Journal of Pathology Informatics 13 (2022) 100133

Contents lists available at ScienceDirect

Journal of Pathology Informatics

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

Original Research Article

A self-supervised contrastive learning approach for whole slide image representation in digital pathology

Parsa Ashrafi Fashi [a], Sobhan Hemati [a,b], Morteza Babaie [a,b,*], Ricardo Gonzalez [a,c], H.R. Tizhoosh [a,b,d]

[a] Kimia Lab, University of Waterloo, Waterloo, ON, Canada
[b] Vector Institute, MaRS Centre, Toronto, ON, Canada
[c] McMaster University, Hamilton, ON, Canada
[d] Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA

ARTICLE INFO

Keywords:
Digital pathology
Representation learning
Computational pathology
Self-supervised learning
Image search
Multiple instance learning
Supervised contrastive learning

ABSTRACT

Image analysis in digital pathology has proven to be one of the most challenging fields in medical imaging for AI-driven classification and search tasks. Due to their gigapixel dimensions, whole slide images (WSIs) are difficult to represent for computational pathology. Self-supervised learning (SSL) has recently demonstrated excellent performance in learning effective representations on pretext objectives, which may improve the generalizations of downstream tasks. Previous self-supervised representation methods rely on patch selection and classification such that the effect of SSL on end-to-end WSI representation is not investigated. In contrast to existing augmentation-based SSL methods, this paper proposes a novel self-supervised learning scheme based on the available primary site information. We also design a fully supervised contrastive learning setup to increase the robustness of the representations for WSI classification and search for both pretext and downstream tasks. We trained and evaluated the model on more than 6000 WSIs from The Cancer Genome Atlas (TCGA) repository provided by the National Cancer Institute. The proposed architecture achieved excellent results on most primary sites and cancer subtypes. We also achieved the best result on validation on a lung cancer classification task.

Introduction

The emergence of digital pathology has opened new horizons in medical image analysis for diagnostic purposes. Histopathology images, also known as whole slide images (WSIs), are generally accompanied with information about the site and type of diseases and malignancies. The recent advances in digital technology enable the fast digital scanning of tissue slides to generate high-quality WSIs. As a result, the volume of WSI archives in hospitals and clinics has been drastically increasing. Consequently, the necessity of timely analysis of WSIs has become apparent to address urgent needs in daily workflow of modern pathology. Hence, the digital scanning of slides, alongside the other benefits of pathology, has made computerized techniques a favorite approach for image analysis and diagnosis. The field of digital pathology has been drastically changing due to the recent success of artificial neural networks in the field of AI. Deep learning can facilitate various pathology tasks such as segmentation, classification of regions and nuclei, and searching among WSIs to find similar morphology. However, the representation of digitized pathology slides has proven to be rather challenging due to the large data size of WSIs (generally larger than 50 000 × 50 000 pixels). Besides, the morphological characteristics that discriminate different diagnoses may be microscopically small which causes a fundamental challenge for WSI representation. Creating a single vector representation directly from a WSI is subject to research with current convolutional neural networks (CNNs). A common approach is to break a WSI into many small patches, feed each patch to a CNN, and aggregate the output features to develop a single WSI representation for search and classification. Nonetheless, developing patch-based feature extraction may not be efficient due to its multi-stage architecture. Also, in aggregation stage, information about patch importance and spatial patch knowledge is often ignored. In this paper, we propose an end-to-end architecture that has two main contributions. Firstly, an end-to-end self-supervised, attention-based multiple instance learning (SS-CAMIL) method, that exploits the primary site information of each WSI, which is almost always available during the tissue preparation and subsequent digitization. Furthermore, we show that employing a supervised contrastive learning approach can improve the quality of model embeddings both in WSI classification and search tasks.

