CLASS-M: Adaptive stain separation-based contrastive learning with pseudo-labeling for histopathological image classification

Bodong Zhang[1,2[0000−0001−9815−0303]], Hamid Manoochehri[1,2[0009−0005−0478−7925]], Man Minh Ho[2[0000−0002−7178−8873]], Fahimeh Fooladgar[3[0000−0002−7859−0456]], Yosep Chong[4[0000−0001−8615−3064]], Beatrice S. Knudsen[5[0000−0002−7589−7591]], Deepika Sirohi[5[0000−0002−0848−4172]], and Tolga Tasdizen[1,2[0000−0001−6574−0366]]

[1] Department of Electrical and Computer Engineering, University of Utah, Salt Lake City, UT, USA
[2] Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT, USA
[3] Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada
[4] Department of Hospital Pathology, College of Medicine, The Catholic University of Korea, Seoul, South Korea
[5] Department of Pathology, University of Utah, Salt Lake City, UT, USA
bodong.zhang@utah.edu

Abstract. Histopathological image classification is one of the critical aspects in medical image analysis. Due to the high expense associated with the labeled data in model training, semi-supervised learning methods have been proposed to alleviate the need of extensively labeled datasets. In this work, we propose a model for semi-supervised classification tasks on digital histopathological Hematoxylin and Eosin (H&E) images. We call the new model Contrastive Learning with Adaptive Stain Separation and MixUp (CLASS-M). Our model is formed by two main parts: contrastive learning between adaptively stain separated Hematoxylin images and Eosin images, and pseudo-labeling using MixUp. We compare our model with other state-of-the-art models on clear cell renal cell carcinoma (ccRCC) datasets from our institution and The Cancer Genome Atlas Program (TCGA). We demonstrate that our CLASS-M model has the best performance on both datasets. The contributions of different parts in our model are also analyzed.

Keywords: Contrastive learning· Stain separation· pseudo-labeling· MixUp· Digital histopathological images· Semi-supervised learning

1 Introduction

Digital histopathological image analysis plays a crucial role in disease diagnosis and treatment optimization. The most commonly used histopathological images are Hematoxylin & Eosin (H&E) stained whole slide images (WSIs) that help the differentiation of various tissue sample features [15]. H&E staining involves the use of two dyes, Hematoxylin and Eosin, which selectively stain different components of the tissue sample. Hematoxylin stains the acidic components of the tissue sample, such as the cell nuclei, and has a blue-purple color, while Eosin stains the basic components of the tissue sample, such as the cytoplasm and extracellular matrix, and exhibits a pink color. In recent years, many efforts have been made to the automatic analysis of images.

With the rapid development of computing power and advanced algorithms, deep learning has been widely used in the field of digital histopathological image analysis for automating disease diagnosis and performing auxiliary image analysis [8,36,43,45]. However, achieving a highly accurate model requires a substantial amount of training labels, demanding significant time and effort from human experts. Various strategies have been proposed to mitigate this challenge. Weakly-supervised learning become the accepted method in histopathology because of the easy availability of case-level labels from medical reports. Self-supervised learning is also commonly used to train the backbone. Nevertheless, semi-supervised learning remains a relatively less explored area in histopathology.

One area worthy of further investigation with deep learning models is kidney cancer. Kidney cancer ranks among the most prevalent cancers globally. It is estimated that around 76,080 new cases were diagnosed with cancers of the kidney and renal pelvis and there are around 13,780 deaths resulted from it in the US in 2021 [38]. Clear cell renal cell carcinoma (ccRCC) stands out as the predominant subtype that dominates the kidney cancer cases [4]. Consequently, further research on cancer detection and classification using ccRCC images holds paramount importance for disease diagnosis and early patient treatment. In this paper, we contribute new annotated Utah ccRCC dataset and The Cancer Genome Atlas Program (TCGA) ccRCC dataset. The Utah ccRCC dataset consists of 49 WSIs, while TCGA ccRCC dataset comprises 420 WSIs. Even though multi-instance learning, requiring only slide-level annotations, is widely used in the histopathology field [32,35,42,50], it requires a larger number of WSIs for effective training. Moreover, slide-level multi-instance learning has some prerequisites, like certain classes should have both positive and negative cases, which can't always be fulfilled in our ccRCC datasets. Therefore, we decided to use polygons to mark regions in a subset of WSIs and assign labels to those polygons. The patches can be cropped inside those annotated polygons to collect labeled samples for performing patch-level classification tasks. To fully utilize the WSIs, we collect the patches outside annotated polygons and from WSIs

that are not annotated at all to gather unlabeled samples. Consequently, we form a semi-supervised classification task based on the labeled and unlabeled patches. Notably, a patch-level classification model serves as a precise tool to expedite pathologists' identification of cancerous regions within WSIs.

In our work, we propose a new semi-supervised model for histopathological image classification: Contrastive Learning with Adaptive Stain Separation and MixUp (CLASS-M), where CLASS-M can also be understood as CLASSifying Medical images. The main portions of our model are contrastive learning and pseudo-labeling. A simple but effective margin loss is built for contrastive learning between adaptive stain separated Hematoxylin images and Eosin images. The pseudo-labeling with MixUp is also applied to provide more samples and pseudo-labels in classification training. The two methods focus on different aspects to fully utilize the information in both labeled and unlabeled samples. Our new model is an extension of [54] presented in MICCAI 2022 workshop with substantial modifications and great improvements on classification results. Instead of using a globally fixed stain separation matrix in Optical Density space, a simple and efficient slide-by-slide stain separation from [37] is applied to utilize domain knowledge of color patterns across slides. We not only perform image augmentations on Hematoxylin images and Eosin images, but also apply augmentations on original RGB images before stain separation process to strengthen the augmentation process. Additionally, considering the benefit of pseudo-labeling in famous semi-supervised learning methods [47] [3], pseudo-labeling with MixUp [56] is adopted to mix samples and then create mixed one-hot encoding labels. In this paper, we evaluated multiple state-of-the-art semi-supervised and self-supervised learning models on Utah ccRCC dataset and TCGA ccRCC dataset. Our CLASS-M model demonstrates the best classification results. We further did ablation studies to carefully analyze the contributions of different parts in our model.

In conclusion, the main contributions of our paper are as follows:

- We proposed our new CLASS-M model for semi-supervised learning. We applied contrastive learning on Hematoxylin images and Eosin images after performing slide-level stain separation, and pseudo-labeling with MixUp to generate more mixed samples and pseudo-labels.
- We provide new Utah ccRCC dataset from our institution and TCGA ccRCC dataset that contain tile-level labels on 49 WSIs and 150 WSIs separately, with extra 270 unlabeled WSIs from TCGA. The Utah ccRCC dataset and its annotations are publicly available upon transfer agreement. The annotations on TCGA ccRCC dataset are publicly available.
- We tested various state-of-the-art semi-supervised learning and self-supervised learning methods on ccRCC datasets and benchmarked their performances. The results of CLASS-M always have the best performance among state-of-the-art models.

2 Related work

2.1 Weakly-supervised learning

Weakly-supervised learning [28,39,58] leverages weak labels to help guide training process. A typical case of weakly-supervised learning is multiple instance learning (MIL) [16,24]. Various MIL approaches, such as instance-based and embedding-based methods, have been proposed. The instance-based approach [25,13] involves training a classifier using bag labels at the instance level, then aggregating the predicted instance labels to form bag label predictions. However, due to the noise in instance labels, the performance of this aggregation may be impacted. On the other hand, the embedding-based approach [19,34] first creates a bag representation from individual instance representations, followed by training a classifier on these bag representations. Research has shown that embedding-based approaches are generally more effective than instance-based methods. The attention-based MIL method, proposed in [26], suggests that instead of treating all patches within a bag equally, assigning varying importance scores to patches, particularly the more discriminative ones, is more effective. This method involves computing attention scores from instance representations to reflect the importance of respective patches and forming the bag representation through a weighted average of these instances. [51] proposed a hierarchical attention-guided MIL that effectively identifies important areas at different scales in WSIs. This method combines several attention techniques to form a comprehensive group representation. Furthermore, the recent Self-ViT-MIL approach by [18], which combines self-supervised learning with Vision Transformers (ViTs) [14] and MIL, has shown promising results, even in comparison to fully-supervised methods.

