Article:
Koohbanani, N.A., Unnikrishnan, B., Khurram, S.A. orcid.org/0000-0002-0378-9380 et al. (2 more authors) (2021) Self-Path: self-supervision for classification of pathology images with limited annotations. IEEE Transactions on Medical Imaging, 40 (10). pp. 2845-2856. ISSN 0278-0062

https://doi.org/10.1109/TMI.2021.3056023

IEEE Transactions on Medical Imaging

Self-Path: Self-supervision for Classification of Pathology Images with Limited Annotations

Journal:	*IEEE Transactions on Medical Imaging*
Manuscript ID	TMI-2020-2016.R1
Manuscript Type:	Special Issue on Annotation-Efficient Deep Learning for Medical Imaging
Date Submitted by the Author:	31-Dec-2020

Complete List of Authors:	Alemi, Navid; University of Warwick, Computer Science Unnikrishnan, Balagopal; Institute for Infocomm Research, Machine Intellection Department Khurram, Syed Ali ; The University of Sheffield, The School of Clinical Dentistry Krishnaswamy, pavitra; Institute for Infocomm Research, Machine Intellection Department Rajpoot, Nasir; University of Warwick, Department of Computer Science

Keywords:	Microscopy < Imaging modalities, Neural network < General methodology, Computer-aided detection and diagnosis < General methodology

IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. XX, NO. XX, XXXX 2020 1

Self-Path: Self-supervision for Classification of Pathology Images with Limited Annotations

Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram,
Pavitra Krishnaswamy and Nasir Rajpoot, *Senior Member, IEEE*

***Abstract*— While high-resolution pathology images lend themselves well to 'data hungry' deep learning algorithms, obtaining exhaustive annotations on these images for learning is a major challenge. In this paper, we propose a self-supervised convolutional neural network (CNN) framework to leverage unlabeled data for learning generalizable and domain invariant representations in pathology images. Our proposed framework, termed as Self-Path, employs multi-task learning where the main task is tissue classification and pretext tasks are a variety of self-supervised tasks with labels inherent to the input images. We introduce novel pathology-specific self-supervision tasks that leverage contextual, multi-resolution and semantic features in pathology images for semi-supervised learning and domain adaptation. We investigate the effectiveness of Self-Path on 3 different pathology datasets. Our results show that Self-Path with the pathology-specific pretext tasks achieves state-of-the-art performance for semi-supervised learning when small amounts of labeled data are available. Further, we show that Self-Path improves domain adaptation for histopathology image classification when there is no labeled data available for the target domain. This approach can potentially be employed for other applications in computational pathology, where annotation budget is often limited or large amount of unlabeled image data is available.**

***Index Terms*— Computational pathology, Limited annotation budget, Semi-supervised learning, Domain adaptation.**

I. Introduction

THE recent surge in the area of computational pathology can be attributed to the increasing ubiquity of digital slide scanners and the consequent rapid rise in the amount of raw pixel data acquired by scanning of histology slides into digital whole-slide images (WSIs). These developments make the area of computational pathology ripe ground for deep neural network (DNN) models. In recent years, there have been notable successes in training DNNs for pathology image analysis and automated diagnosis of disease in the histopathology domain [1]. The performance and generalizability of most DNNs is, however, highly dependent on the availability of large and diverse amounts of annotated data. Although the use of digital slide scanners have made large amounts of raw data available, development of DNN based algorithms remains bottlenecked by the need for extensive annotations on diverse datasets.

In pathology, annotation burden can pose a large problem – even more so when compared to natural scene images. WSIs are by nature high resolution images (sometimes with slide dimensions as large as 200,000 $\times$ 150,000 pixels) – this hinders exhaustive annotations. For even simple use cases like detecting tumor regions or isolated tumor cells in WSIs, pathologists annotating the data need to look at regions of the tissue at multiple levels of magnification. So, even simple labeling of regions of interest can be quite demanding. This issue is compounded by the fact that the whole image can only be annotated part by part owing to its large size. Further, the annotation effort requires expert domain knowledge and significant investment on the part of specialized pathologists. To overcome these challenges, when training DNNs on new pathology image datasets, it would be desirable to pursue one or both of the following strategies: (a) labeling small amounts of the new dataset and making use of the larger pool of the unlabeled data, and/or (b) using existing labeled datasets which closely match the new dataset.

For strategy (a), *semi-supervised deep learning* approaches that learn with small amounts of labeled data and leverage larger pools of unlabeled data to boost performance can be employed. These approaches have been widely demonstrated in the computer vision community for natural scene images. Particularly popular techniques include Mean Teacher [2] and Virtual Adversarial Training (VAT) [3]. Recently, these approaches have also been applied to the area of computational pathology to address tasks such as clustering [4], segmentation [5] and image retrieval [6]. However, due to the high dimensionality of the images, the multi-scale nature of the problem, the requirement of contextual information and texture-like nature of sub-patches extracted from slides, the direct translation of popular semi-supervised algorithms into pathology classification tasks is not feasible.

For strategy (b), *domain adaptation* approaches that transfer knowledge from existing resources for related tasks to the classification task-at-hand can be employed. However, due to

Navid Alemi Koohbanani and Nasir Rajpoot are with the Department of Computer Science, University of Warwick, Coventry, CV4 7AL UK (e-mail: n.alemi-koohbananni, n.m.rajpoot@warwick.ac.uk) and also with The Alan Turing Institute, London. NR is also affiliated with the Department of Pathology, University Hospitals Coventry & Warwickshire, UK.

Syed Ali Khurram is with The School of Clinical Dentistry, University of Sheffield, 19 Claremont Crescent, Sheffield, S10 2TA UK (e-mail: s.a.khurram@sheffield.ac.uk).

Balagopal Unnikrishnan and Pavitra Krishnaswamy are with Department of Machine Intellection, Institute for Infocomm Research, (email: balagopalu, pavitrak@i2r.a-star.edu.sg).

variations in tissue, tumor types, and stain appearance during image acquisition, different pathology image datasets appear quite distinct from one another. In addition, for some rare tissue or tumor types, there may be no annotated datasets available for such knowledge transfer. Hence, direct translation of existing domain adaptation algorithms which work for natural vision images may not be possible. Yet, unlabeled data for related tasks are largely available and are less prone to bias [7]. Hence, when dealing with limited annotations, such unlabeled data can be used to capture the shared knowledge or to learn representations that can improve model performance.

To address the dual challenges of low annotations and domain adaptation in histopathology, it is possible to use unlabeled data in a self-supervised manner. In this setup, the model is supervised by labels that come inherently from the data itself without any additional manual annotations. These labels can represent distinct morphological, geometrical and contextual content of the images. Models trained on these 'free' labels can learn representations that can improve performance for a variety of tasks such as classification, segmentation and detection [8]. Self-supervision tasks can be used together with the main supervised task in a multi-task setup to improve performance for semi-supervised learning and domain adaptation [9]. However, self-supervised tasks proposed in the literature so far are mainly based on characteristics of natural scene images, which are very different from histology images. For instance, common self-supervision tasks focus on predicting the degree of rotation, flipping, and/or the relative position of objects. While these are meaningful concepts for natural scene images, they do not carry much relevance for histopathology images. Specifically, while the degree of rotation could help to also learn semantic information present in a natural image, it would not make sense for pathology images because they have no sense of global orientation [10].

In this paper, we propose the **Self-Path** framework to leverage self-supervised tasks customized to the requirements of the histopathology domain, and enhance DNN training in scenarios with limited or no annotated data for the task at hand. Our main contributions are summarized as follows:

- We introduce a generic and flexible self-supervision based framework, Self-Path, for classification of pathology images in the context of limited or no annotations.
- We propose 3 novel pathology pathology specific self-supervision tasks, namely, prediction of magnification level, solving the magnification jigsaw puzzle and prediction of the Hematoxylin channel, aimed at utilizing contextual, multi-resolution and semantic features in histopathology images.
- We conduct a detailed investigation on the effect of various self-supervision tasks for semi-supervised learning and domain adaptation for three datasets.
- We demonstrate that Self-Path achieves state-of-the art performance in limited annotation regime (when 1-2% of the whole dataset is annotated) or even when no annotations are available (in the case of domain adaptation).

A. Related Work

Semi-supervised Learning: Semi-supervised deep learning approaches are widely studied in the computer vision literature [11]. Popular methods utilize forms of pseudo labelling and consistency regularization, and utilize small amounts of labeled data alongside larger pools of unlabeled data for learning. Pseudo-labeling approaches [12] use available labels to train a model and impute labels on the unlabeled samples which are in turn used in training. MixMatch extends pseudo-labeling by adding temperate sharpening along with the mix-up augmentation [13] . Consistency-based methods regularize the model by ensuring stable outputs for various augmentations of the same sample. These can be done by enforcing consensus between temporal ensembles of network outputs like in Pi-Model [14], or between perturbed images fed to a network and its EMA averaged counterpart like in Mean Teacher [2]. Virtual adversarial training(VAT) [3] generates the perturbed images in an adversarial fashion to smooth the margin in the direction of maximum vulnerability. These methods ensure generalizability against significant image perturbations, move the margin away from high-density regions, and enable strong performance on benchmark natural scene image tasks with low annotation budgets.

However, semi-supervised learning has not been sufficiently explored in pathology image analysis. At the time of this writing, only 6 papers investigate semi-supervised learning for the histopathology domain. In [5], Li et. al proposed an EM-based approach for semi-supervised segmentation of histology images. [4] proposed a cluster based semi-supervised approach to identify high-density regions in the data space which were then used by supervised SVM in finding the decision boundary. Jaiswal *et al.* [15] used pseudo-labels for improving the network performance for metastasis detection of breast cancer. Su *et al.* [16] employed global and local consistency losses for mean teacher approach for nuclear classification. Shaw et. al [17] also proposed to use pseudo-labels of unlabeled images for fine-tuning the model iteratively to improve performance for colorectal image classification. Deep multiple instance learning and contrastive predictive coding were used together in [18] to overcome the scarcity of labeled data for breast cancer classification. Yet, there is scope for improvement to close the gap between fully supervised baselines and semi-supervised methods employing just a few labeled pathology images.

Domain Adaptation: Domain adaptation methods focus on adapting models trained on a source dataset to perform well on a target dataset. Leading-edge techniques mainly use adversarial training for aligning the feature distributions of different domains. Popular domain-adversarial learning-based methods [19], [20] use a domain discriminator to classify the domain of images. These methods play a minimax game where the discriminator is trained to distinguish the features from the source or target sample, while the feature generator is trained to confuse the discriminator. [21] employed adversarial learning and minimized Wassertein distance between domains to learn domain-invariant features. Image-translation methods minimize the discrepancy between the two domains at an

image-level [22]. In pathology, Ren *et al.* [23] employed adversarial training for domain adaptation across acquisition devices (scanners) in a prostate cancer image classification task. [24] used CycleGAN to translate across domains for a cell/nuclei detection task. [25] introduced a measure for evaluating distance between domains to enhance the ability to identify out-of-distribution samples in a tumor classification task. Yet, most practical domain adaptation techniques require labeling of target domain data, and the applicability of state-of-the-art unsupervised domain adaptation approaches for histopathology is yet to be widely established.