* Corresponding author.
E-mail address: mbabaie@uwaterloo.ca (M. Babaie).

http://dx.doi.org/10.1016/j.jpi.2022.100133
Received 24 June 2022; Received in revised form 16 August 2022; Accepted 22 August 2022
Available online 28 August 2022

Related work

Patch-level WSI representation

Early representation approaches primarily investigated patch-level classification. In Hou et al,[1] the authors extracted and classified patch-level features with a CNN in an iterative fashion. Then, they implemented a multi-label SVM to create a single WSI vector representation. The authors in Coudray et al,[2] extracted multi-magnification features from 20x and 5x magnification and aggregated the features with an average of the probabilities of the corresponding patches. Kalra et al. first cluster the entire tissue with color clustering, then select patches based on the cluster.[3] They employed patch-level embeddings for WSI search.

Multiple instance learning

Multiple instance learning (MIL) is a specific learning scheme where a label is assigned to bag of instances.[4] Considering the bag of patches representation for each WSI, the MIL framework takes into account multiple instances of a slide, to represent WSIs. Recently, authors in Zaheer et al[5] proposed deep MIL where they demonstrated different pooling layers following a specific form can obtain permutation invariant representations. Following this paper, many MIL-based WSI representation schemes have been proposed. The authors of Ilse et al[6] proposed attention-based multiple instance learning to perform weighted pooling over each instance feature. Another example of attention-based pooling in MIL are memory networks (MEM) for learning permutation invariant representations.[7] In Adnan et al,[8] the authors used graph CNNs to consider each instance as a node in a graph and then learned an adjacency matrix to build a graph representation of WSIs. Just recently, Hemati et al[9] have exploited deep sets[5] for MIL training in histopathology. They employed a conditional prediction layer where predictions of primary site labels guide the primary diagnosis predictions.

Self-supervised learning

Self-supervised learning (SSL) refers to a category of deep learning methods in which a model is trained on a set of well-defined pre-text tasks before being applied to a primary (or downstream) task. Pre-text tasks are trained on purposefully generated "*pseudo-labels*" from the data, in order to acquire visual representations for utilizing the acquired model weights for the main task. Gidaris et al[10] is among the first works in vision-based SSL, where authors define *rotation* classification as a self-supervised task and show that various computer vision tasks such as classification, detection, or segmentation generalize better with self-supervision. There have also been some patch-level self-supervision in histopathology literature. In a recent publication, Koohbanani et al[11] proposed pathology-specific tasks such as magnification classification, JigMag (predicting the magnification order in a shuffled vector of different magnification), and hematoxylin channel prediction.

Contrastive learning

Contrastive learning (CL) is another active field of research where the goal is to pull similar instances together and push the non-related samples away. Training a model with a contrastive loss can help produce a more distinct feature vector for an input. The first usage of a contrastive loss appeared in Chopra et al.[12] The authors proposed a similarity loss function that maps training data into a target space such that the L_1 norm of the target space imitates the semantic distance of the input space. They considered pairwise input and chose to either push away or pull the samples based on similarity. In Hoffer and Ailon,[13] instead of two samples for comparison, authors used one instance as an anchor, one negative and one positive sample for metric learning. Khosla et al[14] recently suggested a fully supervised contrastive loss that draws all clusters of points belonging to the same class together while pushing clusters of samples from other classes apart. In most recent papers, CL is implemented in a self-supervised fashion. Authors of Chen et al[15] propose SimCLR (a simple framework for contrastive learning of visual representations) that uses different augmentations as positive samples. As a pathology example, in Ciga et al[16] authors employed SimCLR and achieved excellent results for multiple histopathology downstream tasks, including classification, regression, and segmentation compared to baseline training methods. Another recent pathology example is introduced in Li et al.[17] The authors perform contrastive learning on different magnification levels separately to create hierarchical representation based on combined magnifications for downstream tasks.

Proposed methodology

In this paper, we propose a novel end-to-end WSI level self-supervised approach based on the primary site information as the pretext task. The primary site information corresponds to the organ type of each digital slide and is always available for most WSIs. Many papers have used the primary site as a "soft label".[18] We show that using the primary site information for the pre-text task helps the model generalize better on the diagnosis classification. Using the primary site can also be helpful in cytology field where the architectural information of a tissue is not always available.[19,20] A motivation for exploitation of anatomic site for self-supervision is to enhance and encourage unsupervised learning in pathology, since the annotations are expensive to acquire.[21] We have also utilized a supervised contrastive learning loss to create a more robust representation at the WSI level. The following section provides the step-by-step explanation of the proposed method. The complete methodology is depicted in Fig. 1.