In digital histopathological image applications, MIL can be applied when only whole slide image-level labels are provided, but the specific regions that contribute to the labels are not given. Despite the convenience of annotation work, weakly-supervised learning can't deal with small datasets so well, as it needs adequate amount of slides to have enough labels. Additionally, weakly-supervised learning struggles with multiple classes which almost always co-exist in slides or a certain class that lacks positive or negative labels in datasets. For example, in our Utah ccRCC dataset, cancer is consistently present in all WSIs, resulting in a lack of negative labels for the cancer category in the context of MIL.

2.2 Self-supervised learning

Self-supervised learning (SSL) [27,30] learns useful patterns from data itself without explicit labels provided by humans. Self-supervised learning allows models to first pre-train on a large unlabeled dataset, where effective feature representations can be learned. Examples include image inpainting, predicting rotations, or colorizing images. Then the final layers of the models can be optimized in specific downstream tasks, reducing the need of extensively labeled data. SimCLR [9,10]

presents a seminal contrastive learning to train feature representation, where augmented views of an image (positive samples) are minimized within that image while being maximized against augmented views of other images (negative samples). MoCo [11,12,21,22,20] strengthens SSL with a dynamic dictionary, tapping into past batches for more negative samples, and a slowly updating momentum encoder for stable global representations. In contrast, Barlow Twins [53] omits negatives, focusing on refining features by measuring cross-correlation between positive samples. BYOL [17] and DINO [7] utilize a teacher-student setup, where the student refines its understanding by predicting the teacher's representation, while teacher network is updated through the slow-moving average of student's parameters. Clustering-based methods like Deep Clustering [5] and SwAV [6] define pseudo-labels, sorting images into clusters, yielding great performance on downstream tasks. Mingu et al. [29] benchmarked SSL methods on pathology datasets, showcasing consistent enhancements in histopathology tasks. In CS-CO [52], the encoders for Hematoxylin images and Eosin images are also adopted. Instead of contrastive loss between Hematoxylin images and Eosin images introduced in our work, CS-CO calculates contrastive loss between different augmentations. In addition, CS-CO is a self-supervised learning method that requires encoders to generate a visual representation containing enough information to recover images in cross-stain prediction, while our semi-supervised CLASS-M only pursues shared latent features containing enough information to perform classification tasks. Self-supervised learning is especially helpful when there is a lack of labeled data for training. While self-supervised learning benefits from unlabled data, semi-supervised learning has the advantages that allows both unlabeled data and labeled data to be trained at the same time and use the knowledge learned from labeled data to better utilize unlabeled data.

2.3 Semi-supervised learning

Semi-supervised learning [40,48] involves training a model with both labeled and unlabeled data to enhance training outcomes, which reduces reliance on a limited number of annotated samples. Semi-supervised learning has superiority especially when obtaining fully labeled data is prohibitively expensive or impractical. In the realm of digital histopathological image analysis, getting a large amount of labeled data requires great effort from well-trained experts, which is expensive and time consuming. However, it is much easier to acquire unlabeled data such as unannotated WSIs. One type of semi-supervised learning is consistency regularization which seeks agreement among the same input with different views, augmentations or epochs. For example, Temporal ensembling [31] aims to reach consensus in prediction of labels between current epochs and previous epochs. [44] tries to make classifier's prediction consistent across various transformations. More up-to-date consistency regularization methods, such as FixMatch [47] and MixMatch [3], have been developed. FixMatch assigns pseudo-labels to unlabeled samples when the predicting confidence is high with weak augmentation of input data, then the unlabeled data will be treated as labeled data and trained after strong augmentation. MixMatch is another consistency regularization method where pairs of labeled/unlabeled input data are linearly combined to create "Mixup" data. The label of new created data is a weighted average of the original labels followed by a sharpening process. Inspired by self-supervised learning, contrastive learning idea is also integrated in semi-supervised learning [1,46,54]. Instead of seeking agreements in label predictions, contrastive learning focuses on forming positive pairs and negative pairs. Positive pairs originate from the same inputs with different views or augmentations, while negative pairs involve different inputs. The method encourages a shorter distance between positive pairs and a longer distance between negative pairs in the feature representation space.

3 Methods

3.1 An overview of CLASS-M

Figure 1 shows the workflow of our Contrastive learning with adaptive stain separation and mixup (CLASS-M). As shown in orange part of Figure 1, slide-level stain separation is applied for generating Hematoxylin images and Eosin images from original RGB images. The original RGB images, Hematoxylin images and Eosin images are all augmented during training to improve the robustness of CLASS-M model. The model takes Hematoxylin images and Eosin images as input pairs to form different views of input data. The purple and pink boxes in Figure 1 illustrate the H and E ResNet [23] encoders that have the same architecture but separate parameters to generate latent features f_H and f_E. A contrastive loss is proposed for shared latent feature space between H channel and E channel to increase the similarity between f_H and f_E. We take the average of features f_H and f_E and pass it to a linear+softmax layer to predict the labels. For labeled samples in training, the cross-entropy loss is introduced to measure the difference between predicitons and true labels. Moreover, the main challenge in semi-supervised learning in general is the limited number of labeled samples. pseudo-labeling on unlabeled samples has been widely used in many state-of-the-art semi-supervised learning methods [3,47,55]. In MixMatch [3], after pseudo-labeling, the model further used MixUp [56] to introduce virtual samples by linear interpolation of two random samples. Inspired by those ideas and to fully utilize unlabeled samples, we first provide pseudo-labels to unlabeled samples, then add MixUp on both labeled and unlabeled samples to create virtual samples. The labels of the generated samples after MixUp are set to weighted averages of original labels/pseudo-labels with a sharpening process. The green part in Figure 1 shows the workflow of this process for pseudo-labeling. We introduce different parts of CLASS-M model in details in the following subsections.

3.2 Adaptive stain separation

Stain separation is introduced to separate different types of stains present in histological images [41,49,57], such as sepa-

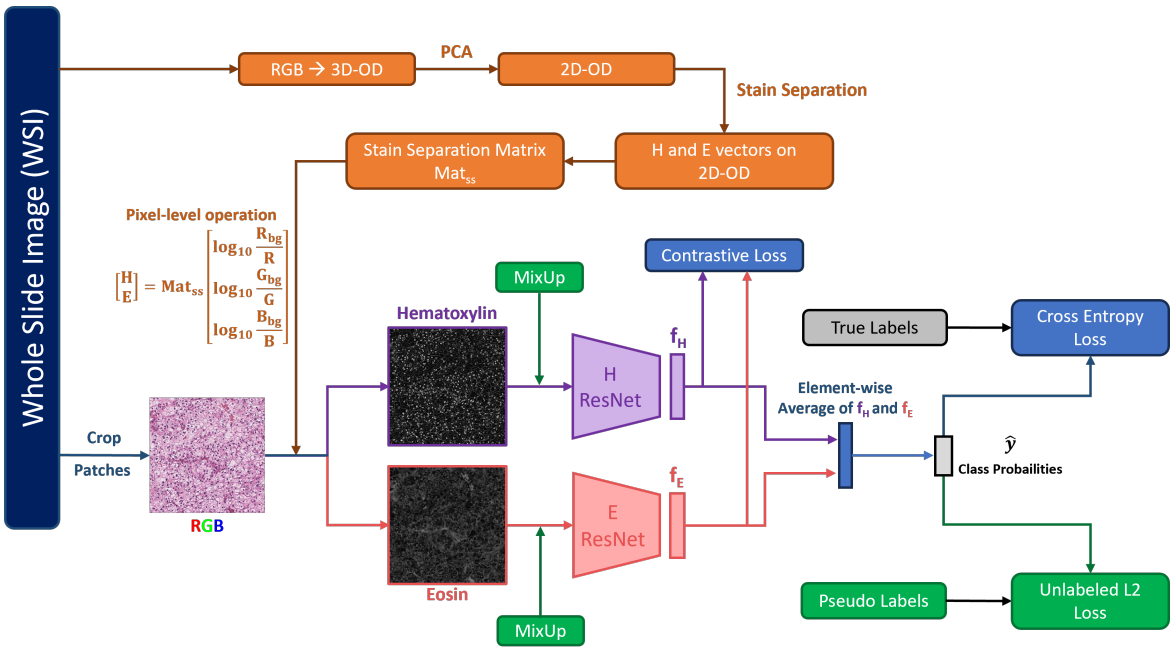

Fig. 1. The block-diagram of our Contrastive learning with adaptive stain separation and mixup (CLASS-M) for semi-supervised histopathological image classification. Orange part shows adaptive stain separation, where OD means Optical Density space. Green part shows mixup on both labeled and unlabeled samples.