Self-Supervision: Self-supervision employs pretext tasks (based on annotations that are inherent to the input data) to learn representations that can enhance performance for the downstream task [8]. Autoencoders [26] are the simplest self-supervised task, where the goal is to minimize reconstruction error and the proxy labels are the values of image pixels. Other self-supervised tasks in the literature are image generation [8], inpainting [27], colorizing grayscale images [28], predicting rotation [29], solving jigsaw puzzle [30], and contrastive predictive coding [31]. Perhaps the main difference between contrastive learning approaches and methods like ours is that while our method caters to a specific use case domain and the task at hand is to come up with self-supervision tasks, the contrastive learning approaches offer the advantage of a more generic framework for learning representations potentially at the cost of losing performance in a very specific use case domain (such as histopathology). Although the classical self-supervision approaches requires no additional annotations, it is also possible to leverage small amounts of labeled data within a self-supervision framework. For example, S4L [9] showed that the pretext task (e.g., rotation, self-supervised exemplar [32]) can benefit from small amount of labeled data alongside larger unlabeled data. Moreover, some works [33], [34] demonstrated the effect of self supervised tasks for domain adaptation, where in [34] the effect of various self-supervised tasks have been shown for domain alignment. Particularly, solving jigsaw puzzle [34] has been proved to be a beneficial pretext task for domain generalization.

As there is no large labeled dataset akin to ImageNet for pretraining in the pathology domain, self supervised learning offers potential to obtain pre-trained model that preserves the useful information about data in itself. Although one recent study [35] explored self-supervised similarity learning for pathology image retrieval, much of the self-supervision literature is focused on computer vision applications. A key challenge in applying self-supervision to pathology-specific applications is to define the pretext task that will be most beneficial. As such, systematic analysis and derivation of pretext tasks customized for a range of histopathology applications would be desirable.

II. Problem Formulation

We now define the problem of semi-supervised learning and domain adaptation for pathology image classification. Consider a whole slide image (WSI) that is comprised of a number of disjoint or overlapping 'patches'. We denote an input image or 'patch' as $\mathbf{x}$ and its associated class label as y.

*a) **Semi-supervised Learning**:* We consider a set of n_l limited labeled images $S_L = \{(\mathbf{x}_i^l, y_i)\}_{i=1l}^{n}$, and a set of $m_l >> n_l$ unlabeled images $S_U = \{(\mathbf{x}_i^u)\}_{i=1l}^{m}$. The semi-supervised framework seeks to leverage the large pool of unlabeled images in S_U to enhance the generalizability of learning with fewer labeled images in S_L. Generally, in the semi-supervised setting, both S_L and S_U are from the same distribution.

*b) **Domain Adaptation**:* We define a source domain S comprising a set of η_s labeled images $D_s = \{\mathbf{x}_i^s, y_i^s\}_{i=1}^{\eta_s}$. Likewise, we have a target domain T comprising a set of η_t unlabeled images $D_t = \{\mathbf{x}_i^t\}_{i=1}^{\eta_t}$. Both source and target domains have the same labels. Further, source and target domains have related task characteristics, but their data distributions are distinct.

III. Methods

Our proposed Self-Path framework is depicted in Figure 1. To address label scarcity for the main classification task (main task), Self-Path leverages self-supervision and informs the supervised learning for the main task with the self-supervised learning for pretext tasks. Further, our proposed framework employs a multi-task learning approach to learn class-discriminative and domain-invariant features that would generalize with limited annotated data. Specifically, Self-Path (a) can leverage one or more pathology-specific or pathology-agnostic pretext tasks, (b) is amenable to adversarial or non-adversarial training, and (c) allows flexibility to incorporate semi-supervised, generative learning and/or domain adaptation approaches. We now formally describe the multi-task learning objective and detail the pretext tasks that are used along with the main task.

A. Multi-task Learning

Our proposed approach trains the model using the main and pretext tasks in conjunction. The framework comprises a shared encoder which learns features that are common to both the pretext task and the main task. Each task usually has a separate head connected to the shared encoder and learning for all tasks is optimized simultaneously. Formally,

$$\begin{aligned}\underset{\theta_c,\theta_e,\,\theta_{p_1},..\theta_{p_k}}{\operatorname{argmin}} \; & \tfrac{1}{n_l}\sum_{i}^{n_l} L_c(F_c^{\theta_c}(F_e^{\theta_e}(\mathbf{x}_i^l)),\; y_i) \\ +\tfrac{1}{n_l}\sum_{k=1}^{K}\alpha_{p_k}\sum_{i}^{n^l} & L_{p_k}(F_{p_k}^{\theta_{p_k}}(F_e^{\theta_e}(\mathbf{x}_i^l)),\; r_{ik}^l) \\ +\tfrac{1}{n_u}\sum_{k=1}^{K}\alpha_{p_k}\sum_{i}^{n_u} & L_{p_k}(F_{p_k}^{\theta_{p_k}}(F_e^{\theta_e}(\mathbf{x}_i^u)),\; r_{ik}^u)\end{aligned} \quad , \tag{1}$$

where K is the number of pretext tasks, r is the label for pretext task; L_c and L_{p_k} are the losses for the main and pretext tasks, respectively; F_e, is the shared encoder, F_c is the function for main task and F_{p_k} is the function of k^{th} pretext task; θ_c, θ_e and θ_{p_n} are parameters of main task classifier, shared encoder and pretext tasks, respectively; α_{p_k} indicates weights for different tasks; and n_l and n_u indicate the number of

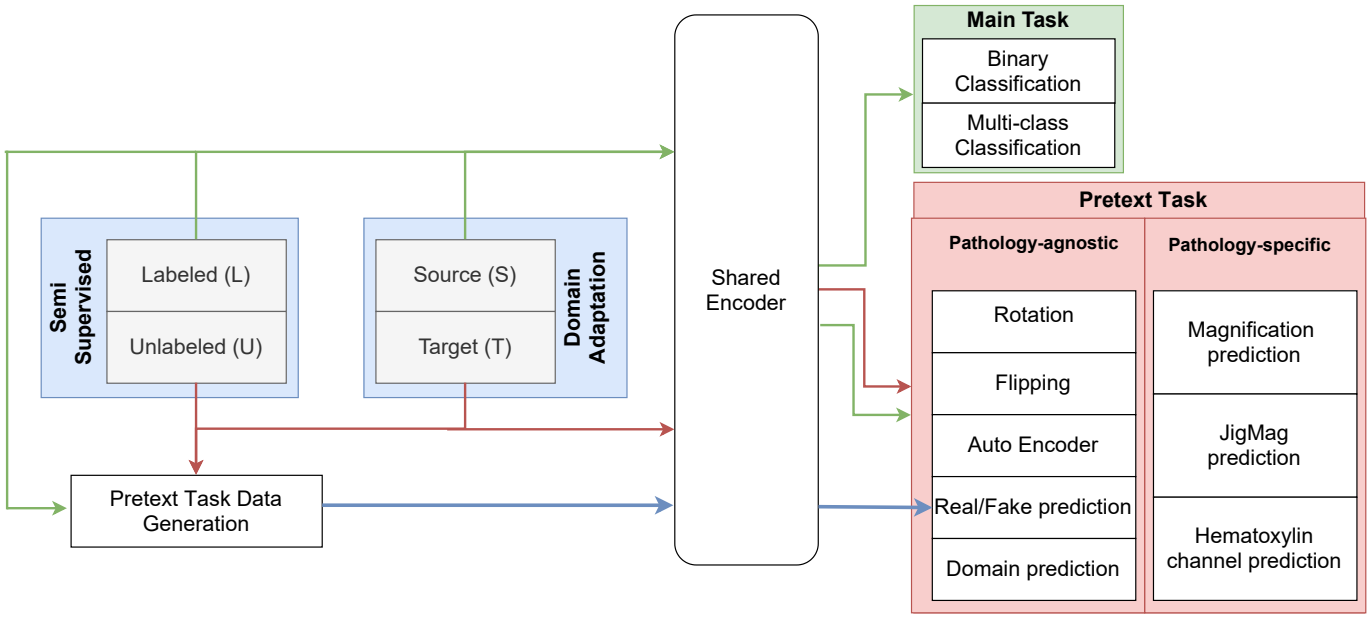

Fig. 1. Overview of Self-Path : The framework employs self-supervised pretext tasks. Pretext tasks can be added atop a shared encoder to learn useful representations and enhance semi-supervised learning or domain-adaptation. Green, red and blue lines indicate the flow of labeled, unlabeled and generated images, respectively. Generated images are used only for the generative task.

labeled and unlabeled images, respectively. When this model is used for semi-supervised learning, the labeled and unlabeled images come from the same domain. When used for domain adaptation, the labeled images come from source domain and unlabeled images come from the target domain.

B. Self-Supervision

The self-supervision utilizes one or more pretext tasks to leverage information in the unlabeled images and improve performance for the main task. Our setup employs both pathology-specific and pathology-agnostic self-supervised tasks. Every pretext task p_k is defined by a transformation function g_k applied to input x, and an implicit label r_k for the transformed input $\tilde{x} = g_k(x)$. Then, the objective function L_{p_k} is the objective for learning the self-supervised classification task that maps $\tilde{x}$ to r_k.

C. Pathology-specific Pretext tasks for Self-supervision

Histopathology images can vary in shape, morphology and arrangement of the nuclei across tissue types and disease conditions. Learning these features or semantic representations of these features can enable generalizable classification models that can more effectively transfer knowledge across domains. Therefore, we design pathology-specific pretext tasks that cater to morphology, context and shapes of nuclei as detailed below :

1) Magnification Prediction: Histopathology images are often generated and viewed at various standard magnification levels. Considering an image of fixed size, higher magnifications provide more details but less context, whereas lower magnifications allow less details but more context of tissue region. Pathologists assessing an image tend to infer important semantic information by iterating between detail and context – i.e., by zooming in and out on WSIs or by looking at different magnification levels [1]. In other words, magnification levels are implicitly correlated with important semantic information. Therefore, to enable the classification model to learn semantic information, we set up a pretext task focused on estimating magnification level of the image. Specifically, the pretext task focuses on classifying the input image to 1 of 4 magnification levels (40×, 20×, 10× and 5×). We extract images or patches from WSIs at these magnification levels Figure 2 (A). If a magnification level is not available, we obtain the patches by (bi-linear) resizing patches from other magnification levels that are available. For example, to obtain 128 × 128 patches at 5×, we extract patches of 1024 × 1024 at 40× and downsample by factor of 4. We then feed the extracted images to the network, which learns by minimizing a cross-entropy objective function.