Patch selection

For selection of the histopathology patches, we employed the patch selection method in Kalra et al.[3] The authors utilized a two-step k-mean clustering. The tissue is grouped in the first step using the color histogram. The patch location is then subjected to a second k-means clustering to select spatially varied patches from each color segment. Each patch represents a different WSI location and color. As a result, more regions of a WSI are likely to be considered during training.

Feature extraction

We first modify the patch order in this phase to be fed into the feature extractor block. Suppose we have b batches, each WSI has n distinct patches, and each patch has a width w and height h. The re-shape layer changes each input from the shape (b, n, w, h) to $(b \times n, w, h)$. The patches are now inputted to an EfficientNet B0[22] model for feature extraction. The features from the final convolutional block are then fed to a global max-pooling layer and a fully connected layer to extract vectors of size 1024 for each patch. Another reshape layer is then utilized to convert the output shape to $(b, n, 1024)$.

Attention-based pooling

As displayed in Fig. 1, the feature vectors serve as the input to an attention block. Two fully connected layers plus an extension layer make up the attention block. The two dense layers produce a mask of size (b, n), which is then duplicated to get a size of $(b, n, 1024)$. This is then multiplied with the attention block input and averaged to generate a 1024 vector representation of each WSI. Instead of a simple average pooling layer, the mask learns the weight of each patch (importance factor) and lets the model pick which patch is more representative of the WSI. The authors of Ilse et al[6] showed that the representation of attention-based pooling is permutation-invariant, meaning that the output does not change when the input patches are reordered, hence establishing a large degree of freedom for patch selection.

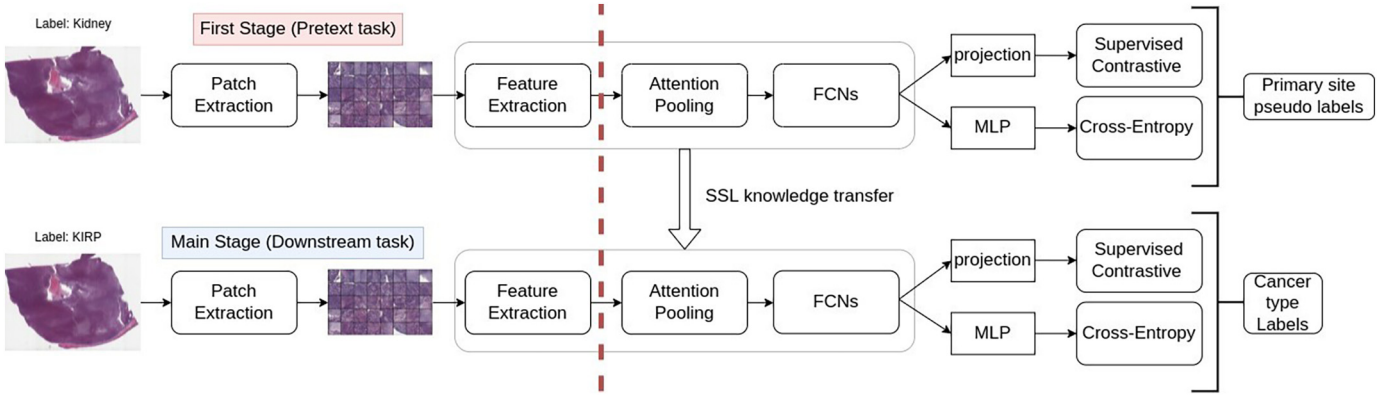

Fig. 1. SS-CAMIL concept. The blocks that the transferred knowledge of pretext task (e.g., for label "kidney" as the primary site) is used for the downstream task (e.g., for label "KIRP", *kidney renal papillary cell carcinoma*, as the primary diagnosis) are outlined with a gray line. For the LUAD/LUSC classification task, only the blocks on the right side of the dashed red line are used for we will be using pre-trained features.