rating Hematoxylin and Eosin(H&E) images into Hematoxylin images and Eosin images. The process has many challenges due to the variation of stains caused by different manufacturers, storage conditions and staining procedures. In our model, a simple and unsupervised method for adaptive stain separation [37] is applied. The main idea is to first project all pixels from RGB space to Optical Density (OD) space, where stain components are formed by linear combinations. Inside the OD space, the appropriate Hematoxylin and Eosin vectors are calculated. Then all the RGB pixels can be projected to the Hematoxyling portion and Eosin portion to form Hematoxylin images and Eosin images. The images are further normalized to mitigate the effect of variations on the brightness and dosage of stains. More detailed algorithm can be found in Appendix A.

3.3 Augmentations on original RGB images, Hematoxylin images and Eosin images

We perform color-jittering augmentations on brightness, contrast and saturation of original images, followed by other image augmentations on Hematoxylin images and Eosin images, including random crops, random rotations and random flips. Independent jittering of brightness on Hematoxylin images and Eosin images is also applied. The extra augmentations on original images provide more diversities in the training samples and help generalize more robust features.

3.4 Contrastive learning

The inputs with two different views are usually adopted in co-training. The general presumptions of co-training are that each input view carries sufficient information for the task and the views should be more independent of each other to achieve good results. Our preliminary work [54] has demonstrated that Hematoxylin and Eosin channels fulfill those requirements perfectly in contrastive co-training, especially compared to using the Red, Green or Blue channels in co-training.

Let $f_H(x)$ and $f_E(x)$ denote the output features after the two ResNet models that take sample x as input. Then the contrastive loss function on sample x_i is written as

$$\mathcal{L}_{ct}(x_i) = \max\left(\| f_H(x_i) - f_E(x_i) \|_2 - \| f_H(x_i) - f_E(x_k) \|_2 + m, 0\right) \quad (1)$$

where x_k represents another random sample, $\| a \|_2$ denotes the L2 norm of vector a, and m serves as a margin hyperparameter. This contrastive loss term forms a positive pair whose $f_H(x)$ and $f_E(x)$ originate from the same sample and a negative pair whose $f_H(x)$ and $f_E(x)$ come from different samples. After training, the two features from the same samples tend to get closer in the shared latent feature space.

3.5 MixUp

MixUp [56] is applied on both labeled and unlabeled samples. First, we assign pseudo-labels to unlabeled samples. Assume x_i is an original image sample from unlabeled training set U,

we augment x_i K times and get the average of softmax predictions from the model. The result is sharpened later to lower the entropy of label distribution. The formulas are as follows:

$$\bar{y}_i = \frac{1}{K} \sum_{1 \le k \le K} P_{model}(x_{i,k}), x_i \in U \tag{2}$$

$$y_i = sharpen(\bar{y}_i, T), \text{where } sharpen(y, T)_{(c)} = \frac{y_{(c)}{}^{\frac{1}{T}}}{\sum_{j=1}^{C} y_{(j)}{}^{\frac{1}{T}}} \tag{3}$$

$x_{i,k}$ means the k-th augmentation of sample x_i, T is a temperature hyperparameter to control the sharpness, $y_{(c)}$ means the c-th element in one-hot label encoding y, C means the total number of classes, and y_i is the virtual one-hot label encoding of sample x_i.

We follow the standard procedure of MixUp to mix samples. Let x_i and x_j be two random samples from labeled training set L or unlabeled training set U, whose one-hot label encodings are y_i and y_j. Let λ denote a random number from Beta distribution $Beta(\alpha, \alpha)$, and set $\lambda' = max(\lambda, 1-\lambda)$. Then the new mixed sample and its label are as follows:

$$x' = \lambda' x_i + (1 - \lambda') x_j \tag{4}$$

$$y' = \lambda' y_i + (1 - \lambda') y_j \tag{5}$$

After MixUp, we define x' as labeled sample if x_i is from labeled set to form new labeled set L', the remaining samples form new unlabeled set U' with virtual labels. Since the λ' is a random number during training, there would be a lot more virtual samples compared to the original samples.

3.6 Loss function

The total loss is formed by the summation of losses from label prediction and the contrastive learning. Inspired by MixMatch [3], we use cross-entropy loss in labeled set L' and squared L_2 loss between predictions and virtual labels in unlabeled set U' considering squared L_2 loss is less sensitive to incorrect predictions. The total loss in a batch can be written as

$$\mathcal{L} = \sum_{x_i \in L'} \frac{y_i \log \hat{y}_i}{|L'|} + \lambda_{U'} \sum_{x_i \in U'} \frac{\| y_i - \hat{y}_i \|_2{}^2}{C|U'|} + \lambda_C \sum_{x_i \in L' \cup U'} \mathcal{L}_{c.t.}(x_i) \tag{6}$$

y_i is one-hot encoding of labels (in L') or virtual labels (in U'), $\hat{y}_j$ is the label prediction from the model, C is the number of classes, $|L'|$ and $|U'|$ means the size of labeled set and unlabeled set in a batch. $\lambda_{U'}$ and λ_C are hyperparameters that control the weights of squared L_2 loss and contrastive learning loss.

4 Experiments

4.1 Datasets

To evaluate our method, we applied our CLASS-M model on Utah clear cell renal cell carcinoma (ccRCC) dataset and The Cancer Genome Atlas Program (TCGA) dataset seperately, and compared results with other semi-supervised and self-supervised classification methods.

In the Utah ccRCC dataset, there are 49 Whole Slide Images (WSIs) from 49 patients. First, a pathologist drew polygons inside WSIs to mark areas with certain growth patterns, which were subsequently verified by another pathologist. We randomly split the WSIs into 32, 10, and 7 WSIs for training, validation and test set. The WSIs were then cropped into 400×400 tiles in 10X resolution with sampling step set to 200 pixels inside each labeled polygon. The same process was applied to crop tiles outside polygons in 32 training WSIs to collect unlabeled tiles to form a semi-supervised learning task for a 4-class classification. In detail, there are 28497, 2044, 2522, 4115 tiles in the category of Normal tissue, Low risk cancer, High risk cancer, and Necrosis, respectively, in the labeled training set extracted from polygons in 32 WSIs. Additionally, there are 171,113 unlabeled training tiles, 5472, 416, 334, 2495 tiles for each category in validation set, and 7263, 598, 389, 924 tiles for each category in test set. Tiles with predominantly background areas were removed, along with those tainted by ink. Tile examples can be seen in Figure 2(a). [6]

For the TCGA ccRCC dataset, we have in total 420 WSIs from 420 patients. 150 of them were labeled by a pathologist through drawing polygons with annotations and verified by another pathologist, and the remaining 270 WSIs were unlabeled. If pathologists had disagreements or concerns about labels of polygons, then those annotations were abandoned. The resolution we trained on was 20X. To make it a more challenging task and show the effectiveness of semi-supervised learning, we split the 150 WSIs into 30, 60 and 60 WSIs for labeled training, validation and test set. The tile cropping process was the same as in the Utah ccRCC dataset except for different sampling steps. We chose 200 pixels as sampling step for labeled training set, but 400 pixels for validation and test set, considering they contain enough tiles. We cropped foreground tiles outside polygons in 30 labeled training WSIs, as well as foreground tiles across the 270 unlabeled WSIs to form unlabeled samples by setting sampling step to 400 pixels. The tiles maintained 400×400 size at all times. We split the labeled tiles into 3 categories: Normal tissue, Cancer and Necrosis to perform 3 class semi-supervised learning classification tasks. Tile examples can be seen in Figure 2(b). In summary, we have 84578, 180471, 7932 labeled training tiles, respectively, in the category of Normal tissue, Cancer and Necrosis, as well as 19638, 79382, 1301 validation tiles, and 15323, 62565, 6168 test tiles for each category. The number of unlabeled training tiles is 1,373,684. The TCGA ccRCC dataset is a publicly available dataset, and our annotations for each WSI are available by contacting authors.