2) Solving Magnification Puzzle (JigMag): A basic problem in pattern recognition is the jigsaw task of retrieving an original image from its shuffled parts [36]. Convolutional neural networks (CNNs) have been employed to solve the jigsaw puzzle [7]. To solve the jigsaw puzzle, it is known that the network should learn the global semantic representation of images. This is achieved by concentrating on the differences between tiles and their positions while avoiding low level statistics [7]. In histopathology, objects are smaller compared to natural scene images, and there is no specific ordering among the objects. For example, the relative positions of different parts of dog in a natural scene image is consistent, however we do not have a similar concept in histopathology. Therefore, solving the jigsaw puzzle is by itself not sufficient

[1]Magnification levels and their corresponding resolutions vary for each scanner. However by observing one particular magnification of an image, other magnifications can be perceived easily for the same scanner.

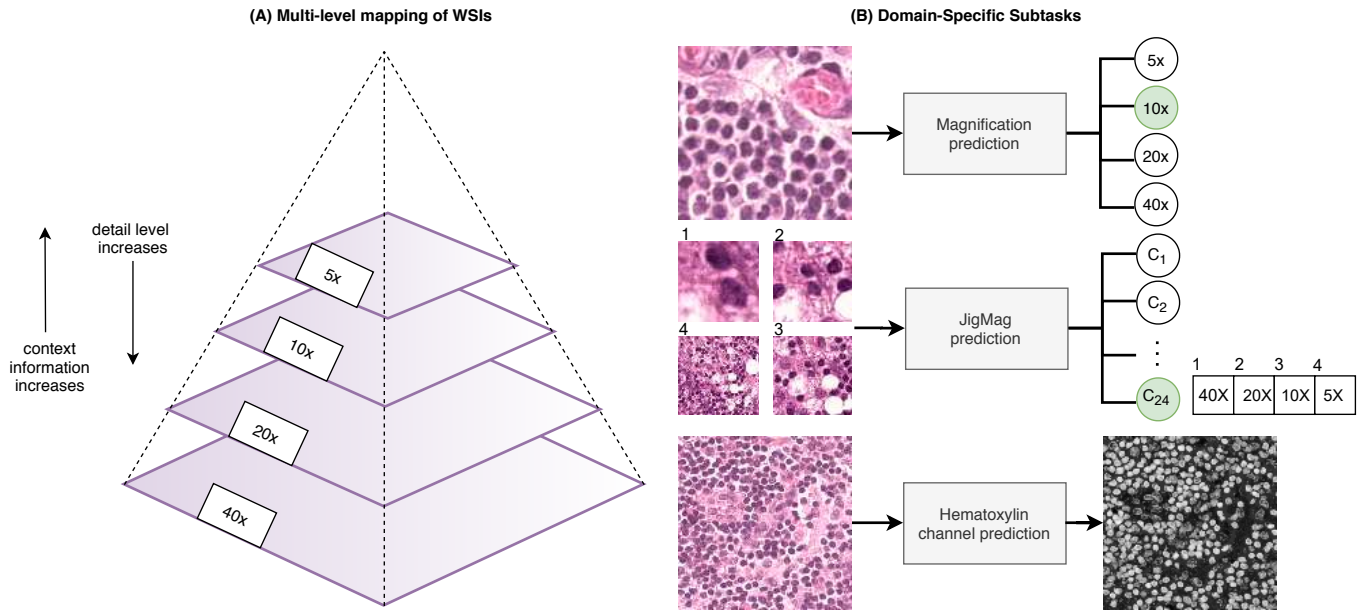

Fig. 2. (A) Whole slide images (WSI) in pathology slides organized hierarchically - each level trades-off the degree of detail against the availability of contextual information. (B) Pathology specific pretext tasks created for Self-Path.

for learning useful semantic representations in histopathology.

Instead, we propose to create a puzzle to reflect the magnification and context characteristics of histopathology images. Conceptually, classification can be enhanced by having the network implicitly learn object size and associated contextual information. Hence, we propose a pretext task focused on solving this magnification and context puzzle. In this puzzle, an image consists of image tiles with various magnifications and the network is tasked with predicting their arrangement. This set up caters also to the need to classify images containing objects with varying shapes and sizes.

Specifically, we define v as a vector of image orders in a 2×2 grid where each grid includes a specific magnification. For example $v = [0, 1, 2, 3]$ defines that image with magnification $5\times$ is on top left corner, $10\times$ is on top right and so on. We consider 24 different orders of magnification. To construct our proposed jigsaw puzzle, we first extract patches of size 512×512 at $40\times$ magnification then each part of the puzzle is constructed by down-sampling and or center-cropping to the size of 64×64, where each reflects specific context and resolution of the the original extracted patch. This pretext task employs a cross entropy loss function.

3) Hematoxylin Channel Prediction: Commonly, histopathology images are stained with Hematoxylin and Eosin (H&E). In H&E images, hematoxylin turns the palish color of nuclei to blue and eosin changes the color of other contents to pink. Color deconvolution methods have been applied to specifically identify cell nuclei in H&E images. Therefore by extracting hematoxylin channel, one can locate the nuclei and their approximate shape. Pathologists often use the location, shape and morphology of nuclei in the hematoxylin channel to diagnose or classify histopathology images (especially for malignant features).

Therefore, one way to enhance learning of useful representations is to enable the classifier to identify the nuclei and their associated characteristics. We choose to define a pretext task focused on predicting the hematoxylin channel from H&E. We use the approach in [37] to extract the hematoxylin channel in our images and define the ground truth for the self-supervision task. We scale the values of hematoxylin channel in the range [0,1] and employ a mean absolute loss for optimizing this task.

D. Pathology-agnostic Self-supervision Tasks

The literature has investigated various pretext tasks like rotation prediction, flipping, image reconstruction [8], [29]. These were however, not tailored for pathology data. Here, we systematically study and benchmark efficacy of these pretext tasks for semi-supervised learning and domain adaptation in histopathology applications.

1) Prediction of Image Rotation: For predicting rotation, the input image is rotated with degrees of $0°$, $90°$, $180°$ and $270°$ corresponding to the labels 0, 1, 2 and 3, respectively [29].

2) Prediction of Image Flipping: The label assigned to the horizontal flipping of image is 1 and 0 if not flipped.

3) Image Reconstruction with Autoencoder: For reconstructing the image, a convolutional decoder is used on top of the feature extractor [26], similar to one for predicting hematoxylin channel however 3 channels is considered for output.

4) Real vs Fake Prediction (Generative): The generative learning literature has shown that predicting whether an image is real or fake can help to learn useful representations for classification [38]. Therefore, we introduce a generative pretext task focused on real vs. fake prediction. To learn this pretext task, we train a generative network in an adversarial fashion by using unlabeled samples. While one could use a shared encoder to extract features, we found that it is easier to employ a simpler encoder/discriminator similar to the generative adversarial network (GAN) in [38].

Formally, real images are drawn from distribution D_{real}, and the generative function learns the distribution D_{gen} where the goal is to align this two distributions ($D_{gen} \sim D_{real}$). The generator $G(.)$ takes predefined noise variables z from a uniform distribution D_{noise}. The objective function is defined as:

$$\begin{aligned} L_{dis} &= -\mathrm{E}_{x\sim\ D_{real}}[\log[1 - F_{Dis}(F_e(x))]] \\ &-\mathrm{E}_{x\sim\ D_{gen}}[\log[F_{Dis}(F_e(x))]] \\ L_{gen} &= \|\mathrm{E}_{x\sim D_{real}}[F_e(x)] - \mathrm{E}_{z\sim D_{noise}}[F_e(G(z))]\|_1 \end{aligned} \quad , \quad (2)$$

where L_{gen} and L_{dis} are the generator and discriminator losses, respectively. $F_e(x)$ is the feature from intermediate layer of feature extractor (last layer before fully connected layers) and $F_{Dis}(F_e(x))$ is the output of the discriminator (fake/real head).

5) Domain Prediction: In order to learn useful representations to facilitate domain adaptation, it is useful to have a network learn the common features between source and target domains. Therefore, we introduce a pretext task to predict if the image belongs to source or target domain, and employ it in combination with other pretext tasks for the domain adaptation experiments.

For this pretext task, we employ a domain adversarial neural network (DANN) [20]. DANN includes a minimax game where discriminator H_d (domain prediction head) is trained to distinguish between the source and target domain, and the feature extractor is simultaneously trained to confuse the discriminator. Therefore, to extract the common or domain-invariant features, the parameters of feature extractor θ_e (shared encoder in the multi-task setup) are learned by maximizing the loss of domain discriminator L_d, while parameters of the domain discriminator are learned by minimizing the loss of domain discriminator. Parameters of the main task F_c are also minimized to ensure good performance on the main task. Formally:

$$\begin{aligned} \underset{\theta_c,\theta_e}{\operatorname{argminmax}}\ \underset{\theta_d}{} \frac{1}{\eta_s}\sum_{i=0}^{\eta_s} L_c(F_c^{\theta_c}(F_e^{\theta_e}(\mathbf{x}_i^s)),\ y_i) + \\ -\frac{\alpha_d}{\eta_s+\eta_t}(\sum_{i=1}^{\eta_s+\eta_t} L_d(F_d^{\theta_d}(F_e^{\theta_e}(\mathbf{x}_i)),\ d_i) \end{aligned} \quad , \quad (3)$$

where d_i is the domain label for $\mathbf{x}_i$ and α_d is a coefficient for discriminator loss. In practice, we apply domain confusion using the Gradient Reversal Layer (GRL), where the gradients of L_d with respect to the gradients of feature extractor parameters θ_e ($\frac{\partial L_d}{\partial \theta_e}$) are reversed during back-propagation.

IV. Experiments

A. Datasets

1) Camelyon16: We used the Camelyon 16 challenge dataset [39] that contains 399 H&E stained WSIs obtained on patients with breast cancer metastasis in the lymph nodes. The WSIs were acquired from 2 different centers, namely: Radboud University Medical Center (RUMC) and University Medical Center Utrecht (UMCU). RUMC images were generated by a digital slide scanner (Pannoramic 250 Flash ; 3DHISTECH) with a 20× objective lens (0.243 μm × 0.243 μm) and UMCU images were produced using a digital slide scanner (NanoZoomer-XR Digital slide scanner C12000-01; Hamamatsu Photonics) with a 40× objective lens (0.226μm × 0.226 μm). The tumor regions are exhaustively annotated by pathologists. We used the official training and testing splits comprising 270 and 129 WSIs, respectively. We randomly sampled 34 WSIs of the training set for validation. For our experiments, we randomly extracted patches from both normal and tumor regions (Table I).

TABLE I
Number of WSIs and patches in each dataset.