Self-supervision and contrastive learning based on primary site information

The main contribution of this paper is to introduce the exploitation of primary site information as pseudo-labels in self-supervised learning setup (the first training stage). To our knowledge, previous SSL methods in pathology used data augmentation-based self-supervision as pre-text tasks. Primary site information of a WSI is an information that is almost always available and can be used as a pseudo-label (in rare cases, the metadata on an slide may be lost, but in most of these cases, this is not an issue). This is in contrary to tumor labels that requires pathologist annotations, and thus cannot be exploited as a self-supervision pretext label. We have shown in our experiments that transferring the primary site information improves the performance of our model. To evaluate the impact of self-supervision, we conducted the experiments in two phases. First, the results of basic attention-based MIL without self-supervision (CAMIL) are reported. Then, we compare the result of primary site self-supervision on CAMIL (SS-CAMIL). Also, compared to all previous patch-based SSL methods, our self-supervision approach is performed in an end-to-end fashion on WSI-level. The second contribution is the utilization of supervised contrastive learning[14] for both pre-text and downstream tasks. After extracting WSI feature vectors, the features are passed to a projection head and a contrastive loss based on the primary site labels. However, using CL for a MIL setup has a bottleneck. One of the necessities of CL is large batch sizes (commonly more than 256), which may be infeasible due to the extensive set size of each WSI. To overcome this challenge, we added a cross-entropy loss to the contrastive loss function. This is because, CL cannot find enough positive samples within small batch sizes. Adding cross-entropy loss helps the positive instances to be close to a specific point in the embedding space. After the training with the above setting, the model is trained on the downstream task with diagnostic labels (i.e., primary diagnosis). After the training, the features extracted from the last fully connected layer before the projection head are utilized for WSI search and classification.

Experiments and results

Dataset and setup

We exploited 6746 diagnostic WSIs from The Cancer Genome Atlas Program (TCGA) and used 85, 5, and 10 percent (imposed by the ratios published in benchmarking literature) of the dataset for training, validation, and testing, respectively. The dataset consisted of WSIs of 24 primary sites with 30 distinct primary diagnoses. In the training stage, we set the batch size to 16 and the WSI set size (number of patches per WSI) to 40. It should be mentioned that number of slides per primary site and diagnostic is variable, but the number of patches per slide is fixed. We extracted patches of sizes 1000 × 1000 and resized them to 224 × 224 mainly due to memory limits (downsampling patches is quite common in literature[23,24]). For data augmentation, we applied horizontal and vertical flip, 90 degree rotation, shifting, and scaling to the data from the Albumentations library.[25] We used an exponential decay learning rate scheduler with a base of 0.96 and a coefficient of 0.0001. We trained each of the presented results with 150 epochs trained on three Tesla V 100 GPUs in parallel mode. We set the temperature to 0.1 for contrastive learning in both pre-text and downstream tasks. For testing, we established horizontal (site identification) and vertical (subtype identification) WSI-search tasks. The precision with which we can locate a tumor type across the full test database is referred to as horizontal search. Vertical search, on the other hand, measures how well we can identify the proper cancer subtype of a tumor type from a set of slides from a single primary site, which may have a variety of initial diagnoses. For both search tasks, we employ k-NN algorithm with $k = 3$ to find the three instances closest to each test sample. We use the leave-one-out technique and provide the average scores due to the limited size of the test set. We also omitted the results for tumor types with only one subtype, since it will always have the perfect accuracy of 100%.

In another experiment, we employed our model on a classification task of Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). The dataset had 2574 lung tissues, which is distinct from the previous dataset. LUAD/LUSC classification is a challenging classification task that requires visual inspection of the tissue by expert pathologist.[2] We use 1800 slides for training, and 774 slides for test.[7] We freeze the convolutional feature extraction block to demonstrate the learned features from the previous setup. The batch and set sizes are the same as in the above setup. In the final experimetn setup, we test our model's performance on Liver, Kidney, and Stomach (LKS) Immunofluorescence dataset introduced by Maskoud et al. in 2020.[26] It should be mentioned that in this paper, we did not address the imbalanced classes.