[6] The Utah ccRCC dataset is available through a transfer agreement by contacting authors.

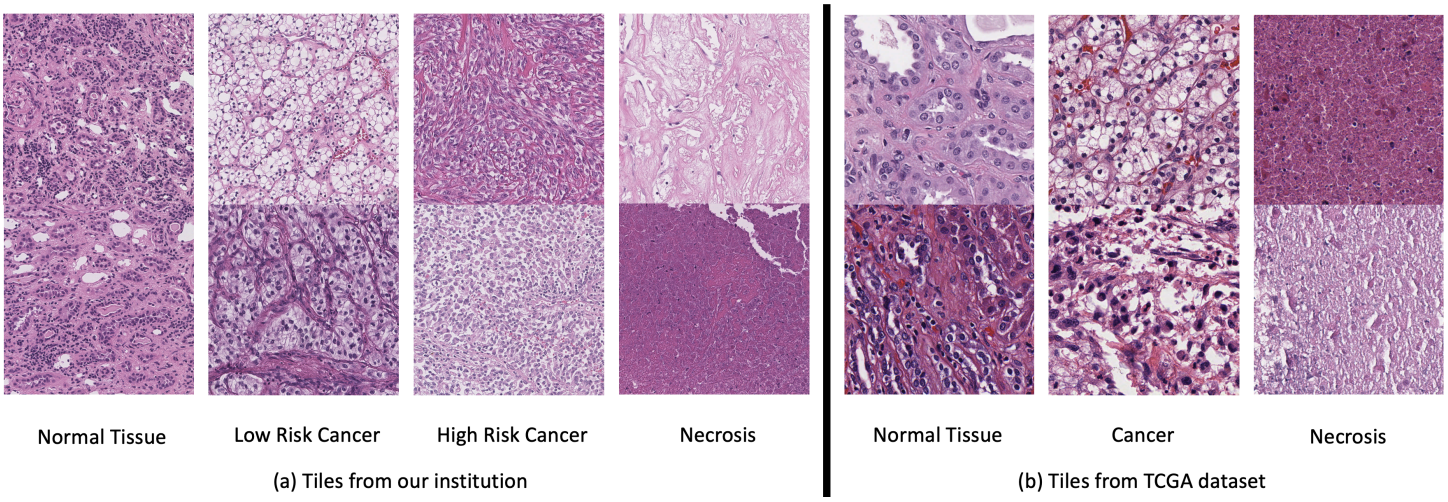

Fig. 2. Examples of tiles with the size of 400×400 from (a) Utah ccRCC dataset (in 10X) and (b) TCGA ccRCC dataset (in 20X).

4.2 Adaptive stain separation

In experiments, we maintained slide-by-slide level adaptive stain separation based on the following considerations: The stain styles are various across whole slide images due to different conditions. But within one whole slide image, most conditions, such as stain manufacturer, storage condition, and staining procedure, remain consistent. On the other hand, if adaptive stain separation is applied patch-by-patch, there are not enough pixel samples inside each patch to obtain robust stain separation results. For example, in Figure 2(a), the top image in Necrosis category has very limited Hematoxylin portion, which makes stain separation less trustful at patch-by-patch level.

The RGB to H/E conversion matrices for each WSI were already calculated before the CLASS-M model training. We provide our original code for adaptive stain separation based on algorithms in [37]. Please contact authors to get access to the code.

4.3 Experiment settings

For a fair comparison, we used ImageNet pretrained ResNet18 model [23] as backbone for all Convolutional Neural Network (CNN) models in semi-supervised learning. We compared our CLASS-M model with other state-of-the-art semi-supervised learning models, including FixMatch [47] and MixMatch [3]. In CLASS-M model training, we performed jittering of brightness, contrast and saturation on original RGB images, as well as image augmentations on H images and E images after stain separation, such as random rotation, random crop to 256×256, random flip and jittering of brightness. The image augmentations on H images and E images are independent, for example, they could have different rotation angles in random rotation. Since H images and E images are one-channel images, we summed the pretrained weights on the first layer of ResNet18 as initialization weights.

We used Root Mean Squared Propagation optimization method with a decaying learning rate. A batch size of 64 was chosen for experiments. Balanced sampler was used to address the imbalance in the number of labels in the training set. In experiments with the Utah ccRCC dataset, we chose to have 32 labeled samples in each batch, with 8 samples for each category, while the remaining 32 samples were unlabeled. In experiments with the TCGA ccRCC dataset, we chose to have 33 labeled samples in each batch, with 11 samples for each category, and the remaining 31 samples would be unlabeled. In validation and test, we first calculated classification accuracies for each category, and then averaged them to get balanced validation accuracy and test accuracy. This approach ensures that each category holds equal importance, irrespective of the actual number of tiles for each category. The hyperparameters were fine-tuned for all models to get the best validation accuracy, then the test accuracy on that hyperparameter setting was the final performance. We found that most hyperparameters didn't need to be changed between Utah ccRCC dataset and TCGA ccRCC dataset. We chose decaying learning rate with initial value set to 10^{-4}. We set number of augmentation times K to 2, temperature T to 0.5, α to 2, margin m in contrastive loss to 37 and unlabeled L_2 loss weight to 7.5 for both datasets. We set contrastive loss weight to 0.1 for experiments on the Utah ccRCC dataset, and contrastive loss weight to 0.001 for the TCGA ccRCC dataset. For each epoch, we ran 1000 iterations and checked validation accuracy once. The total epochs were large enough to ensure convergence. The training would stop if the best validation accuracy was no longer updated for more than 100 epochs. Each experiment was repeated three times to obtain the average test accuracy and standard deviation.

The experiment platform is Python 3.7.11, Pytorch 1.9.0, torchvision 0.10.0, and CUDA 10.2. The GPUs we used are NVIDIA TITAN RTX. During CLASS-M model training, it usually took 13 GB GPU memory and 12 minutes for each

Table 1. Performance of different classification models on Utah ccRCC dataset and TCGA ccRCC dataset. Mean accuracy and standard deviation on test set are calculated. The encoders of self-supervised learning foundational models including Barlow Twins, SwAV, MoCov2, and ViT-DINO are pretrained on large amount of histopathological images by [29] and then frozen. The results with the best mean accuracy are shown in bold.

	Models	Test accuracy (Utah)	Test accuracy (TCGA)
Fully-supervised	ResNet	$88.85 \pm 2.66\%$	$72.11 \pm 0.41\%$
	ViT	$84.69 \pm 1.33\%$	$73.50 \pm 0.94\%$
Transfer learning	ResNet (Freeze)	$85.68 \pm 0.18\%$	$73.17 \pm 0.42\%$
	ViT (Freeze)	$82.51 \pm 0.15\%$	$73.17 \pm 0.46\%$
Self-supervised	BT	$90.64 \pm 1.21\%$	$80.81 \pm 1.79\%$
	SwAV	$91.22 \pm 0.67\%$	$77.23 \pm 3.61\%$
	MoCoV2	$87.35 \pm 0.21\%$	$80.54 \pm 1.75\%$
	ViT-DINO	$92.11 \pm 0.86\%$	$77.61 \pm 3.64\%$
Semi-supervised	FixMatch	$91.58 \pm 0.65\%$	$83.34 \pm 2.53\%$
	MixMatch	$92.94 \pm 1.54\%$	$88.35 \pm 1.39\%$
	CLASS	$94.92 \pm 0.67\%$	$83.06 \pm 0.47\%$
	CLASS-M	$\mathbf{95.35 \pm 0.46\%}$	$\mathbf{92.13 \pm 0.89\%}$

training epoch. The code of our CLASS-M model is publicly available, please contact authors for the access to the code.