		Train	Validation	Test
Camelyon16	WSIs	236	34	129
	patches	67054	15586	16562
LNM-OSCC	WSIs	100	14	103
	patches	55416	7224	14472
Kather	patches	79994	20006	7180

2) LNM-OSCC: LNM-OSCC is an in-house dataset comprising 217 H&E WSIs obtained on patients with Oral Squamous Cell Carcinoma (OSCC). Of these 217 patients, 140 have metastases in the cervical lymph nodes and 77 do not manifest metastases in the cervical lymph nodes. The WSIs were acquired from 2 hospitals using 2 different scanners – (a) 98 WSIs scanned with 40× objective lens using IntelliSite Ultra Fast Scanner (0.25 μm/pixel) at University Hospital Conventry and Warwickshire (UHCW), and (b) 119 WSIs scanned at the School of Medical Dentistry in Sheffield University by Aperio/Leica CS2 with 20× objective lens (0.2467 μm/pixel). The training set comprises 100 WSIs, the validation set 14 WSIs and testing set 103 WSIs. For those cases in the training and validation sets that have metastases, a sampling of the tumor and normal regions were delineated with bounding box annotations by pathologists. For the testing set, the tumor regions were exhaustively annotated at the pixel-level.

3) Kather: This dataset contains 107,180 image patches from H&E stained WSIs comprising human colorectal cancer (CRC) and normal tissue. For this dataset, only patches were available (no WSIs).The dataset covers 9 tissue classes: Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM). We used the official data splits comprising 100k patches for training and 7180 patches for testing. We randomly sampled 20k patches of the training set for validation.

B. Data Summary

Figure 3 shows some illustrative examples of the different datasets used in our study. The overall data statistics are shown in Table I. For Camleyon16 and LNM-OSCC datasets, we extracted patches from the WSIs, and patches are distributed equally for each class. For our main task the patch extraction size is 128 × 128 at 10×. The Kather dataset patches are sized 224 × 224 and we resized to 128 × 128 for our experiments.

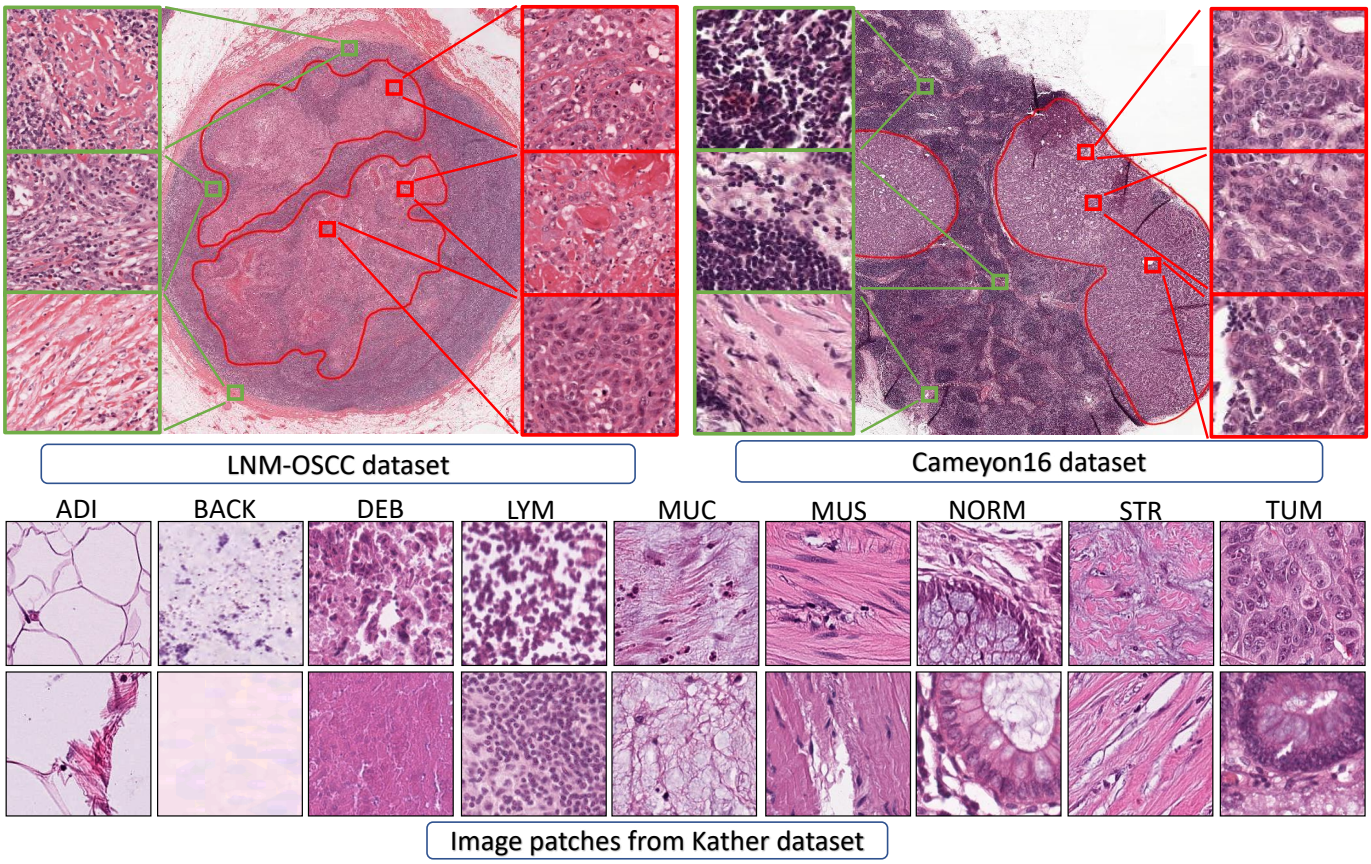

Fig. 3. Exemplar images of different datasets that are used in this study. Red and green boxes denote the tumor and normal image patches.

C. Experimental Setup

1) Networks: We chose Resnet50 [40] as the feature extraction backbone for all our experiments. The classifier head consists of adaptive average pooling which is followed by fully connected layer and softmax. The decoder head for reconstructing image and predicting hematoxylin channel is similar to the UNet decoder [41] (Supplementary Material) without using any skip connections. While using the real vs fake pretext task for image generation, we utilize the architecture presented in [38] (Supplementary Material) and find that this simpler feature extractor allows easy and robust convergence for the image generator.

2) Implementation Details: When Resnet50 is used as the shared encoder, we trained the network for 200 epochs. Our experiments used batch size 64, Adam optimizer, and learning rate of 10^{-3}. We fed batches of labeled and unlabeled images to the network separately. Therefore an epoch is defined as one full step through all the unlabeled images. Since our self-supervised experiments utilize fewer labeled images than unlabeled images, the labeled images are repeated in an epoch. Experiments related to real vs fake prediction used number of epochs and batch size of 500 and 32, respectively; and employed Adam optimizer with learning rate of 3×10^{-4}. For training model in multitask setup, we separately input batches of images for each task to the network and then sum their losses with their corresponding weights. Finally we back-propagate the whole loss through the network.

D. Results of Semi-Supervised Experiments

Here, we compare the effect of different self-supervision tasks for semi-supervised learning. We compare our models against the popular semi-supervised benchmarks, namely Mean Teacher [2] and VAT [3]. We also compare with teacher-student chain [17] (TSchain). TSchain is a recent semi-supervised approach for histopathology domain, that predicts the pseudo-labels for the unlabeled data and then uses all images for iteratively retraining the model. For performance evaluations, we follow the typical protocol of varying the annotation budget for the training set while maintaining a fixed validation set, and reporting AUCs (average across 3 seeds) on the test set.

1) Results for LNM-OSCC Dataset: We report performance of each of the self-supervised tasks on LNM-OSCC dataset in Table II. We have evaluated the model performance in terms of AUROC (Area Under the Receiver Operating Characteristic) for different annotation budgets (1%, 4%, 5%, 10% and 20% of the available WSIs). The semi-supervised approaches train on a combination of the labeled and unlabeled WSIs. The supervised baseline is only trained on labeled images without utilizing any unlabeled images.

We observe from Table II that at very low annotation budgets, pathology specific self-supervised tasks outperform the baselines and the pathology agnostic self-supervised tasks. For instance, at annotation budgets of 1% (1 labeled WSI, 134 labeled patches) and 4% (4 labeled WSIs, 1120 labeled

TABLE II
LNM-OSCC RESULTS FOR DIFFERENT ANNOTATION BUDGETS. ANNOTATION BUDGET IS DEFINED AS THE PERCENTAGE OF AVAILABLE WSIS THAT ARE LABELED. THE NUMBER OF PATCHES ASSOCIATED WITH EACH BUDGET ARE INDICATED IN THE PARENTHESES. THE SUPERVISED UPPER BOUND PERFORMANCE WHEN USING ALL LABELED DATA IS 98.4%.

% Labeled WSIs (No. Patches)	1%(134)	2%(1024)	5%(1880)	10%(3334)	20%(7558)
	AUROC(%)	AUROC(%)	AUROC(%)	AUROC(%)	AUROC(%)
	Baselines				
supervised baseline	73.4 ± 2.0	76.1 ± 5.3	85.3 ± 6.3	86.3 ± 2.7	96.3 ± 0.3
mean teacher [2]	75.1 ± 4.5	78.4 ± 5.6	86.2 ± 7.6	91.4 ± 1.2	**97.4 ± 0.3**
VAT [3]	74.5 ± 5.6	77.4 ± 3.3	85.3 ± 4.3	92.1 ± 1.2	96.5 ± 0.9
TS chain [17]	75.3 ± 2.4	79.3 ± 2.5	85.2 ± 3.1	94.1 ± 1.7	97.2 ± 0.2
	Pathology-Agnostic Self-supervised Tasks				
rotation	74.5 ± 5.6	76.3 ± 4.2	88.4 ± 1.5	93.2 ± 0.3	96.2 ± 0.1
flipping	74.6 ± 4.0	74.2 ± 5.3	85.3 ± 4.1	91.4 ± 0.4	94.2 ± 0.4
autoencoder	73.0 ± 6.5	75.1 ± 3.5	84.2 ± 3.3	90.3 ± 1.5	94.3 ± 0.2
generative	73.4 ± 7.1	79.3 ± 4.1	**90.3 ± 2.4**	**95.4 ± 0.2**	97.1 ± 0.3
	Pathology-Specific Self-supervised Tasks				
magnification	76.3 ± 4.0	76.6 ± 3.6	87.4 ± 2.3	92.5 ± 0.2	94.1 ± 0.4
JigMag	**80.6 ± 3.5**	**81.8 ± 5.3**	89.5 ± 5.4	92.4 ± 0.5	96.5 ± 0.2
hematoxylin	75.3 ± 7.6	80.2 ± 5.3	87.5 ± 1.2	94.4 ± 1.3	**97.4 ± 0.5**
Best self-supervised	**80.6 ± 3.5**	**81.8 ± 5.3**	**90.3 ± 2.4**	**95.4 ± 0.2**	**97.4 ± 0.5**

patches), JigMag task has the best performance. At annotation budgets of 1% and 2%, Hematoxylin and magnification tasks outperform pathology agnostic tasks and generative tasks. When annotation budget increases to 10%, we observe that the generative task performs much better (AUC 95.4%), suggesting that the generated images can help the classifier to boost the performance. Overall, our LNM-OSCC experiments suggest that for limited annotation budgets, pathology specific pretext tasks are helpful for enhancing the model performance, with JigMag outperforming other approaches.