WSI search results

Tables 1 and 2 show the horizontal and vertical search results, respectively. We compare our performance with Kalra et al[3] and Hemati et al.[9] In both tables, CAMIL is the baseline attention-based MIL with CL and without self-supervision, and SS-CAMIL is the same as CAMIL setup but uses the weights of self-supervision of primary sites (See Table 3).

For horizontal search, we can observe that the SS-CAMIL model has the best results among the four setups in 10 out of 13 cases. In one of remaining three cases (Prostate/Testis), CAMIL is the dominant model. In the rest of tumor types, SS-CAMIL has shown competitive results. One of the interesting observations is that although CNN-DS utilizes primary site information as prior information for the classification of tumor subtypes, the result of CAMIL is better in most cases. This observation demonstrates the effect of

Table 1
Horizontal search results. F1-scores of majority-3 (in %) are reported.

Tumor type	n_{slides}	Yottixel	CNN-DS	**CAMIL**	**SS-CAMIL**
Brain	46	73	91	100	100
Breast	77	45	77	91	91
Endocrine	71	61	66	86	89
Gastro.	69	50	75	84	86
Gynaec.	18	16	33	56	62
Head/neck	23	17	69	74	92
Liver	44	43	56	77	84
Melanocytic	18	16	50	61	78
Mesenchymal	12	8	100	92	92
Prostate/testis	44	47	81	91	89
Pulmonary	68	58	91	81	87
Urinary tract	112	67	76	92	95
Haematopoietic	42	0	24	50	50

attention-based pooling compared to simple average pooling. Another observation is the improvement in performance with self-supervision on the primary sites. It can be observed that in most tumor types, SSCAMIL has performed better than CAMIL. This observation indicates that the primary site, within the self-supervision framework, can help the model generalize better when deciding on the subtypes.

For the case of vertical search, SS-CAMIL achieves the best, comparable, and worst F1-score in 11, 3, and 10 subtypes out of 24 in comparison to other baselines (the subtypes BLCA, THYM, HNSC, SARC, SKCM, and UVM are not included in the table since they are the only subtype in their tumor types). For five subtypes, CAMIL has performed better. Small sample sizes seem to be a recurrent pattern when our model does not perform well, meaning that the model did not have the chance to learn distinct features from these subtypes. Again, here we can see that in 15 subtypes, self-supervision has helped the model perform better than CAMIL. To show the effect of CL, we show the 2D t-SNE plot of Hemati et al[9] and SS-CAMIL in Fig. 2. We observe that SS-CAMIL clusters are tighter and more separable than CNN-DS.

LUAD/LUSC classification

The results of LUAD/LUSC classification are shown in 3. In this section, we use the features extracted from a DenseNet model,[27] as per Hemati et al.[9] We can observe that our suggested strategy outperformed earlier approaches for LUAD/LUSC classification by 2% (delivering 88 %), that underlines the performance of attention-pooling and constrastive learning.

Table 2
Vertical search results. F1-scores of majority-3 (in %) are reported.

Tumor Type	Subtype	n_{slides}	Yottixel	CNN-DS	**CAMIL**	**SS-CAMIL**
Gastrointestinal tract	COAD	22	62	69	72	73
	STAD	27	61	64	79	92
	ESCA	10	12	44	55	89
	READ	10	30	55	26	0
Pulmonary	LUAD	30	62	61	71	76
	LUSC	35	69	60	76	75
	MESO	3	0	50	50	33
Liver, pancreaticobiliary	LIHC	32	82	95	95	95
	PAAD	8	94	94	94	94
	CHOL	4	26	0	0	0
Endocrine	THCA	50	92	98	99	100
	PCPG	15	61	81	86	90
	ACC	6	25	28	50	77
Urinary tract	KIRP	25	75	84	84	88
	KIRC	47	91	87	92	92
	BLCA	31	89	95	94	98
	KICH	9	70	53	88	80
Brain	LGG	23	78	89	91	89
	GBM	23	82	89	91	90
Prostate/testis	PRAD	31	98	97	94	100
	TGCT	13	96	93	96	100
Gynaecological	OV	9	80	82	76	80
	CESC	6	92	66	44	44
	UCS	3	75	80	100	50

Table 3
LUAD/LUSC classification.