4.4 Results

We tested our CLASS-M model along with other state-of-the-art semi-supervised models, including FixMatch and MixMatch by conducting semi-supervised ccRCC histopathological image classification tasks on Utah ccRCC dataset and TCGA ccRCC dataset. The CLASS-M without pseudo-labeling using MixUp was tested as a reference, and we call it CLASS model. The baseline ResNet18 [23] and Vision Transformer (ViT) [14] were also tested as references, where only labeled samples were used in training. The ResNet18 and ViT were pre-trained on ImageNet and then trained with two options, train all the parameters (fully-supervised learning), or freeze the encoder part and only train the final layer (transfer learning). All experiments were run three times to obtain mean test accuracy and standard deviation. Each test accuracy was calculated based on the mean of each category's test accuracy to address the unbalanced nature of the numbers of each category in the test set.

As shown in Table 1, our CLASS-M model outperforms other state-of-the-art models by a large margin on both datasets. We achieved a test accuracy of $95.35 \pm 0.46\%$ on Utah ccRCC dataset, and $92.13 \pm 0.89\%$ on TCGA ccRCC dataset. The ResNet18 results from Table 1 indicates that it is a more challenging task on TCGA ccRCC dataset, whose test accuracy dropped by more than 15%. One possible reason is that the samples from Utah ccRCC dataset have less variation. For instance, the slides follow the same staining procedure, and the storage conditions are uniform. However, on TCGA ccRCC dataset, slides come from different institutions, leading to more variation in the quality of slides. Additionally, the Necrosis tiles on TCGA training set only come from 3 WSIs, which causes more difficulties in learning robust Necrosis features. As shown in the recall section of Table 4 in Appenix B, the pseudo-labeling with MixUp raised the Necrosis's test accuracy from around 60% to around 87% on TCGA ccRCC dataset, demonstrating that MixUp is especially effective on classes with very limited number of samples. In conclusion, our CLASS-M model achieves top results and proves to be the best on a variety of tasks. We also provide recall, precision and F-score for each run, with details available in Appendix B.

4.5 Quantitative comparison to Self-supervised learning foundation models

To demonstrate our superior performance in histopathology image classification by adopting adaptive stain separation-based contrastive learning and pseudo-labeling with MixUp, we did extra self-supervised learning experiments for comparison. We employed a fully-connected classification network built upon pre-trained models from [29], including Barlow Twins [53], SwAV [6], MoCov2 [12], and ViT-DINO [7]. The encoder part was frozen after pre-training on a large amount of histopathological images by [29], and the final fully-connected classification layer was trained on our labeled training samples. As shown in Table 1, our proposed models, CLASS and CLASS-M outperformed those self-supervised learning models on ccRCC datasets even though they were pre-trained on very large histological images. The recall, precision and F-score of those self-supervised learning models can also be found in Appendix B.

4.6 Ablation studies

We conducted ablation studies to further validate our model's H/E contrastive learning structure and identify the components contributing to the classification results.

One of the critical parts of our model is the H/E contrastive loss that provides a regularization term and helps to generate representative features. As shown in Table 2, removing the contrastive loss term from total loss led to a significant drop

Table 2. Ablation studies of CLASS-M models on Utah ccRCC dataset and TCGA ccRCC dataset. Mean accuracy and standard deviation on test set are calculated. The results with best mean accuracy are shown in bold.

Models	Test accuracy (Utah)	Test accuracy (TCGA)
CLASS-M	$\mathbf{95.35 \pm 0.46}\%$	$\mathbf{92.13 \pm 0.89}\%$
CLASS	$94.92 \pm 0.67\%$	$83.06 \pm 0.47\%$
CLASS (use Red/Green channels as two views)	$90.75 \pm 0.13\%$	$81.14 \pm 0.34\%$
CLASS (use Red/Blue channels as two views)	$89.06 \pm 0.54\%$	$80.14 \pm 2.66\%$
CLASS (use Green/Blue channels as two views)	$83.43 \pm 3.63\%$	$80.25 \pm 1.15\%$
CLASS without contrastive loss	$84.57 \pm 2.40\%$	$74.46 \pm 2.03\%$
CLASS without aug on RGB images	$94.99 \pm 0.58\%$	$82.78 \pm 1.10\%$
CLASS without adaptive stain separation (use fixed stain seperation)	$92.43 \pm 0.34\%$	$79.89 \pm 0.69\%$
CLASS without aug on RGB images and adaptive stain separation	$91.53 \pm 1.52\%$	$75.67 \pm 1.76\%$

in the test accuracy. On the Utah ccRCC dataset, the test accuracy dropped from 94.92% to 84.57% in CLASS model. On TCGA ccRCC dataset, the test accuracy declined from 83.06% to 74.46%. Choosing Hematoxylin and Eosin channels as two views also plays a crucial role. Instead of choosing Hematoxylin and Eosin as two channels, we did extra experiments by directly selecting two channels from Red, Green and Blue channels in RGB images to form two views in CLASS model. As a result, we saw obvious reduction in the test accuracies. In addition, splitting the original RGB images into Hematoxylin and Eosin channels is more logical and makes more sense in the aspect of pathology and the true nature of H&E images. The effect of adaptive stain separation was also verified. We used globally fixed stain separation matrix to replace adaptive stain separation. The CLASS model's test accuracy dropped from 94.92% to 92.43% on Utah ccRCC dataset and dropped from 83.06% to 79.89% on TCGA ccRCC dataset. We further did ablation study on the augmentations on original RGB images. The augmentation proved beneficial and improved classification accuracy from 91.53% to 92.43% on dataset from Utah ccRCC dataset and from 75.67% to 79.89% on TCGA ccRCC dataset with fixed stain separation in CLASS model. However, the improvement was limited when adaptive stain separation was adopted. One possible reason is that the augmentation disturbed the color pattern and made adaptive stain separation produce less accurate Hematoxylin and Eosin results.

We finally analyzed the role of pseudo-labeling using MixUp. Table 2 shows that, by comparing CLASS-M with CLASS, we saw a substantial improvement by adding MixUp augmentation, especially on the TCGA ccRCC dataset.

5 Conclusion

In this paper, we proposed a novel semi-supervised classification model called CLASS-M for histopathological image classification by adopting adaptive stain separation-based contrastive learning and pseudo-labeling using MixUp. We provided newly annotated Utah ccRCC dataset and TCGA ccRCC dataset. Experiments on them have shown that our CLASS-M model consistently reached the best classification results compared to other state-of-the-art models. Our model demonstrates the capability to perform accurate patch-level classification at various resolutions with only rough annotations on approximately 30 WSIs in training. The code for our model is also publicly available.

In theory, the advantage of semi-supervised learning is the end-to-end training despite the sacrifice of conveniency compared to self-supervised learning. The self-supervised learning freezes the pre-trained encoders in final training with labeled data, which lacks the flexibility to fine-tune the whole model. However, unfreezing the encoders in final training leads to overfitting as there is no longer access to the large unlabeled dataset, especially when labeled data is limited.

Future work may involve addressing the challenge of handling noisy labels. In our datasets, there is a small portion of tiles that only contain blood vessels, which makes labels inaccurate. A model with more capabilities to tolerate noisy labels may be an interesting road to explore. Our method can also be readily applied to other types of histopathological images, such as immunohistochemistry (IHC) stained images.

Acknowledgments

The work was funded in part by NIH 1R21CA277381. We also acknowledge the support of the Computational Oncology Research Initiative (CORI) at the Huntsman Cancer Institute, ARUP Laboratories, and the Department of Pathology at the University of Utah. The project utilized H&E images from the KIRK cohort of the TCGA Research Network: https://www.cancer.gov/tcga

Appendix A. Detailed algorithm of Adaptive Stain Seperation

First, the RGB pixel values from slide images are transformed into Optical Density (OD) space, where stain components are formed by linear combinations. The corresponding Beer-Lambert law [2,33] for this operation is below:

$$OD_C = \log_{10} \frac{I_{0,C}}{I_C} \tag{7}$$

The letter C represents a specific channel, like Red, Green, or Blue channel. The value $I_{0,C}$ means the intensity of light before passing through the specimen, which is background intensity, and the value I_C is the intensity of light after passing through the specimen. We can treat the OD value as a measurement of the absorption of light on different channels. Zero values in all channels on OD space means pure white background where there is no tissue absorbing light on that pixel. So for each RGB pixel, we use a 3×1 vector $V_{OD} = (OD_R, OD_G, OD_B)^\top$ to show the values on OD space.