2) Results for Camelyon16 Dataset: We report performance of each of the self-supervised tasks on Camelyon16 dataset in Table III. We have evaluated the model performance in terms of AUROC (Area Under the Receiver Operating Characteristic) for different annotation budgets (1%, 2%, 5%, 10% and 20% of the available WSIs). The semi-supervised approaches train on a combination of the labeled and unlabeled WSIs. The supervised baseline is only trained on labeled images without utilizing any unlabeled images.

Similar to LNM-OSCC dataset, pathology specific tasks outperform other semi supervised methods. In particular, the JigMag task improves the performance over the supervised baseline by 13.4%, 11.8% and 6.2% at 1% (2 WSIs), 2% (4 WSIs) and 5% (8 WSIs) annotation budgets, respectively. At 1% annotation budget, only magnification and JigMag outperform mean teacher and supervised baseline. Unlike LNM-OSCC, the generative model cannot achieve highest AUROC for any annotation budget, but it's performance is competitive with mean teacher and VAT. Similar to LNM-OSCC, JigMag could achieve highest performance overall, and the main boost is obtained at very low annotation budgets.

3) Results for Kather Dataset: We report performance of each of the self-supervised tasks on Kather dataset in Table IV. Since there are 9 classes in the Kather dataset, Macro AUROC is used for evaluation of classification performance. Unlike the other 2 datasets, only patches were available for this dataset, therefore the annotation budget only reflects the proportion of the overall patches that is labeled. Further, we observe that at 2% annotation budget, the performance of supervised baseline is still high (Macro AUC of 98%). Hence using semi-supervised approaches would not add much benefit. Hence, we focus on the very low annotation budget regime where some degradation of Macro-AUC can be observed for supervised model – i.e., annotation budgets of 0.1%(100 labeled) and at 1% (800 labeled images). Moreover, as this dataset does not include WSIs, we were unable to extract large patches or patches at different magnificationsand hence could not evaluate JigMag and magnification self-supervised tasks on this dataset.

From Table IV, we observe that at 0.1% annotation budget, predicting hematoxylin channel as a self-supervised task improves the performance by 2.8% and 1.2% compared to the baseline and mean teacher, respectively. At 1% annotation budget, we see that the various self-supervised tasks can again improve performance compared to the baseline. Predicting hematoxylin channel can also give the superior performance, suggesting that the prediction of rough nuclear segmentations can be helpful for semi-supervised learning.

E. Domain Adaptation Experiments

We conduct two domain transfer experiments, (i) Camelyon16 to LNM-OSCC (Cam16→LNM-OSCC) and (ii) LNM-OSCC to Camelyon16 (LNM-OSCC→Cam16). In both cases, we do unsupervised domain transfer, where the source is the labeled set and the target set is completely unlabeled.

We evaluate our approach against the naive supervised baseline, and two other domain adaptation methods WDGRL [21] and DANN [20]. The supervised baseline employs Resnet50 and is trained with source domain data only. WDGRL trains a domain critic network to estimate the Wasserstein distance between the source and target feature representations. The feature extractor network will then be optimized to minimize the estimated Wasserstein distance in an adversarial manner. By iterative adversarial training, WDGRL learns feature representations invariant to the covariate shift between domains. DANN is a domain prediction approach based on the GRL unit and was mentioned in Section III-D.

We report the results obtained with Self-Path (using different pretext tasks) and the comparisons with the supervised and

TABLE III

CAMELYON16 RESULTS FOR DIFFERENT ANNOTATION BUDGETS. ANNOTATION BUDGET IS DEFINED AS THE PERCENTAGE OF AVAILABLE WSIS THAT ARE LABELED. THE NUMBER OF PATCHES ASSOCIATED WITH EACH BUDGET ARE INDICATED IN THE PARENTHESES. THE SUPERVISED UPPER BOUND PERFORMANCE WHEN USING ALL LABELED DATA IS 94.2%.

Labeled WSIs (No. Patches)	1%(600)	2%(1000)	5%(2600)	10%(6400)	20%(13540)
	AUROC(%)	AUROC(%)	AUROC(%)	AUROC(%)	AUROC(%)
	Baselines				
supervised baseline	68.3 ± 5.1	74.5 ± 5.8	81.2 ± 2.5	88.4 ± 2.3	92.1 ± 0.5
Mean Teacher [2]	73.7 ± 3.8	78.5 ± 2.6	84.5 ± 2.4	92.7 ± 1.9	93.1 ± 0.9
VAT [3]	70.9 ± 5.8	77.4 ± 3.3	81.3 ± 5.2	90.3 ± 2.3	92.8 ± 1.5
TS chain [17]	74.9 ± 6.9	76.9 ± 3.2	83.8 ± 2.1	93.1 ±2.5	**93.9 ± 1.3**
	Pathology-Agnostic Self-supervised Tasks				
rotation	69.8 ± 4.8	74.5 ± 3.1	80.4 ± 2.5	90.1 ± 2.0	92.4 ± 2.5
flipping	70.2 ± 6.2	75.4 ± 3.5	81.6 ± 5.1	89.4 ± 0.6	92.3 ± 1.6
autoencoder	70.1 ± 2.4	75.6 ± 4.1	82.3 ± 4.5	90.5 ± 2.3	92.4 ± 1.1
generative	72.5 ± 5.5	77.6 ± 5.4	82.4 ± 7.2	92.6 ± 3.2	93.6 ± 1.5
	Pathology-Specific Self-Supervised Tasks				
magnification	77.5 ± 3.1	84.6 ± 5.2	85.1 ± 3.6	**93.2 ± 3.4**	93.4 ± 2.5
JigMag	**81.7 ± 3.8**	**86.3 ± 5.2**	**87.4 ± 4.5**	90.6 ± 4.6	92.8 ± 2.4
hematoxylin	72.8 ± 4.6	78.3 ± 4.5	84.6 ± 3.4	92.3 ± 4.1	**93.7 ± 2.5**
Best Self-supervised	**81.7 ± 3.8**	**86.3 ± 5.2**	**87.4 ± 4.5**	**93.2 ± 3.4**	**93.7 ± 2.5**

TABLE IV

KATHER RESULTS FOR DIFFERENT ANNOTATION BUDGETS. ANNOTATION BUDGET IS DEFINED AS THE PERCENTAGE OF AVAILABLE WSIS THAT ARE LABELED. THE NUMBER OF PATCHES ASSOCIATED WITH EACH BUDGET ARE INDICATED IN THE PARENTHESES. THE SUPERVISED UPPER BOUND PERFORMANCE WHEN USING ALL LABELED DATA IS 99.4%.

Labeled WSIs (No. Patches)	0.1%(100)	1%(800)
	AUROC(%)	AUROC(%)
	Baselines	
supervised baseline	87.5 ± 2.0	92.5 ± 1.2
mean teacher [2]	89.1 ± 1.5	93.9 ± 0.3
VAT [3]	88.5 ± 1.4	92.6 ± 0.4
TS chain [17]	88.9 ± 0.3	93.5 ± 0.2
	Self-supervised tasks	
generative	88.4 ± 3.5	92.3 ± 2.6
rotation	87.4 ± 1.6	93.3 ± 0.4
flipping	88.6 ± 0.8	93.0 ± 0.9
autoencoder	89.3 ± 1.3	94.3 ± 1.2
hematoxylin	90.3 ± 0.7	95.1 ± 0.5
Best self-supervised	**90.3 ± 0.7**	**95.1 ± 0.5**

TABLE V

AUROC RESULTS FOR DOMAIN ADAPTATION

	Cam16→LNM-OSCC	LNM-OSCC→Cam16
	Baselines	
supervised baseline	79.53 ± 0.2	63.73 ± 0.5
DANN	89.23 ± 1.5	71.15 ± 0.6
WDGRL	89.64 ± 2.6	72.65 ± 2.2
	Pathology-Agnostic Self-supervised Tasks	
rotation	86.14 ± 3.4	66.91 ± 4.1
flipping	82.14 ± 3.6	65.95 ± 4.4
autoencoder	89.90 ± 2.8	71.62 ± 2.6
generative	91.54 ± 3.5	74.14 ± 2.7
	Pathology-Specific Self-supervised Tasks	
magnification	89.69 ± 3.6	73.62 ± 4.1
JigMag	92.34 ± 4.4	74.51 ± 3.6
hematoxylin	90.47 ± 4.5	73.24 ± 3.8
mag+hem+JigMag	92.85 ± 3.6	74.95 ± 3.5

domain adaptation baselines in Table V. We observe that the pathology-specific pretext tasks can help the model outperform the baseline by a large margin. For Cam16→LNM-OSCC, the pathology-specific pretext tasks provide more than 10% boost in AUROC over the supervised baseline. The combination of all pathology specific pretext tasks achieves the best performance. Amongst the individual pretext tasks, JigMag achieves the best performance (∼2% better than DANN and WDGRL). Further, we note that the pathology agnostic generative model also performs well – with 1.9% higher AUROC than WDGRL and 11% higher AUROC over the supervised baseline. This suggests that the images from the generator can contribute to learning useful domain-invariant features as well. We see similar trends for LNM-OSCC→Cam16 – where again combining pathology specific tasks has the best performance and JigMag provides the second best performance. We highlight that we have used domain prediction with GRL layer in all non-generative methods as it improves the performance. Generative models, owing to adversarial training can still achieve very high performance, even without GRL.

1) WSI Analysis: While the results thus far are reported at the patch level, it is also useful to consider the WSI-level performance. For the Cam16→LNM-OSCC domain adaptation task, we now report the WSI-level results for the top two best performing Self-Path settings i.e., combination of all pathology specific pretext tasks and JigMag pretext task. We also provide comparisons with the supervised baseline (source only), WDGRL, and the pathology agnostic generative pretext task.