Method	Accuracy
MEM[7]	84%
CNN-DS[9]	86%
CAMIL	**88**%

We have also, employed the SS-CAMIL blocks in this task and it improved the performance to 89%, but since we are not sure whether it has seen the data in the search task (both datasets are from TCGA repository), we did not report those numbers in the table.

Lung–kidney–stomach immunofluorescence

In this experiment setup, we test our model's performance on a different dataset than before. We exploit the Liver, Kidney, and Stomach (LKS) Immunofluorescence dataset introduced by Maskoud et al. in 2020.[26] The data contains immunofluorescence WSIs of the liver, kidney, and stomach that are widely utilized in studying autoimmune liver disease. The dataset consists of 684 immunofluorescence slides split into 479 train and 205 test WSIs. Each WSI in the dataset contains a low-resolution thumbnail image and 1600 high-resolution patches extracted from the slide. Each slide falls into one of the following four classes: Negative (Neg), Anti-Mitochondrial Antibodies (AMA), Vessel-Type Anti-Smooth Muscle Antibodies (SMA-V), and Tubule-Type Anti-Smooth Muscle Antibodies (SMA-T).

To be more memory efficient, we select 40 patches from the 1600 patches of each slide. First, we sort the patches based on their entropy which measures the *information* or uncertainty of a specific signal.

Therefore, we measure how much information each patch contains so that we exclude background and unnecessary patches. We set the batch size to 16 and exploited the same model we used for image search training. We used the trained weights from the image search task. The comparison between reported numbers of Maksoud et al[26] and our model can be seen in Table 4.

The numbers from 4 are taken from Maksoud et al[26] paper directly. Image-level results are related to classification using only the low-level image. Patch-level accuracy is the classification result using high-resolution patches. Multiscale is a conventional classification method using both high- and low-resolution images. RDMS is the abbreviation for Reinforced Dynamic Multi-Scale and is a derivative of Dong et al[28] method. Finally, SOS is the original paper's method that stands for Selective Objective Switch and switches to high-resolution images, if only the prediction confidence of the low-resolution is low.[26]

As observed, our method produces comparable results with regard to the mentioned method. However, some notes should be stated. First, in our model, we only use high-resolution patches. Second, we only use 40 patches among 1600 patches available for each dataset. Therefore, we have produced this result with only 2.5% of the available. Finally, we use an end-to-end classification method compared to SOS, which is a multi-stages classification scheme. With all these notes, our method performed excellently compared to these mentioned methods.

Attention pooling effectiveness

We have also investigated the effectiveness of the attention-pooling layer. As we mentioned, we have used yottixel method[3] for extracting patches to ensure diversity of patches for each slide. We chose nine random WSIs from Lung, Kidney, and Brain organs. A pathology expert scored the effectiveness of the 40 patches from each WSI with labels 1, 2, and 3, meaning "not useful", "somewhat useful", and "very useful", respectively. We multiplied the normalized scores and the output of attention block for each WSI and compared the results with uniform importance (with all patches having the same weight). We then divide the scores by the optimal importance (weights of patches are proportional to effectiveness label) scores to get normalized numbers. The results are shown in Table 5. This