Second, dimensionality reduction is applied. In an ideal situation of H&E staining, V_{OD} is a linear combination of fixed Hematoxylin and Eosin unit vectors: $V_{OD} = \alpha_H V_H + \alpha_E V_E$, where V_H and V_E are 3×1 unit vectors on 3D OD space. So all OD vectors should approximately fall into the same 2D plane for all pixels in one image. We use Principal Component Analysis (PCA) on 3D OD space to get a 2D plane formed by two covariance matrix's eigenvectors with the largest two eigenvalues. Figure 3.a is an example of the distribution of OD vectors that map onto the 2D plane. The brightness shows the density of pixels. The x-axis is the direction of the eigenvector with the largest eigenvalue. The y-axis is the direction of the eigenvector with the second-largest eigenvalue. The third dimension would be treated as a residual part and be abandoned. The x and y axes are orthogonal because the eigenvectors of the covariance matrix are orthogonal to each other. By applying the PCA method, the 2D plane we found is more informative, increasing the possibility that the separated H and E images contain more information and details.

The third step is to find stain vectors V_H and V_E. [37] assumes every pixel sample in OD space must exist between the two stain vectors, considering the fact that each component should be non-negative. However, noise can't be fully avoided. As a result, we allow 1% of samples to be outside of V_H and V_E respectively. Moreover, pixel samples that are within a distance of 0.1 from original point are not taken into account as they contain less stain and are more easily influenced by noise. An example of V_H and V_E is shown in Figure 3.b. The positive directions of x-axis and y-axis were carefully chosen such that H unit vector falls into quadrant IV and E unit vector falls into quadrant I.

After acquiring unit vectors V_H and V_E, we are able to reconstruct the transformation matrix. In the previous PCA process, we got three unit eigenvectors V_x, V_y, $V_{Residual}$, which are orthogonal to each other and ordered by their eigenvalues from maximum to minimum. The V_H and V_E on x-y 2D OD space can be written as $V_H = \cos\theta_H V_x + \sin\theta_H V_y$, $V_E = \cos\theta_E V_x + \sin\theta_E V_y$. The transformation matrix from H, E, Residual to RGB on OD space would be

$$Mat_{HERes \to RGB\, on\, OD} = [V_H, V_E, V_{Residual}] \tag{8}$$

Therefore:

$$\begin{aligned}&(OD_R, OD_G, OD_B)^\mathsf{T} = \\ &Mat_{HERes \to RGB\, on\, OD} \times (\alpha_H, \alpha_E, \alpha_{Residual})^\mathsf{T}\end{aligned} \tag{9}$$

If we define $Mat_{RGB \to HERes\, on\, OD}$ as inverse matrix of $Mat_{HERes \to RGB\, on\, OD}$, then:

$$\begin{aligned}&(\alpha_H, \alpha_E, \alpha_{Residual})^\mathsf{T} = \\ &Mat_{RGB \to HERes\, on\, OD} \times (OD_R, OD_G, OD_B)^\mathsf{T}\end{aligned} \tag{10}$$

If we apply this formula to all pixels in an image, then the α_H forms the Hematoxylin image and α_E forms the Eosin image.

Finally, to further normalize Hematoxylin images and Eosin images, the 99th percentile of intensity values is used as an approximation of maximum value. We normalize it to 0.5 and clip to 1.0 for any numbers larger than 1.0 after normalization.

The main advantages of this stain separation method are simplicity and efficiency. We are able to perform it without the need of model training or complex calculations. Moreover, finding optimum H vectors and E vectors helps to handle stain variations. The final normalization based on the calculated maximum value further mitigates the effect of variations on the brightness and hue of stains.

Appendix B. More details on experiment results.

Please check Table 3 and Table 4 for recall, precision and F-score of each model on Utah ccRCC dataset and TCGA ccRCC dataset.

References

1. Alonso, I., Sabater, A., Ferstl, D., Montesano, L., Murillo, A.C.: Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8219–8228 (2021)
2. Beer: Bestimmung der absorption des rothen lichts in farbigen flüssigkeiten. Annalen der Physik **162**(5), 78–88 (1852)
3. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems **32** (2019)
4. Bian, Z., Fan, R., Xie, L.: A novel cuproptosis-related prognostic gene signature and validation of differential expression in clear cell renal cell carcinoma. Genes **13**(5), 851 (2022)
5. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV). pp. 132–149 (2018)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
8. Chan, H.P., Samala, R.K., Hadjiiski, L.M., Zhou, C.: Deep learning in medical image analysis. Deep Learning in Medical Image Analysis: Challenges and Applications pp. 3–21 (2020)

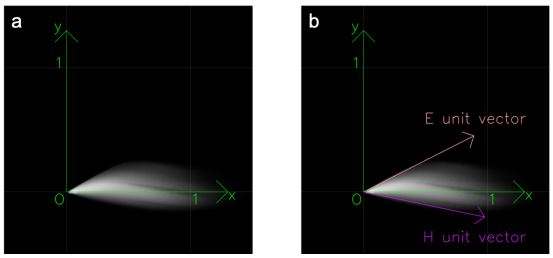

Fig. 3. Mapping pixels onto the 2D OD space

Table 3. Recall, precision and F-score of different classification models on Utah ccRCC dataset. Mean and standard deviation on test set are calculated based on 3 runs. Blue numbers show fully-supervised learning results. Brown numbers show transfer learning results. Green numbers show self-supervised learning results. Black numbers show semi-supervised learning results.