In order to quantify WSI-level performance, we aggregate patches belonging to a WSI and construct a WSI-level heat map based on the patch level predictions. For heat map generation, there are two steps. First, we extract patches of 128 × 128 at 10× magnification with overlap of 50% from tissue regions of WSIs. Second, we aggregate the prediction of each patch together to build the final heat map of WSIs. We then post-process these heat maps to obtain the WSI-level prediction. The post-processing steps are uniform for all models in this section, and as follows: we extract 10 morphological and geometrical features from objects within binarized heat map at three thresholds of 0.25, 0.5 and 0.9. Then we calculate the mean, stddev, minimum and maximum of object features for each WSI. Therefore, in total we use 120 features for constructing feature vectors. Afterwards, we

TABLE VI
CAM16 → LNM-OSCC DOMAIN ADAPTATION RESULTS ON THE WSI-LEVEL. THE UPPER BOUND PERFORMANCE USING ALL LABELS FOR TARGET DOMAIN IN SUPERVISED FASHION IS 93.3%.

	AUROC(%)	Average Precision(%)
supervised baseline (source only)	75.2	81.7
WDGRL	85.8	91.6
generative	90.4	95.2
JigMag	91.6	96.7
mag+JigMag+hem	91.6	96.3

employ the random forest algorithm for classification of the features. Finally, we evaluate the model on the test set of LNM-OSCC.

The results are shown in Table VI. The supervised baseline has WSI-level AUROC of 75.2% whereas Self-Path with JigMag pretext task and Self-Path with the combination of all pathology specific pretext tasks each improve the performance by 16.4%. Further, we note that Self-Path with JigMag improves performance over WDGRL by 2% at the patch-level and a ∼6% improvement at the WSI-level. This suggests that the magnification puzzle and the pretext tasks that can help learn from various image resolutions in a self-supervised manner enable strong performance boost at WSI-level (beyond patch-level).

These improvements are also evident in the WSIs overlaid with the heatmaps, as visualized in Figure 4. This figure shows that the supervised baseline (source only) model (middle column) has many false negatives and often misses tumor regions. However, WDGRL, Self-Path with JigMag, and Self-Path with generative pretext task can all increase true positives while decreasing false negatives. We note that WDGRL and Self-Path with generative pretext task do not perform as well as Self-Path with JigMag - mainly because they suffer larger number of false positives at the patch-level classification.

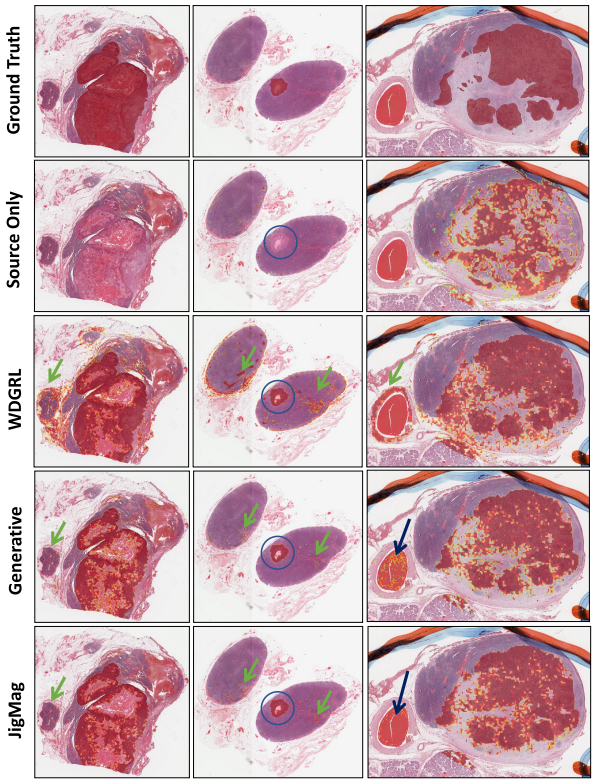

Fig. 4. Three WSI samples and their overlaid heatmaps. from top to bottom, first row: the overlaid ground-truth mask, second row: overlaid heat map of model predictions when it is trained using only Camlelyon16 data, third row: Overlaid heatmap of WDGRL predictions, fourth row depicts the overlaid predictions of Self-path using generative task and the last row shows the heatmaps generated Self-path using JigMag task. The circle indicates a region which is missed using the supervised baseline (source only) model and green arrows point to the false positive regions generated by WDGRL where using generative task and JigMag task eliminate those regions. Black arrow also shows regions that are misclassified by generative model but are correctly classified as normal regions by Jig-Mag. (Best viewed in color, zoom in to see more details)

V. DISCUSSION

In this section we describe sensitivity analyses and discuss the model performance by changing the values of loss weights, decreasing the annotation budget and combing all pathology specific tasks. Moreover, we conduct an experiment to show the usefulness of transfer learning using our proposed self-supervised tasks. For following experiments, we choose Camelyon dataset. Since the variation of hyperparameters are studied, it is expected that these trends will be similar on other dataset.

A. Effect of Loss Weight for Each Task

We consider the task of training with 1% of annotation budget on Camelyon16 dataset. To understand the effect of loss weights for each pretext task, we experiment with different values of α and show the results in Table VII. Overall, assigning more weights on each task shows better performance. More precisely, when α is set to 1, maximum value of AUROC is obtained. Therefore we can conclude when we are using only one pretext task, the pretext task and the main task should have similar weight to be effective for semi-supervised learning. The optimum value of α may change when we use all tasks together which we investigate in the next section. In here, by choosing the alpha values greater than one, the pretext task will be dominant. Therefore the main task does not learn discriminant features for separating the classes. Moreover, we are interested to see the values of alpha up to one (when it is similar to the main task).

B. Combining tasks

We now evaluate the effect of the loss weights (α's) when combining all pathology specific tasks. We consider the task of training with 1% and 2% of annotation budget on Camelyon16 dataset, and experiment with different combinations of loss coefficients. The results, in Table VIII, suggest that assigning high weights (similar to main task) to all pretext tasks can degrade the performance. For example, if all tasks are given $\alpha = 1$, overall the weights for pretext tasks would be $3\times$ more than the main task which would cause drop in performance. However by assigning smaller weight values for each task, we can achieve better performance. Particularly, best performance

TABLE VII
AUROC PERFORMANCE OF PATHOLOGY SPECIFIC TASKS WITH DIFFERENT VALUES OF α ON CAMELYON16 DATASET.

α	magnification	JigMag	hematoxylin
1	77.5 ± 3.1	81.7 ± 3.8	72.8 ± 4.6
0.8	77.1 ± 2.8	81.5 ± 3.4	71.3 ± 2.4
0.6	76.4 ± 4.0	78.8 ± 2.6	70.2 ± 3.5
0.5	74.6 ± 3.4	78.4 ± 2.4	70.3 ± 4.6
0.2	72.5 ± 3.7	74.1 ± 4.6	69.5 ± 4.4

TABLE VIII
USING ALL PATHOLOGY SPECIFIC TASKS FOR SEMI-SUPERVISED LEARNING ON CAMELYON16 DATASET. α_{mag}, α_{JigMag} AND α_{hem} INDICATE THE LOSS COEFFICIENT FOR MAGNIFICATION, JIGMAG AND HEMATOXYLIN TASKS, RESPECTIVELY.

α_{mag}	α_{JigMag}	α_{hem}	1%	2%
1	1	1	79.1 ± 4.5	83.5 ± 5.1
0.25	0.5	0.25	83.2 ± 4.3	86.3 ± 5.3
0.5	0.25	0.25	80.2 ± 2.5	85.4 ± 3.1
0.25	0.25	0.5	79.6 ± 2.7	84.3 ± 5.5
0.25	0.25	0.25	80.3 ± 3.4	85.5 ± 1.8

is obtained when more weight is assigned to JigMag task and lower weights to Hematoxylin and magnification tasks. This is in line with previous experiments which showed that JigMag had better performance as compared to other tasks. We can, therefore, recommend that a good strategy can be to start with heavy weight to JigMag for computational pathology tasks before combining it with other self-supervision tasks.

C. Performance at Very Low Annotation Budget

In section IV-D, we evaluated the performance of self-supervised tasks with different annotation budgets. we observed, despite high boost in performance by applying self-supervised tasks, the supervised baseline also gives reasonable results (e.g., 73.4% on LNM-OSCC for 134 patches). To assess performance at even lower annotation budget, we further decreased number of patches annotated (while maintaining the same number of WSIs) to 50 for LNM-OSCC and Camelyon datasets. As shown in Table IX, Self-Path with pathology-specific pretext tasks can improve the AUC by about 10% over the supervised baseline. Again, the JigMag pretext task is the best performing pretext task. Moreover, we also note that combining all pathology specific tasks (with loss weights 0.25, 0.25 and 0.5 for hematoxylin, magnification and JigMag respectively) can result in even better performance.

D. Transfer Learning

We finally investigate the usefulness of the representations learned by Self-Path for related tasks. For this, we conduct a transfer learning experiment using Camelyon16 dataset. We first train Self-Path with each self-supervised pretext task on the entire dataset, and then fine-tune the backbone (the model excluding the final linear layer/decoder) for the main task. We compare the performance against the naive method of training the network from scratch with random weight initializations (Scratch). The results for different pretext tasks at varying annotation budgets are shown in Table X. We can see that the representations learned by Self-Path with transfer

TABLE IX
AUROC RESULTS FOR VERY LOW BUDGET OF ANNOTATION:HERE ONLY 25 IMAGE PATCHES ARE USED IN EACH CLASS

	Camelyon16	LNM-OSCC
	Baselines	
supervised baseline	55.3 ± 5.1	54.8 ± 8.1
mean Teacher	65.4 ± 4.8	60.4 ± 5.4
VAT	64.3 ± 6.4	58.6 ± 6.5
TS chain	62.4 ± 10.6	59.4 ± 7.7
	Pathology-Agnostic Self-supervised Tasks	
rotation	62.6 ± 4.6	58.7 ± 4.6
flipping	65.7 ± 9.3	58.9 ± 5.3
autoencoder	65.1 ± 6.4	59.6 ± 4.3
generative	64.2 ± 5.7	60.1 ± 10.3
	Pathology-Specific Self-supervised Tasks	
magnification	65.3 ± 7.5	62.2 ± 6.7
JigMag	**66.2 ± 6.4**	**63.5 ± 7.9**
hematoxylin	64.2 ± 7.4	62.4 ± 4.6
mag+hem+JigMag	**66.5 ± 5.5**	**64.1 ± 5.5**

TABLE X
RESULTS OF TRANSFER LEARNING OF SELF-SUPERVISED TASKS WITH DIFFERENT BUDGET OF ANNOTATIONS USING CAMELYON16 DATASET.

	1%	2%	5%	10%	20%
Scratch	68.3	74.5	81.2	88.4	92.1
magnification	72.6	77.4	84.8	89.9	92.2
JigMag	73.3	79.4	85.8	90.4	92.7
hematoxylin	72.9	79.5	85.9	88.6	92.3

learning enable performance improvement over 'Scratch' in each case. Again, Self-Path with JigMag achieves the best performance. The improvements with fine-tuning is largest in the low annotation regime, and drops off when more annotated data are available. These results suggest that the pretext tasks in Self-Path enable learning of useful representations. Overall, with annotation budget of over 20%, fine-tuning gives the same result as training from scratch. Therefore multi-task approach where self-supervision task and main task are trained together leads to better results than fine-tuning. Therefore multi-task approach where self-supervision task and main task are trained together leads to better results than fine-tuning. This phenomenon is also shown by [42].

