patches as inputs except for PathDino, where 512x512 patches were used after image resizing. We also included an option in our code to add the augmentation contrastive loss, CL_{auq} . Moreover, we tracked the contrastive loss on the validation set and selected the training epoch with the lowest contrastive loss on the validation set. In the downstream tasks on the TCGA KIRC and Utah KIRC, we used a balanced sampler for training and balanced accuracy for evaluation, as the patch labels were highly imbalanced. For the downstream tasks on the Camelyon 16 dataset, we employed DTFD-MIL [38], as it has been shown to be one of the best multi-instance learning models for Camelyon 16 and other datasets. Stain reconstruction augmentation was only applied during pretraining, not in any downstream tasks, where the MoCo v3's image classification code was used.

All experiments were conducted using Python 3.11.4, PyTorch 2.0.1, torchvision 0.15.2, and CUDA 11.8 on NVIDIA RTX A6000 GPUs. With a batch size of 512, the MoCo v3 pre-training required a total of around 63 GB of memory and approximately 20 hours of runtime across two parallel GPUs.

4.3 Pre-training and downstream task on same datasets

We first evaluated SRA-MoCo v3 by performing contrastive feature representation learning and downstream classification on the same datasets (TCGA KIRC or Utah

Pre-trained Dataset	Model	Balanced Accuracy (TCGA KIRC)	
ImageNet	ResNet50	69.97 ± 5.59	
TCGA + TULIP 32.6M patches	Barlow Twins MoCo v2 SwAV	81.59 ± 2.65 79.04 ± 0.11 77.43 ± 1.15	
11,765 TCGA Slides 6.1M patches	PathDino	76.92 ± 6.22	
TCGA KIRC 300 Slides 1.6M patches	MoCo v3 MoCo v3 + TSA SRA-MoCo v3	$79.37 \pm 1.18 \\81.50 \pm 0.23 \\83.62 \pm 0.28$	

Table 3: Performance of pre-trained models on TCGA KIRC dataset compared to foundation models (20X magnification). TSA represents traditional stain augmentation methods [29, 30]

KIRC). Since there is no domain shift in this scenario, we consistently used the first option's target range for stain strengths, which has narrower limits (H: [0.5, 2.0], E: [0.2, 2.0]), and set the probability of generating pure Hematoxylin or pure Eosin images to zero. After pretraining, we froze the pre-trained encoders and performed classification on the same dataset. For reference, we also compared the classification results with other state-ofthe-art foundation encoder models, including PathDino pre-trained on 11,765 TCGA slides by [2], as well as Barlow Twins, SwAV, and MoCo v2 pre-trained on 20,994 TCGA slides and 15672 from the TULIP dataset by [21]. Additionally, we included the basic ResNet50 model in a fully supervised learning setting for comparison.

As shown in table 3 and table 4, our SRA-MoCo v3 significantly outperforms standard MoCo v3 and stain augmentations from [29, 30]. On the TCGA KIRC dataset, SRA-MoCo v3 surpasses MoCo v3 by 4.25% (83.62% vs. 79.37%) in balanced accuracy, and SRA also outperforms the traditional stain augmentation (TSA) [29, 30] by 2.12% (83.62% vs. 81.50%). On the Utah KIRC dataset, SRA-MoCo v3 exceeds MoCo v3 by 2.08% (95.85% vs. 93.77%), and MoCo v3 plus TSA by 1.85% (95.85% vs. 94.00%). The comparison with other foundation models further demonstrates the advantages of SRA-MoCo v3. Despite these state-of-the-art models being pre-trained on much larger datasets, SRA-MoCo v3 still reaches the best.