	Models	Normal Tissue	Low Risk Cancer	High Risk Cancer	Necrosis	average (all classes)
Recall	ResNet	0.9308 ± 0.0358	0.9186 ± 0.0254	0.9717 ± 0.0384	0.7330 ± 0.1172	0.8885 ± 0.0266
	ViT	0.8906 ± 0.0138	0.9571 ± 0.0179	0.7112 ± 0.0351	0.8286 ± 0.0278	0.8469 ± 0.0133
	ResNet(Frz)	0.8534 ± 0.0035	0.9326 ± 0.0035	0.7781 ± 0.0121	0.8633 ± 0.0056	0.8568 ± 0.0018
	ViT(Frz)	0.8486 ± 0.0089	0.9415 ± 0.0073	0.7241 ± 0.0193	0.7861 ± 0.0088	0.8251 ± 0.0015
	BT	0.8987 ± 0.0466	0.9810 ± 0.0075	0.8612 ± 0.0268	0.8849 ± 0.0708	0.9065 ± 0.0121
	SwAV	0.9363 ± 0.0205	0.9537 ± 0.0126	0.8946 ± 0.0380	0.8640 ± 0.0542	0.9122 ± 0.0067
	MoCoV2	0.9586 ± 0.0081	0.9426 ± 0.0042	0.7352 ± 0.0068	0.8575 ± 0.0204	0.8735 ± 0.0021
	ViT-DINO	0.9734 ± 0.0089	0.9950 ± 0.0029	0.9632 ± 0.0097	0.7529 ± 0.0411	0.9211 ± 0.0086
	FixMatch	0.8322 ± 0.0759	0.9693 ± 0.0459	0.9563 ± 0.0602	0.9051 ± 0.0558	0.9158 ± 0.0065
	MixMatch	0.9004 ± 0.0261	0.9047 ± 0.0420	0.9588 ± 0.0044	0.9538 ± 0.0092	0.9294 ± 0.0154
	CLASS	0.9753 ± 0.0061	0.9660 ± 0.0092	0.9563 ± 0.0051	0.8994 ± 0.0260	0.9492 ± 0.0067
	CLASS-M	0.9252 ± 0.0194	0.9576 ± 0.0174	0.9880 ± 0.0065	0.9430 ± 0.0051	0.9535 ± 0.0046
Precision	ResNet	0.9702 ± 0.0148	0.8665 ± 0.1087	0.4963 ± 0.1317	0.8930 ± 0.0493	0.8065 ± 0.0493
	ViT	0.9832 ± 0.0034	0.5781 ± 0.0642	0.3941 ± 0.0429	0.8649 ± 0.0547	0.7051 ± 0.0132
	ResNet(Frz)	0.9902 ± 0.0004	0.4538 ± 0.0090	0.3983 ± 0.0088	0.8628 ± 0.0091	0.6762 ± 0.0006
	ViT(Frz)	0.9790 ± 0.0012	0.5345 ± 0.0078	0.3128 ± 0.0158	0.7876 ± 0.0077	0.6535 ± 0.0048
	BT	0.9868 ± 0.0075	0.8436 ± 0.0464	0.5036 ± 0.0831	0.7130 ± 0.1271	0.7618 ± 0.0442
	SwAV	0.9875 ± 0.0030	0.8012 ± 0.0868	0.6008 ± 0.0612	0.8228 ± 0.0980	0.8031 ± 0.0306
	MoCoV2	0.9829 ± 0.0044	0.7937 ± 0.0255	0.5673 ± 0.0307	0.9063 ± 0.0246	0.8126 ± 0.0186
	ViT-DINO	0.9729 ± 0.0048	0.9479 ± 0.0228	0.7767 ± 0.0840	0.8782 ± 0.0295	0.8939 ± 0.0294
	FixMatch	0.9913 ± 0.0059	0.6529 ± 0.1011	0.3679 ± 0.0784	0.7837 ± 0.1874	0.6990 ± 0.0776
	MixMatch	0.9945 ± 0.0015	0.8400 ± 0.1067	0.4072 ± 0.0615	0.8660 ± 0.0518	0.7770 ± 0.0401
	CLASS	0.9918 ± 0.0026	0.9165 ± 0.0103	0.7164 ± 0.0411	0.9439 ± 0.0119	0.8921 ± 0.0111
	CLASS-M	0.9980 ± 0.0005	0.9263 ± 0.0102	0.7316 ± 0.0048	0.6760 ± 0.0666	0.8330 ± 0.0192
F-score	ResNet	0.9497 ± 0.0137	0.8903 ± 0.0706	0.6483 ± 0.1052	0.8005 ± 0.0641	0.8222 ± 0.0256
	ViT	0.9346 ± 0.0063	0.7192 ± 0.0495	0.5065 ± 0.0400	0.8452 ± 0.0152	0.7513 ± 0.0156
	ResNet(Frz)	0.9167 ± 0.0019	0.6104 ± 0.0074	0.5268 ± 0.0062	0.8630 ± 0.0023	0.7292 ± 0.0017
	ViT(Frz)	0.9091 ± 0.0048	0.6819 ± 0.0080	0.4365 ± 0.0118	0.7868 ± 0.0014	0.7036 ± 0.0059
	BT	0.9401 ± 0.0222	0.9066 ± 0.0235	0.6318 ± 0.0565	0.7811 ± 0.0512	0.8149 ± 0.0290
	SwAV	0.9611 ± 0.0099	0.8687 ± 0.0476	0.7164 ± 0.0297	0.8383 ± 0.0264	0.8462 ± 0.0196
	MoCoV2	0.9706 ± 0.0020	0.8616 ± 0.0151	0.6402 ± 0.0220	0.8809 ± 0.0044	0.8383 ± 0.0097
	ViT-DINO	0.9731 ± 0.0023	0.9707 ± 0.0107	0.8586 ± 0.0544	0.8097 ± 0.0117	0.9030 ± 0.0177
	FixMatch	0.9035 ± 0.0415	0.7785 ± 0.0834	0.5263 ± 0.0777	0.8273 ± 0.0878	0.7589 ± 0.0572
	MixMatch	0.9451 ± 0.0137	0.8692 ± 0.0697	0.5699 ± 0.0600	0.9072 ± 0.0281	0.8228 ± 0.0357
	CLASS	0.9835 ± 0.0019	0.9406 ± 0.0093	0.8187 ± 0.0267	0.9208 ± 0.0080	0.9159 ± 0.0044
	CLASS-M	0.9602 ± 0.0106	0.9416 ± 0.0037	0.8407 ± 0.0050	0.7864 ± 0.0479	0.8822 ± 0.0150

Table 4. Recall, precision and F-score of different classification models on TCGA ccRCC dataset. Mean and standard deviation on test set are calculated based on 3 runs. Blue numbers show fully-supervised learning results. Brown numbers show transfer learning results. Green numbers show self-supervised learning results. Black numbers show semi-supervised learning results.

	Models	Normal Tissue	Cancer	Necrosis	average (all classes)
Recall	ResNet	0.7136 ± 0.0191	0.9032 ± 0.0105	0.5466 ± 0.0172	0.7211 ± 0.0041
	ViT	0.7502 ± 0.0164	0.8986 ± 0.0130	0.5563 ± 0.0242	0.7350 ± 0.0094
	ResNet(Frz)	0.7239 ± 0.0054	0.8837 ± 0.0031	0.5876 ± 0.0141	0.7317 ± 0.0042
	ViT(Frz)	0.7383 ± 0.0072	0.9233 ± 0.0038	0.5333 ± 0.0212	0.7316 ± 0.0046
	BT	0.7906 ± 0.0274	0.9650 ± 0.0029	0.6688 ± 0.0681	0.8081 ± 0.0179
	SwAV	0.7462 ± 0.0577	0.9436 ± 0.0227	0.6271 ± 0.0695	0.7723 ± 0.0361
	MoCoV2	0.8377 ± 0.0545	0.9633 ± 0.0093	0.6152 ± 0.0072	0.8054 ± 0.0175
	ViT-DINO	0.7969 ± 0.0962	0.9782 ± 0.0049	0.5531 ± 0.0179	0.7760 ± 0.0364
	FixMatch	0.8582 ± 0.0548	0.9565 ± 0.0165	0.6855 ± 0.0664	0.8334 ± 0.0253
	MixMatch	0.9198 ± 0.0334	0.8910 ± 0.0250	0.8397 ± 0.0156	0.8835 ± 0.0139
	CLASS	0.9291 ± 0.0377	0.9580 ± 0.0212	0.6047 ± 0.0260	0.8306 ± 0.0047
	CLASS-M	0.9477 ± 0.0173	0.9496 ± 0.0060	0.8665 ± 0.0106	0.9213 ± 0.0089
Precision	ResNet	0.6168 ± 0.0214	0.9130 ± 0.0048	0.7622 ± 0.0128	0.7640 ± 0.0062
	ViT	0.6281 ± 0.0302	0.9166 ± 0.0035	0.7879 ± 0.0902	0.7776 ± 0.0385
	ResNet(Frz)	0.5883 ± 0.0071	0.9391 ± 0.0016	0.5735 ± 0.0188	0.7003 ± 0.0044
	ViT(Frz)	0.6919 ± 0.0112	0.9093 ± 0.0011	0.7886 ± 0.0154	0.7966 ± 0.0079
	BT	0.8824 ± 0.0415	0.9324 ± 0.0072	0.7535 ± 0.0809	0.8561 ± 0.0163
	SwAV	0.8290 ± 0.0886	0.9401 ± 0.0106	0.5345 ± 0.0915	0.7679 ± 0.0394
	MoCoV2	0.8391 ± 0.0288	0.9362 ± 0.0140	0.8759 ± 0.0072	0.8837 ± 0.0061
	ViT-DINO	0.8722 ± 0.0096	0.9214 ± 0.0199	0.9432 ± 0.0085	0.9123 ± 0.0070
	FixMatch	0.8651 ± 0.0677	0.9530 ± 0.0151	0.7338 ± 0.1257	0.8506 ± 0.0224
	MixMatch	0.7505 ± 0.0380	0.9939 ± 0.0031	0.5674 ± 0.0380	0.7706 ± 0.0235
	CLASS	0.8727 ± 0.0472	0.9636 ± 0.0108	0.6986 ± 0.1187	0.8449 ± 0.0452
	CLASS-M	0.8592 ± 0.0249	0.9913 ± 0.0037	0.7420 ± 0.0295	0.8641 ± 0.0130
F-score	ResNet	0.6616 ± 0.0183	0.9081 ± 0.0076	0.6365 ± 0.0117	0.7354 ± 0.0051
	ViT	0.6836 ± 0.0222	0.9075 ± 0.0056	0.6503 ± 0.0308	0.7471 ± 0.0192
	ResNet(Frz)	0.6490 ± 0.0029	0.9106 ± 0.0023	0.5803 ± 0.0132	0.7133 ± 0.0039
	ViT(Frz)	0.7143 ± 0.0093	0.9162 ± 0.0016	0.6361 ± 0.0151	0.7554 ± 0.0040
	BT	0.8331 ± 0.0058	0.9484 ± 0.0051	0.7044 ± 0.0267	0.8286 ± 0.0117
	SwAV	0.7815 ± 0.0275	0.9417 ± 0.0062	0.5768 ± 0.0825	0.7667 ± 0.0362
	MoCoV2	0.8370 ± 0.0128	0.9494 ± 0.0029	0.7227 ± 0.0060	0.8364 ± 0.0070
	ViT-DINO	0.8310 ± 0.0563	0.9488 ± 0.0086	0.6971 ± 0.0130	0.8257 ± 0.0258
	FixMatch	0.8588 ± 0.0120	0.9546 ± 0.0062	0.7016 ± 0.0439	0.8383 ± 0.0104
	MixMatch	0.8260 ± 0.0273	0.9395 ± 0.0132	0.6767 ± 0.0282	0.8141 ± 0.0209
	CLASS	0.8987 ± 0.0084	0.9606 ± 0.0061	0.6435 ± 0.0429	0.8343 ± 0.0186
	CLASS-M	0.9011 ± 0.0150	0.9700 ± 0.0017	0.7993 ± 0.0211	0.8901 ± 0.0109