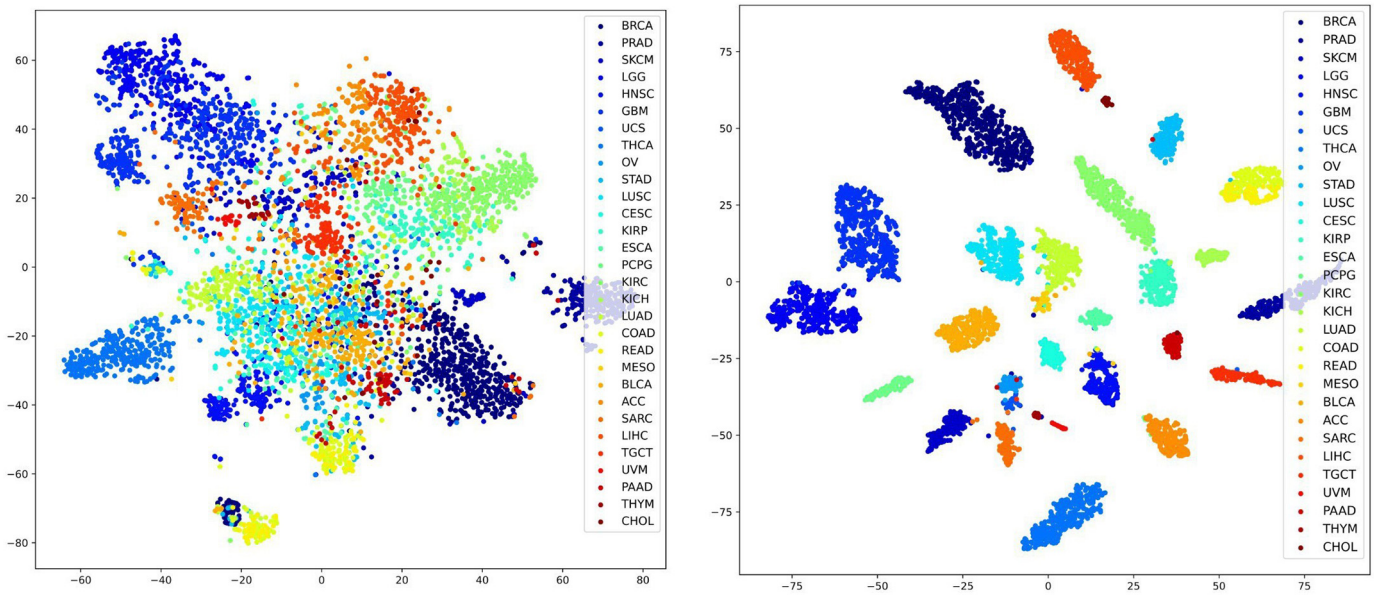

Fig. 2. t-SNE of CNN-DS[9] (left) and SS-CAMIL (right).

Table 4
LKS classification results.

Method	F1-Score
Image-Level	81.95
Patch-Level	69.27
Multi-Scale	85.37
RDMS	88.78
SOS	90.73
SS-CAMIL	88.89

Table 5
Attention pooling scores of 9 different WSIs.

Weighting	Lung			Kidney			Brain			Avg
	1	2	3	1	2	3	1	2	3	
Uniform	0.97	0.89	0.80	0.89	0.70	0.89	0.94	0.87	0.88	0.87
SS-CAMIL	0.98	0.90	0.83	0.91	0.79	0.91	0.96	0.86	0.90	0.89

suggests that our model has learned the relative importance of patches in the attention block.

Conclusions

In this paper, we proposed a self-supervised multiple instance learning model based on primary site information. We showed that our WSI-level representation model generalizes better on tumor subtypes comparing to two previous approaches. We demonstrated our performance on two tasks; WSI search and classification and showed that our model has a dominant performance on both tasks.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Parsa Ashrafi Fashi, Sobhan Hemati, Morteza Babaie, Ricardo Gonzalez, H.R. Tizhoosh reports financial support was provided by Government of Ontario. Parsa Ashrafi Fashi, Sobhan Hemati, Morteza Babaie, Ricardo Gonzalez, H.R. Tizhoosh reports financial support was provided by Mayo Clinic Rochester. All authors are affiliated with Kimia Lab, University of Waterloo. Professor Tizhoosh and Dr. Gonzales are also affiliated with Mayo Clinic, Rochester.

Appendix A. Cancer subtype abbreviations

Table A.6
TCGA cancer subtype abbreviations.