4.4 Pre-training and downstream task on different datasets

Although our SRA-MoCo v3 demonstrates outstanding results as discussed in the previous subsection, the more common application of self-supervised learning is to pretrain an encoder on large datasets and then apply the

Pre-trained Dataset	Model	Balanced Accuracy (Utah KIRC)	
ImageNet	ResNet50	87.76 ± 0.10	
TCGA + TULIP 32.6M patches	Barlow Twins MoCo v2 SwAV	90.23 ± 1.81 91.45 ± 0.44 94.96 ± 1.04	
11,765 TCGA Slides 6.1M patches	PathDino	92.14 ± 1.65	
Utah KIRC 49 Slides 0.2M patches	MoCo v3 MoCo v3 + TSA SRA-MoCo v3	$93.77 \pm 0.86 \\94.00 \pm 0.26 \\95.85 \pm 0.34$	

Table 4: Performance of pre-trained models on Utah KIRC dataset compared to foundation models (10X magnification). TSA represents traditional stain augmentation methods [29, 30].

frozen encoder to other datasets. To evaluate SRA-MoCo v3 in a more practical setting, we pre-trained SRA-MoCo v3 on 300 training slides from the TCGA KIRC dataset and subsequently evaluated the encoders on the Utah KIRC dataset and the Camelyon 16 dataset separately.

Considering the presence of domain shift, we used the second option's wider target range for stain strength in the stain reconstruction augmentation (H: [0.1, 2.5], E: [0.1, 2.5]) and set the probability of generating pure Hematoxylin or pure Eosin images to 0.1. We also compared the classification results with other state-of-the-art foundation models pre-trained on large datasets, including PathDino, Barlow Twins, SwAV, and MoCo v2. For the downstream multi-instance learning on the Camelyon 16 dataset, we also used the ImageNet pre-trained ResNet50 as the encoder within the DTFD-MIL framework as a baseline.

Based on table 5, on the Utah KIRC dataset, we observed 2.8% improvement with SRA-MoCo v3 compared to MoCo v3. Both SRA-MoCo v3 and MoCo v3 pretrained on the TCGA KIRC dataset outperform those pre-trained on the Utah KIRC dataset itself. The most likely reason is that the TCGA KIRC dataset contains significantly more patches. Even though domain shift is present, both datasets focus on kidney cancer, which mitigates the impact. When comparing the results with foundation models, it becomes evident that pre-training and downstream classification both on kidney images offer distinct advantages.

The multi-instance learning results on the Camelyon 16 dataset can also be found in table 5. The MoCo v3 encoder pre-trained on the TCGA KIRC dataset shows lower performance compared to encoders pre-trained on ImageNet. One possible explanation is the substantial domain shift between kidney cancer and breast cancer slides. This is further illustrated in fig. 5, which highlights the significant differences between TCGA kidney

Pre-trained Dataset	Model	Balanced Accuracy Utah KIRC	F1-score Camelyon16	Accuracy Camelyon16	Balanced Accuracy Camelyon16	
ImageNet	ResNet50	87.76 ± 0.10	0.8372 ± 0.0149	88.37 ± 0.78	86.54 ± 1.35	
	Barlow Twins	90.23 ± 1.81	0.9019 ± 0.0100	93.02 ± 0.78	91.35 ± 0.67	
22 6M patabas	MoCo v2	91.45 ± 0.44	0.9291 ± 0.0049	94.83 ± 0.45	93.73 ± 0.13	
52.0M patches	SwAV	94.96 ± 1.04	0.9264 ± 0.0098	94.57 ± 0.78	93.65 ± 0.63	
11,765 TCGA Slides	PathDino	02.14 ± 1.65	0.0176 ± 0.0252	03.80 ± 2.05	03.15 ± 1.74	
6.1M patches	1 atmbillo	52.14 ± 1.00	0.3170 ± 0.0252	35.00 ± 2.05	95.10 ± 1.14	
TCGA KIRC	MoCo v3	95.32 ± 0.30	0.8075 ± 0.0135	85.79 ± 1.79	84.33 ± 0.89	
300 Slides	MoCo v $3 + TSA$	94.17 ± 0.82	0.8268 ± 0.0163	87.60 ± 2.05	85.65 ± 0.94	
0.4M/1.6M patches	SRA-MoCo v 3	98.12 ± 0.15	0.9207 ± 0.0084	94.31 ± 0.44	92.91 ± 0.95	

Table 5: Performance of pre-trained models on Utah KIRC dataset and Camelyon16 dataset, compared to foundation models. TSA represents traditional stain augmentation methods [29, 30]. In 300 slides from TCGA KIRC, 363,225 10X patches are used in pre-training for downstream task on Utah KIRC, 1,646,665 20X patches are used in pre-training for downstream task on Camelyon 16.