VI. CONCLUSIONS

In this paper, we proposed Self-Path – a generic framework based on self-supervision tasks for histopathology image classification – to address the challenge of limited annotations in the area of computational pathology. We introduced 3 novel self-supervision tasks to cater to the contextual, multi-resolution and semantic features in pathology images. We showed that such pathology specific self-supervision tasks can improve the classification performance for both semi-supervised learning and domain adaptation. Moreover, we thoroughly investigated general self-supervised approaches such as generative models within this pipeline and showed that using the pathology-specific tasks, despite being simple and easy to implement, can improve performance over generic self-supervision in many scenarios involving limited annotation budget or domain shift. In particular, we note that the JigMag self-supervision can be extremely helpful when the amount of labeled data is very small. Unlike baseline methods that are highly dependent on hyperparameters values, our method can

achieve good performance without exhaustive hyperparameter tuning. Self-Path can be applied to other problems in computational pathology, where annotation budget is often limited or large amounts of unlabeled image data are available. In our sensitivity analyses, we considered only domain specific tasks and showed that their combination leads to better performance compared to using only one pretext task in the multitask setup. Using all domain agnostic task as pretext task can also potentially increase the performance and requires further exploration. Other future directions include employing other self-supervision tasks (such as predicting the Eosin channel or a combination of Hematoxylin and Eosin after estimating the two channels, rather than keeping them fixed), increasing the number of magnification levels, increasing the JigMag grids to incorporate wider and more complex puzzles for the network to solve, exploring different variations of orders for JigMag (here all 24 orders were used) and a deeper investigation into other domain adaptation tasks.

REFERENCES

[1] O.-J. Skrede, S. De Raedt, A. Kleppe, T. S. Hveem, K. Liestøl, J. Maddison, H. A. Askautrud, M. Pradhan, J. A. Nesheim, F. Albregtsen, *et al.*, "Deep learning for prediction of colorectal cancer outcome: a discovery and validation study," *The Lancet*, vol. 395, no. 10221, pp. 350–360, 2020.

[2] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, pp. 1195–1204, 2017.

[3] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[4] M. Peikari, S. Salama, S. Nofech-Mozes, and A. L. Martel, "A cluster-then-label semi-supervised learning approach for pathology image classification," *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.

[5] J. Li, W. Speier, K. C. Ho, K. V. Sarma, A. Gertych, B. S. Knudsen, and C. W. Arnold, "An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies," *Computerized Medical Imaging and Graphics*, vol. 69, pp. 125–133, 2018.

[6] R. Sparks and A. Madabhushi, "Out-of-sample extrapolation utilizing semi-supervised manifold learning (ose-ssl): content based image retrieval for histopathology images," *Scientific reports*, vol. 6, p. 27306, 2016.

[7] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*, pp. 69–84, Springer, 2016.

[8] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[9] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4L: Self-supervised semi-supervised learning," in *Proceedings of the IEEE international conference on computer vision*, pp. 1476–1485, 2019.

[10] S. Graham, D. Epstein, and N. Rajpoot, "Dense steerable filter cnns for exploiting rotational symmetry in histology images," *arXiv preprint arXiv:2004.03037*, 2020.

[11] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[12] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, 2013.

[13] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, pp. 5049–5059, 2019.

[14] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[15] A. K. Jaiswal and et al., "Semi-supervised learning for cancer detection of lymph node metastases," *arXiv preprint arXiv:1906.09587*, 2019.

[16] H. Su and et al., "Local and global consistency regularized mean teacher for semi-supervised nuclei classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 559–567, Springer, 2019.

[17] S. Shaw, M. Pajak, A. Lisowska, S. A. Tsaftaris, and A. Q. O'Neil, "Teacher-student chain for efficient semi-supervised histology image classification," *arXiv preprint arXiv:2003.08797*, 2020.

[18] M. Y. Lu and et al., "Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding," *arXiv preprint arXiv:1910.10825*, 2019.

[19] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

[20] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[21] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[22] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3722–3731, 2017.

[23] J. Ren, I. Hacihaliloglu, E. A. Singer, D. J. Foran, and X. Qi, "Adversarial domain adaptation for classification of prostate histopathology whole-slide images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 201–209, Springer, 2018.

[24] F. Xing, T. Bennett, and D. Ghosh, "Adversarial domain adaptation and pseudo-labeling for cross-modality microscopy image quantification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 740–749, Springer, 2019.

[25] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström, "A closer look at domain shift for deep learning in histopathology," *arXiv preprint arXiv:1909.11575*, 2019.

[26] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[27] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

[28] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6874–6883, 2017.

[29] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[30] M. Noroozi, H. Pirsiavash, and P. Favaro, "Representation learning by learning to count," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5898–5906, 2017.

[31] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. Eslami, and A. v. d. Oord, "Data-efficient image recognition with contrastive predictive coding," *arXiv preprint arXiv:1905.09272*, 2019.

[32] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with convolutional neural networks," in *Advances in neural information processing systems*, pp. 766–774, 2014.

[33] Y. Sun, E. Tzeng, T. Darrell, and A. A. Efros, "Unsupervised domain adaptation through self-supervision," *arXiv preprint arXiv:1909.11825*, 2019.

[34] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2229–2238, 2019.

[35] J. Gildenblat and E. Klaiman, "Self-supervised similarity learning for digital pathology," *arXiv preprint arXiv:1905.08139*, 2019.

[36] H. Freeman and L. Garder, "Apictorial jigsaw puzzles: The computer solution of a problem in pattern recognition," *IEEE Transactions on Electronic Computers*, no. 2, pp. 118–127, 1964.

[37] A. C. Ruifrok, D. A. Johnston, *et al.*, "Quantification of histochemical staining by color deconvolution," *Analytical and quantitative cytology and histology*, vol. 23, no. 4, pp. 291–299, 2001.

[38] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, pp. 2234–2242, 2016.

[39] B. E. Bejnordi, M. Veta, and V. Diest, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
[41] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.
[42] A. Newell and J. Deng, "How useful is self-supervised pretraining for visual tasks?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7345–7354, 2020.

Response to Reviewers' Comments

"*Self-Path: Self-supervision for Classification of Pathology Images with Limited Annotations*" TMI-2020-2016

We thank the reviewers and the Associate Editor for their valuable time and insightful comments on our manuscript that have certainly helped us improve the quality of the manuscript.

We have addressed comments by all reviewers and incorporated all the recommended changes, as detailed in the point-by-point response below. All modifications in the revised manuscript have been highlighted in blue color.

Reviewer# 2

The authors satisfactory addressed most of the previous comments and extended the paper with new results that shed light on the effects of the different elements of their semi-supervised learning framework. I trust that this work will be of high interest to the journal readership, while complete enough to provide a good understanding of the techniques described.

> We thank the reviewer for their compliments about our work being of high interest and complete enough for providing a good understanding of the methods presented in the paper.

I would still like to suggest the following comments in order to further improve the current form of the paper:

2.1: Regarding the response R6.2 of the authors to my previous concern, I am glad that the cross-validation results are similar or better than the reported ones. However, I would suggest including such results in the paper itself. I would argue that repeating experiments with different seeds does not lead to a fair comparison. Such fairness would require a proper statistical testing on independent samples, such as those presented in this response with cross-validation. In addition, I would strongly suggest applying a proper hypothesis testing method among those samples (e.g. Wilcoxon test) when claiming superiority of any of the proposed methods over others.

R2.1: We appreciate the reviewer's concern about fairness of comparison and perhaps about reproducibility. In the interest of the short turnaround time of 2 weeks for revising the manuscript and for the benefit of reproducibility, we have made the code of our method available at the following address: https://github.com/navidstuv/self_path/

2.2: The illustration of JigMag in Figure 2 seems to have the wrong order. In my understanding, it should be 40x-20x-5x-10x instead of 40x-20x-10x-5x

R2.2: We thank the reviewer for bringing this to our attention. We have made the order clearer in Figure 2 of the revised manuscript.

2.3: Tables V-VIII & X show the results with only one experiment, whereas the rest of tables in the paper show results as mean and standard deviation. I would suggest keeping the consistency on the experimental settings and displaying of the results.

R2.3: We thank the reviewers for their suggestion. As suggested, we have added the sd for sensitivity analyses.

2.4: In the new Section V.A the authors conclude that having similar weights for the main and auxiliary tasks leads to the best results. However, only values of alpha smaller or equal than 1 are investigated in Table VII. Besides, the table shows the trend that bigger alpha leads to better results. Consequently, it would be worth investigating what happens when alpha>1.

R2.4: In here, by choosing the alpha values greater than one, the pretext task will be dominant. Therefore the main task does not learn discriminant features for separating the classes. Moreover, we are interested to see the values of alpha up to one (when it is similar to the main task).

2.5: Section V.B is a nice inclusion that improves the paper. However, only pathology specific tasks are considered. Since combining these leads to better results, it would be worth exploring if simultaneously accounting for pathology-agnostic tasks could further improve them.

R2.5: We agree with the reviewer that it may be worth exploring if simultaneously accounting for pathology-agnostic tasks could further improve the results. We have added this as a potential future direction in the Conclusions section (Sec VI, page 11).

2.6: The new data in Section V.B suggests that combining different tasks provides the best results in this work. Thus, the paper would look more conclusive by providing the results of this model on the different datasets in the different data regimes (e.g. by comparing with the results in Tables II-IV or as has been done in Table IX).

R2.6: In our sensitivity analyses, we chose the Camelyon data set as it is publicly available. Moreover, since the variation of hyperparameters are studied, it is expected that these trends will be similar on other dataset.

2.7: Section V.D also provides additional information that makes the work more complete. I would only suggest adding some comments that better connect it with the rest of the paper, e.g., that the multi-task approach proposed leads to better results than fine-tuning.

R2.7: We have added a sentence towards the end of Section V to clarify this point.

Typos:
1. Fig. 4 caption: “the The circle” instead of “The circle”
2. Table VIII caption: “tsks” instead of “tasks”

We thank the reviewer for bringing these typos to our attention. These have been fixed now.

Reviewer# 4

4.1: The authors addressed most of my concerns. The only one I still unconvincing is 3.4. The authors just stated "all possible variations of orders can give us better results as well" and changed 12 to 24. No explanation about this choice is given and more insightful analysis should be preferred.

R4.1: Intuitively, at first only 12 random variations of orders were considered where there was no guarantee that these variations are the most suitables choices. With 4 magnification choices, one could come up with 24 permutations. In the work, we had earlier chosen only 12 permutations of these to experiment with. As per the suggestions from the reviewers - we have utilized all 24 permutations and have improved our results. Hence we report the same. However, this may be worth exploring in detail and hence we have added this as another potential future direction in the Conclusions section.