Abbreviation	Primary diagnosis
ACC	Adrenocortical Carcinoma
BLCA	Bladder Urothelial Carcinoma
CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenoc.
CHOL	Cholangiocarcinoma
COAD	Colon Adenocarcinoma
ESCA	Esophageal Carcinoma
GBM	Glioblastoma Multiforme
KICH	Kidney Chromophobe
KIRC	Kidney Renal Clear Cell Carcinoma
KIRP	Kidney Renal Papillary Cell Carcinoma
LGG	Brain Lower Grade Glioma
LIHC	Liver Hepatocellular Carcinoma
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
MESO	Mesothelioma
OV	Ovarian Serous Cystadenocarcinoma
PAAD	Pancreatic Adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate Adenocarcinoma
READ	Rectum Adenocarcinoma
STAD	Stomach Adenocarcinoma
TGCT	Testicular Germ Cell Tumors
THCA	Thyroid Carcinoma
UCS	Uterine Carcinosarcoma

References

1. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based convolutional neural network for whole slide tissue image classification. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 2424–2433.
2. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med 2018;24(10):1559–1567.
3. Kalra S, Tizhoosh HR, Choi C, et al. Yottixel–an image search engine for large archives of histopathology whole slide images. Med Image Anal 2020;65, 101757.
4. Dietterich TG, Lathrop RH, Lozano-Perez T. Solving the multiple instance problem with axis-parallel rectangles. Artif Intel 1997;89(1–2):31–71.
5. Zaheer M, Kottur S, Ravanbakhsh S, Poczos B, Salakhutdinov RR, Smola AJ. Deep sets. Adv Neural Inform Process Syst 2017;30.

6. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. International conference on machine learning. PMLR; 2018. p. 2127–2136.
7. Kalra S, Adnan M, Taylor G, Tizhoosh HR. Learning permutation invariant representations using memory networks. European Conference on Computer Vision. Springer; 2020. p. 677–693.
8. Adnan M, Kalra S, Tizhoosh HR. Representation learning of histopathology images using graph neural networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020. p. 988–989.
9. Hemati S, Kalra S, Meaney C, Babaie M, Ghodsi A, Tizhoosh H. Cnn and deep sets for end-to-end whole slide image representation learning. Medical Imaging with Deep Learning; 2021.
10. Gidaris S, Singh P, Komodakis N. Unsupervised representation learning by predicting image rotations. International Conference on Learning Representations; 2018.
11. Koohbanani NA, Unnikrishnan B, Khurram SA, Krishnaswamy P, Rajpoot N. Self-path: self-supervision for classification of pathology images with limited annotations. IEEE Trans Med Imaging 2021;40(10):2845–2856.
12. Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1. IEEE; 2005. p. 539–546.
13. Hoffer E, Ailon N. Deep metric learning using triplet network. International workshop on similarity-based pattern recognition. Springer; 2015. p. 84–92.
14. Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning. Adv Neural Inform Process Syst 2020;33.
15. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. International conference on machine learning. PMLR; 2020. p. 1597–1607.
16. Ciga O, Xu T, Martel AL. Self supervised contrastive learning for digital histopathology. Mach Learn Appl 15 March 2022;7, 100198.
17. Li B, Li Y, Eliceiri KW. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 14318–14328.
18. Riasatian A, Babaie M, Maleki D, et al. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. Med Image Anal 2021;70, 102032.
19. Girolami I, Marletta S, Pantanowitz L, et al. Impact of image analysis and artificial intelligence in thyroid pathology, with particular reference to cytological aspects. Cytopathology 2020;31(5):432–444.
20. Eccher A, Girolami I. Current state of whole slide imaging use in cytopathology: pros and pitfalls. Cytopathology 2020;31(5):372–378.
21. Nam S, Chong Y, Jung CK, et al. Introduction to digital pathology and computer-aided pathology. J Pathol Transl Med 2020;54(2):125–134.
22. Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. International Conference on Machine Learning. PMLR; 2019. p. 6105–6114.
23. Tizhoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: challenges and opportunities. J Pathol inform 2018;9.
24. Marini N, Otalora S, Muller H, Atzori M. Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: an experiment on prostate histopathology image classification. Med Image Anal 2021;73, 102165.
25. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: fast and flexible image augmentations. Information 2020;11(2):125.
26. Maksoud S, Zhao K, Hobson P, Jennings A, Lovell BC. Sos: selective objective switch for rapid immunofluorescence whole slide image classification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 3862–3871.
27. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4700–4708.
28. Dong N, Kampffmeyer M, Liang X, Wang Z, Dai W, Xing E. Reinforced auto-zoom net: towards accurate and fast breast cancer segmentation in whole-slide images. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer; 2018. p. 317–325.