Range	Range	p(only	Extra	Balanced Acc.	F1-score	Accuracy	Balanced Acc.
$coef_H$	$coef_E$	H or E)	Loss	Utah KIRC	Camelyon16	Camelyon16	Camelyon16
N/A	N/A	0	-	95.32 ± 0.30	0.8075 ± 0.0135	85.79 ± 1.79	84.33 ± 0.89
N/A	N/A	0	CL_{aug}	95.34 ± 0.41	0.8457 ± 0.0253	88.11 ± 1.19	87.78 ± 2.44
[0.2, 2.0]	[0.5, 2.0]	0	_	96.51 ± 0.37	0.8244 ± 0.0047	87.34 ± 0.45	85.58 ± 0.42
[0.1, 2.5]	[0.1, 2.5]	0	_	96.95 ± 0.76	0.8494 ± 0.0154	89.67 ± 0.89	87.19 ± 1.25
[0.2, 2.0]	[0.5, 2.0]	0	CL_{aug}	96.86 ± 0.17	0.8341 ± 0.0060	88.37 ± 0.78	86.15 ± 0.33
[0.1, 2.5]	[0.1, 2.5]	0	CL_{aug}	98.09 ± 0.12	0.8596 ± 0.0213	89.15 ± 2.05	88.75 ± 1.40
[0.2, 2.0]	[0.5, 2.0]	10%	CL_{aug}	97.41 ± 0.08	0.9079 ± 0.0150	93.28 ± 1.18	92.08 ± 1.02
[0.1, 2.5]	[0.1, 2.5]	10%	CL_{aug}	98.12 ± 0.15	0.9207 ± 0.0084	94.31 ± 0.44	92.91 ± 0.95

Table 6: Ablation study results showing the impact of each component of SRA-MoCo v3 on Utah KIRC dataset and Camelyon 16 dataset.

cancer patches and Camelyon 16 breast cancer patches. However, with our stain reconstruction augmentation and our additional augmentation contrastive loss, SRA-MoCo v3 improves the balanced accuracy by 8.53% compared to MoCo v3, while MoCo v3 + TSA exceeds MoCo v3 by only 1.82%. When comparing results with other foundation models pre-trained on much larger datasets with a greater variety of organ types, SRA-MoCo v3 surpasses Barlow Twins and achieves comparable F1 score, accuracy, and balanced accuracy to PathDino, SwAV, and MoCo v2. Our pre-trained SRA-MoCo v3 encoder will be released upon publication.

4.5 Ablation studies

We also conducted ablation studies to carefully analyze the impact of each component in SRA-MoCo v3.

The ablation studies for transfer learning from the TCGA KIRC dataset to the Utah KIRC dataset are presented in table 6. From the results, we found that simply adding the augmentation contrastive loss to MoCo v3 does not yield any improvement. However, this loss becomes effective when combined with stain reconstruction augmentation. Using a wider target range for stain strength and incorporating the possibility of generating pure Hematoxylin or pure Eosin images in stain reconstruction augmentation also proved beneficial. During transfer learning, the slides exhibit significantly more variation across datasets, and SRA-MoCo v3 addresses this by generating highly augmented images that are also clinically meaningful.

Lastly, we evaluated the contribution of each component in SRA-MoCo v3 through downstream tasks on the Camelyon 16 dataset. We observed that both the augmentation contrastive loss and stain reconstruction augmentation independently improve performance. When combining the augmentation contrastive loss with stain reconstruction augmentation using a wider target range of stain variation, we achieved a 4.43% improvement in balanced accuracy. Unlike transfer learning from the TCGA KIRC to the Utah KIRC dataset, where setting the probability of converting augmented images to pure Hematoxylin or pure Eosin to 0.1 resulted in a minor boost, we observed a 4.16% increase in balanced accuracy on Camelyon 16 by simply adjusting this probability from 0 to 0.1, which implies that stronger augmentations are more beneficial on cases with stronger domain shift. All the adjustments contributed to SRA-MoCo v3 outperforming MoCo v3 by 8.58%.