9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
10. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems **33**, 22243–22255 (2020)
11. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
12. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
13. Chikontwe, P., Kim, M., Nam, S.J., Go, H., Park, S.H.: Multiple instance learning with center embeddings for histopathology classification. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23. pp. 519–528. Springer (2020)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Fischer, A.H., Jacobson, K.A., Rose, J., Zeller, R.: Hematoxylin and eosin staining of tissue and cell sections. CSH Protoc. **2008**, db.prot4986 (May 2008). https://doi.org/10.1101/pdb.prot4986
16. Foulds, J., Frank, E.: A review of multi-instance learning assumptions. The knowledge engineering review **25**(1), 1–25 (2010)
17. 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.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
18. Gul, A.G., Cetin, O., Reich, C., Flinner, N., Prangemeier, T., Koeppl, H.: Histopathological image classification based on self-supervised vision transformer and weak labels. In: Medical Imaging 2022: Digital and Computational Pathology. vol. 12039, pp. 366–373. SPIE (2022)
19. Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I.: Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3852–3861 (2020)
20. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
21. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. arXiv preprint arXiv:1911.05722 (2019)
22. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
24. Herrera, F., Ventura, S., Bello, R., Cornelis, C., Zafra, A., Sánchez-Tarragó, D., Vluymans, S., Herrera, F., Ventura, S., Bello, R., et al.: Multiple instance learning. Springer (2016)
25. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2424–2433 (2016)
26. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
27. Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F.: A survey on contrastive self-supervised learning. Technologies **9**(1), 2 (2020)
28. Kanavati, F., Toyokawa, G., Momosaki, S., Rambeau, M., Kozuma, Y., Shoji, F., Yamazaki, K., Takeo, S., Iizuka, O., Tsuneki, M.: Weakly-supervised learning for lung carcinoma classification using deep learning. Scientific reports **10**(1), 9297 (2020)
29. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3344–3354 (2023)
30. Krishnan, R., Rajpurkar, P., Topol, E.J.: Self-supervised learning in medicine and healthcare. Nature Biomedical Engineering **6**(12), 1346–1352 (2022)
31. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
32. Laleh, N.G., Muti, H.S., Loeffler, C.M.L., Echle, A., Saldanha, O.L., Mahmood, F., Lu, M.Y., Trautwein, C., Langer, R., Dislich, B., et al.: Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. Medical image analysis **79**, 102474 (2022)
33. Lambert, J.H.: Photometria sive de mensura et gradibus luminis, colorum et umbrae. sumptibus vidvae E. Klett, typis CP Detleffsen (1760)
34. Li, J., Li, W., Sisk, A., Ye, H., Wallace, W.D., Speier, W., Arnold, C.W.: A multi-resolution model for histopathology image classification and localization with multiple instance learning. Computers in biology and medicine **131**, 104253 (2021)
35. Li, Z., Cong, Y., Chen, X., Qi, J., Sun, J., Yan, T., Yang, H., Liu, J., Lu, E., Wang, L., et al.: Vision transformer-based weakly supervised histopathological image analysis of primary brain tumors. IScience **26**(1) (2023)
36. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
37. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE international symposium on biomedical imaging: from nano to macro. pp. 1107–1110. IEEE (2009)
38. Motzer, R.J., Jonasch, E., Agarwal, N., Alva, A., Baine, M., Beckermann, K., Carlo, M.I., Choueiri, T.K., Costello, B.A., Derweesh, I.H., et al.: Kidney cancer, version 3.2022, nccn clinical practice guidelines in oncology. Journal of the National Comprehensive Cancer Network **20**(1), 71–90 (2022)
39. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 685–694 (2015)

40. Ouali, Y., Hudelot, C., Tami, M.: An overview of deep semi-supervised learning. arXiv preprint arXiv:2006.05278 (2020)
41. Pérez-Bueno, F., Serra, J.G., Vega, M., Mateos, J., Molina, R., Katsaggelos, A.K.: Bayesian K-SVD for H and E blind color deconvolution. applications to stain normalization, data augmentation and cancer classification. Comput. Med. Imaging Graph. **97**, 102048 (Apr 2022). https://doi.org/10.1016/j.compmedimag.2022.102048
42. Qian, Z., Li, K., Lai, M., Chang, E.I.C., Wei, B., Fan, Y., Xu, Y.: Transformer based multiple instance learning for weakly supervised histopathology image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 160–170. Springer (2022)
43. Razzak, M.I., Naz, S., Zaib, A.: Deep learning for medical image processing: Overview, challenges and the future. Classification in BioApps: Automation of Decision Making pp. 323–350 (2018)
44. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. Advances in neural information processing systems **29** (2016)
45. Sarvamangala, D., Kulkarni, R.V.: Convolutional neural networks in medical image understanding: a survey. Evolutionary intelligence **15**(1), 1–22 (2022)
46. Singh, A., Chakraborty, O., Varshney, A., Panda, R., Feris, R., Saenko, K., Das, A.: Semi-supervised action recognition with temporal contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10389–10399 (2021)
47. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
48. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine learning **109**(2), 373–440 (2020)
49. Vasiljević, J., Feuerhake, F., Wemmert, C., Lampert, T.: Towards histopathological stain invariance by unsupervised domain augmentation using generative adversarial networks. Neurocomputing **460**, 277–291 (Oct 2021). https://doi.org/10.1016/j.neucom.2021.07.005
50. Wang, X., Chen, H., Gan, C., Lin, H., Dou, Q., Huang, Q., Cai, M., Heng, P.A.: Weakly supervised learning for whole slide lung cancer image classification. In: Medical imaging with deep learning (2022)
51. Xiong, C., Chen, H., Sung, J., King, I.: Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. arXiv preprint arXiv:2301.08125 (2023)
52. Yang, P., Yin, X., Lu, H., Hu, Z., Zhang, X., Jiang, R., Lv, H.: Cs-co: A hybrid self-supervised visual representation learning method for h&e-stained histopathological images. Medical Image Analysis **81**, 102539 (2022)
53. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. pp. 12310–12320. PMLR (2021)
54. Zhang, B., Knudsen, B., Sirohi, D., Ferrero, A., Tasdizen, T.: Stain based contrastive co-training for histopathological image analysis. In: Workshop on Medical Image Learning with Limited and Noisy Data. pp. 106–116. Springer (2022)
55. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. Advances in Neural Information Processing Systems **34**, 18408–18419 (2021)
56. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
57. Zhang, R., Cao, Y., Li, Y., Liu, Z., Wang, J., He, J., Zhang, C., Sui, X., Zhang, P., Cui, L., Li, S.: MVFStain: Multiple virtual functional stain histopathology images generation based on specific domain mapping. Med. Image Anal. **80**, 102520 (Aug 2022). https://doi.org/10.1016/j.media.2022.102520
58. Zhou, Z.H.: A brief introduction to weakly supervised learning. National science review **5**(1), 44–53 (2018)