5 Conclusion

In this paper, we propose a novel framework for contrastive representation learning in pathology images. Specifically, we introduce a new stain reconstruction augmentation method and an augmentation contrastive loss. Experiments on downstream tasks, including image classification and multi-instance learning, demonstrate significant improvements when applying our algorithms to MoCo v3. Despite being pre-trained on much smaller datasets, SRA-MoCo v3 achieves comparable or superior performance in various self-supervised learning and downstream tasks compared to other state-of-the-art foundational encoders.

Acknowledgements

The work was funded in part by NIH 1R21CA277381. The Computational Oncology Research Initiative (CORI) at the Huntsman Cancer Institute, The Department of Pathology at the University of Utah and ARUP Laboratories had also supported the work. Part of the results presented in this paper are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga

References

- [1] CAMELYON16 Grand Challenge. https: //camelyon16.grand-challenge.org/Data/. Accessed: 2023-12-15.
- [2] Saghir Alfasly, Abubakr Shafique, Peyman Nejat, Jibran Khan, Areej Alsaafin, Ghazal Alabtah, and Hamid R Tizhoosh. Rotation-agnostic image representation learning for digital pathology. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11683–11693, 2024.
- [3] Beer. Bestimmung der absorption des rothen lichts in farbigen flüssigkeiten. Ann. Phys., 162(5):78–88, Jan. 1852.
- [4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Oct. 2021.
- [6] Wenhao Chai and Gaoang Wang. Deep vision multimodal learning: Methodology, benchmark, and trend. Appl. Sci. (Basel), 12(13):6588, June 2022.
- [7] Jia-Ren Chang, Min-Sheng Wu, Wei-Hsiang Yu, Chi-Chung Chen, Cheng-Kung Yang, Yen-Yu Lin, and Chao-Yuan Yeh. Stain mix-up: Unsupervised domain generalization for histopathology images. In *Medical Image Computing and Computer Assisted Intervention – MIC-CAI 2021*, Lecture notes in computer science, pages 117– 126. Springer International Publishing, Cham, 2021.

- [8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé Iii and Aarti Singh, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 1597–1607. PMLR, 2020.
- [9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems, 33:22243–22255, 2020.
- [10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv [cs.CV], Mar. 2020.
- [11] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Oct. 2021.
- [12] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024.
- Babak Ehteshami Bejnordi, Mitko Veta, Paul Jo-[13]hannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A W M van der Laak, the CAMELYON16 Consortium, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, Oscar Geessink, Nikolaos Stathonikos, Marcory Crf van Dijk, Peter Bult, Francisco Beca, Andrew H Beck, Davong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Aoxiao Zhong, Qi Dou, Quanzheng Li, Hao Chen, Huang-Jing Lin, Pheng-Ann Heng, Christian Haß, Elia Bruni, Quincy Wong, Ugur Halici, Mustafa Umit Oner, Rengul Cetin-Atalay, Matt Berseth, Vitali Khvatkov, Alexei Vylegzhanin, Oren Kraus, Muhammad Shaban, Nasir Rajpoot, Ruqayya Awan, Korsuk Sirinukunwattana, Talha Qaiser, Yee-Wah Tsang, David Tellez, Jonas Annuscheit, Peter Hufnagl, Mira Valkonen, Kimmo Kartasalo, Leena Latonen, Pekka Ruusuvuori, Kaisa Liimatainen, Shadi Albarqouni, Bharti Mungal, Ami George, Stefanie Demirci, Nassir Navab, Seiryo Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Hady Ahmady Phoulady, Vassili Kovalev, Alexander Kalinovsky, Vitali Liauchuk, Gloria Bueno, M Milagro Fernandez-Carrobles, Ismael Serrano, Oscar Deniz, Daniel Racoceanu, and Rui Venâncio. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA, 318(22):2199-2210, Dec. 2017.
- [14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.
- [15] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends.

IEEE Trans. Pattern Anal. Mach. Intell., PP(99):1–20, June 2024.

- [16] Sai Chowdary Gullapally, Yibo Zhang, Nitin Kumar Mittal, Deeksha Kartik, Sandhya Srinivasan, Kevin Rose, Daniel Shenker, Dinkar Juyal, Harshith Padigela, Raymond Biju, Victor Minden, Chirag Maheshwari, Marc Thibault, Zvi Goldstein, Luke Novak, Nidhi Chandra, Justin Lee, Aaditya Prakash, Chintan Shah, John Abel, Darren Fahy, Amaro Taylor-Weiner, and Anand Sampat. Synthetic DOmain-targeted augmentation (S-DOTA) improves model generalization in digital pathology. arXiv [cs. CV], May 2023.
- [17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. arXiv [cs.CV], Nov. 2019.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. arXiv [cs. CV], pages 770–778, Dec. 2015.
- [19] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies* (*Basel*), 9(1):2, Dec. 2020.
- [20] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(11):4037– 4058, Nov. 2021.
- [21] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3344–3354. IEEE, June 2023.
- [22] J H Lambert. Photometria sive de mensura et gradibus luminis, colorum et umbrae. sumptibus vidvae E. Klett, typis C.P. Detleffsen, 1760.
- [23] Marc Macenko, Marc Niethammer, J S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE, June 2009.
- [24] National Cancer Institute. National cancer institute GDC data portal. https://portal.gdc.cancer.gov/. Accessed: 2023-11-2.
- [25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research, 2024.
- [26] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: contrastive 3d representation learning guided by generative pretraining. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.
- [27] A C Ruifrok and D A Johnston. Quantification of histochemical staining by color deconvolution. Anal. Quant. Cytol. Histol., 23(4):291–299, Aug. 2001.

- [28] Yiqing Shen, Yulin Luo, Dinggang Shen, and Jing Ke. RandStainNA: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 212–221. Springer Nature Switzerland, Cham, 2022.
- [29] David Tellez, Maschenka Balkenhol, Nico Karssemeijer, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. H and E stain augmentation improves generalization of convolutional networks for histopathological mitosis detection. In John E. Tomaszewski and Metin N. Gurcan, editors, *Medical Imaging 2018: Digital Pathology*, volume 10581, page 105810Z. International Society for Optics and Photonics, SPIE, 2018.
- [30] David Tellez, Maschenka Balkenhol, Irene Otte-Holler, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Trans. Med. Imaging*, 37(9):2126–2136, Sept. 2018.
- [31] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.*, 58(101544):101544, Dec. 2019.
- [32] Diane Wagner, Fabio Ferreira, Danny Stoll, Robin Tibor Schirrmeister, Samuel Müller, and Frank Hutter. On the importance of hyperparameters and data augmentation for self-supervised learning. arXiv [cs.LG], July 2022.
- [33] Luyang Wang, Feng Liang, Yangguang Li, Honggang Zhang, Wanli Ouyang, and Jing Shao. Repre: Improving self-supervised vision transformer with reconstructive pre-training. In *International Joint Conference on Artificial Intelligence*, 2022.
- [34] Pengshuai Yang, Xiaoxu Yin, Haiming Lu, Zhongliang Hu, Xuegong Zhang, Rui Jiang, and Hairong Lv. CS-CO: A hybrid self-supervised visual representation learning method for H&E-stained histopathological images. *Med. Image Anal.*, 81(102539):102539, Oct. 2022.
- [35] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. ArXiv, abs/2103.03230, 2021.
- [36] Bodong Zhang, Beatrice Knudsen, Deepika Sirohi, Alessandro Ferrero, and Tolga Tasdizen. Stain based contrastive co-training for histopathological image analysis. In *Medical Image Learning with Limited and Noisy Data*, pages 106–116. Springer Nature Switzerland, 2022.
- [37] Bodong Zhang, Hamid Manoochehri, Man Minh Ho, Fahimeh Fooladgar, Yosep Chong, Beatrice S Knudsen, Deepika Sirohi, and Tolga Tasdizen. CLASS-M: Adaptive stain separation-based contrastive learning with pseudolabeling for histopathological image classification. arXiv [cs. CV], Dec. 2023.
- [38] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In 2022 IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR), pages 18780–18790, 2022.

[39] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pretraining with online tokenizer. *International Conference* on Learning Representations (ICLR), 2022.