Reviewer# 6

6.1: Authors have tried using all 3 augmentation techniques together with different weights. I feel the authors can try using only 2 tasks JigMag and hematoxylin as magnification is already part of JigMag.

R6.1: We agree with the reviewer that combining JigMag and Hematoxylin may give similar performance to that obtained by combining the three tasks because JigMag is indeed a major contributor to the boost in performance (as evident in Table IX). Therefore, we have added a recommendation (at the end of section V.B) on starting with JigMag before combining it with other self-supervision tasks. It is also worth mentioning that JigMag is a more challenging task for a network to solve, because the network should consider what magnification locates where, whereas in case of magnification task, the network is trying to predict only magnification.

IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. XX, NO. XX, XXXX 2020 1

Self-Path: Self-supervision for Classification of Pathology Images with Limited Annotations

Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy and Nasir Rajpoot, *Senior Member, IEEE*

***Abstract*— While high-resolution pathology images lend themselves well to 'data hungry' deep learning algorithms, obtaining exhaustive annotations on these images for learning is a major challenge. In this paper, we propose a self-supervised convolutional neural network (CNN) framework to leverage unlabeled data for learning generalizable and domain invariant representations in pathology images. Our proposed framework, termed as Self-Path, employs multi-task learning where the main task is tissue classification and pretext tasks are a variety of self-supervised tasks with labels inherent to the input images. We introduce novel pathology-specific self-supervision tasks that leverage contextual, multi-resolution and semantic features in pathology images for semi-supervised learning and domain adaptation. We investigate the effectiveness of Self-Path on 3 different pathology datasets. Our results show that Self-Path with the pathology-specific pretext tasks achieves state-of-the-art performance for semi-supervised learning when small amounts of labeled data are available. Further, we show that Self-Path improves domain adaptation for histopathology image classification when there is no labeled data available for the target domain. This approach can potentially be employed for other applications in computational pathology, where annotation budget is often limited or large amount of unlabeled image data is available.**

***Index Terms*— Computational pathology, Limited annotation budget, Semi-supervised learning, Domain adaptation.**

I. Introduction

THE recent surge in the area of computational pathology can be attributed to the increasing ubiquity of digital slide scanners and the consequent rapid rise in the amount of raw pixel data acquired by scanning of histology slides into digital whole-slide images (WSIs). These developments make the area of computational pathology ripe ground for deep neural network (DNN) models. In recent years, there have been notable successes in training DNNs for pathology image analysis and automated diagnosis of disease in the histopathology domain [1]. The performance and generalizability of most DNNs is, however, highly dependent on the availability of large and diverse amounts of annotated data. Although the use of digital slide scanners have made large amounts of raw data available, development of DNN based algorithms remains bottlenecked by the need for extensive annotations on diverse datasets.

In pathology, annotation burden can pose a large problem – even more so when compared to natural scene images. WSIs are by nature high resolution images (sometimes with slide dimensions as large as 200,000 $\times$ 150,000 pixels) – this hinders exhaustive annotations. For even simple use cases like detecting tumor regions or isolated tumor cells in WSIs, pathologists annotating the data need to look at regions of the tissue at multiple levels of magnification. So, even simple labeling of regions of interest can be quite demanding. This issue is compounded by the fact that the whole image can only be annotated part by part owing to its large size. Further, the annotation effort requires expert domain knowledge and significant investment on the part of specialized pathologists. To overcome these challenges, when training DNNs on new pathology image datasets, it would be desirable to pursue one or both of the following strategies: (a) labeling small amounts of the new dataset and making use of the larger pool of the unlabeled data, and/or (b) using existing labeled datasets which closely match the new dataset.

For strategy (a), *semi-supervised deep learning* approaches that learn with small amounts of labeled data and leverage larger pools of unlabeled data to boost performance can be employed. These approaches have been widely demonstrated in the computer vision community for natural scene images. Particularly popular techniques include Mean Teacher [2] and Virtual Adversarial Training (VAT) [3]. Recently, these approaches have also been applied to the area of computational pathology to address tasks such as clustering [4], segmentation [5] and image retrieval [6]. However, due to the high dimensionality of the images, the multi-scale nature of the problem, the requirement of contextual information and texture-like nature of sub-patches extracted from slides, the direct translation of popular semi-supervised algorithms into pathology classification tasks is not feasible.

For strategy (b), *domain adaptation* approaches that transfer knowledge from existing resources for related tasks to the classification task-at-hand can be employed. However, due to

Navid Alemi Koohbanani and Nasir Rajpoot are with the Department of Computer Science, University of Warwick, Coventry, CV4 7AL UK (e-mail: n.alemi-koohbananni, n.m.rajpoot@warwick.ac.uk) and also with The Alan Turing Institute, London. NR is also affiliated with the Department of Pathology, University Hospitals Coventry & Warwickshire, UK.

Syed Ali Khurram is with The School of Clinical Dentistry, University of Sheffield, 19 Claremont Crescent, Sheffield, S10 2TA UK (e-mail: s.a.khurram@sheffield.ac.uk).

Balagopal Unnikrishnan and Pavitra Krishnaswamy are with Department of Machine Intellection, Institute for Infocomm Research, (email: balagopalu, pavitrak@i2r.a-star.edu.sg).

variations in tissue, tumor types, and stain appearance during image acquisition, different pathology image datasets appear quite distinct from one another. In addition, for some rare tissue or tumor types, there may be no annotated datasets available for such knowledge transfer. Hence, direct translation of existing domain adaptation algorithms which work for natural vision images may not be possible. Yet, unlabeled data for related tasks are largely available and are less prone to bias [7]. Hence, when dealing with limited annotations, such unlabeled data can be used to capture the shared knowledge or to learn representations that can improve model performance.

To address the dual challenges of low annotations and domain adaptation in histopathology, it is possible to use unlabeled data in a self-supervised manner. In this setup, the model is supervised by labels that come inherently from the data itself without any additional manual annotations. These labels can represent distinct morphological, geometrical and contextual content of the images. Models trained on these 'free' labels can learn representations that can improve performance for a variety of tasks such as classification, segmentation and detection [8]. Self-supervision tasks can be used together with the main supervised task in a multi-task setup to improve performance for semi-supervised learning and domain adaptation [9]. However, self-supervised tasks proposed in the literature so far are mainly based on characteristics of natural scene images, which are very different from histology images. For instance, common self-supervision tasks focus on predicting the degree of rotation, flipping, and/or the relative position of objects. While these are meaningful concepts for natural scene images, they do not carry much relevance for histopathology images. Specifically, while the degree of rotation could help to also learn semantic information present in a natural image, it would not make sense for pathology images because they have no sense of global orientation [10].

In this paper, we propose the **Self-Path** framework to leverage self-supervised tasks customized to the requirements of the histopathology domain, and enhance DNN training in scenarios with limited or no annotated data for the task at hand. Our main contributions are summarized as follows:

- We introduce a generic and flexible self-supervision based framework, Self-Path, for classification of pathology images in the context of limited or no annotations.
- We propose 3 novel pathology pathology specific self-supervision tasks, namely, prediction of magnification level, solving the magnification jigsaw puzzle and prediction of the Hematoxylin channel, aimed at utilizing contextual, multi-resolution and semantic features in histopathology images.
- We conduct a detailed investigation on the effect of various self-supervision tasks for semi-supervised learning and domain adaptation for three datasets.
- We demonstrate that Self-Path achieves state-of-the art performance in limited annotation regime (when 1-2% of the whole dataset is annotated) or even when no annotations are available (in the case of domain adaptation).

A. Related Work

Semi-supervised Learning: Semi-supervised deep learning approaches are widely studied in the computer vision literature [11]. Popular methods utilize forms of pseudo labelling and consistency regularization, and utilize small amounts of labeled data alongside larger pools of unlabeled data for learning. Pseudo-labeling approaches [12] use available labels to train a model and impute labels on the unlabeled samples which are in turn used in training. MixMatch extends pseudo-labeling by adding temperate sharpening along with the mix-up augmentation [13] . Consistency-based methods regularize the model by ensuring stable outputs for various augmentations of the same sample. These can be done by enforcing consensus between temporal ensembles of network outputs like in Pi-Model [14], or between perturbed images fed to a network and its EMA averaged counterpart like in Mean Teacher [2]. Virtual adversarial training(VAT) [3] generates the perturbed images in an adversarial fashion to smooth the margin in the direction of maximum vulnerability. These methods ensure generalizability against significant image perturbations, move the margin away from high-density regions, and enable strong performance on benchmark natural scene image tasks with low annotation budgets.

However, semi-supervised learning has not been sufficiently explored in pathology image analysis. At the time of this writing, only 6 papers investigate semi-supervised learning for the histopathology domain. In [5], Li et. al proposed an EM-based approach for semi-supervised segmentation of histology images. [4] proposed a cluster based semi-supervised approach to identify high-density regions in the data space which were then used by supervised SVM in finding the decision boundary. Jaiswal *et al.* [15] used pseudo-labels for improving the network performance for metastasis detection of breast cancer. Su *et al.* [16] employed global and local consistency losses for mean teacher approach for nuclear classification. Shaw et. al [17] also proposed to use pseudo-labels of unlabeled images for fine-tuning the model iteratively to improve performance for colorectal image classification. Deep multiple instance learning and contrastive predictive coding were used together in [18] to overcome the scarcity of labeled data for breast cancer classification. Yet, there is scope for improvement to close the gap between fully supervised baselines and semi-supervised methods employing just a few labeled pathology images.

Domain Adaptation: Domain adaptation methods focus on adapting models trained on a source dataset to perform well on a target dataset. Leading-edge techniques mainly use adversarial training for aligning the feature distributions of different domains. Popular domain-adversarial learning-based methods [19], [20] use a domain discriminator to classify the domain of images. These methods play a minimax game where the discriminator is trained to distinguish the features from the source or target sample, while the feature generator is trained to confuse the discriminator. [21] employed adversarial learning and minimized Wassertein distance between domains to learn domain-invariant features. Image-translation methods minimize the discrepancy between the two domains at an

image-level [22]. In pathology, Ren *et al.* [23] employed adversarial training for domain adaptation across acquisition devices (scanners) in a prostate cancer image classification task. [24] used CycleGAN to translate across domains for a cell/nuclei detection task. [25] introduced a measure for evaluating distance between domains to enhance the ability to identify out-of-distribution samples in a tumor classification task. Yet, most practical domain adaptation techniques require labeling of target domain data, and the applicability of state-of-the-art unsupervised domain adaptation approaches for histopathology is yet to be widely established.

Self-Supervision: Self-supervision employs pretext tasks (based on annotations that are inherent to the input data) to learn representations that can enhance performance for the downstream task [8]. Autoencoders [26] are the simplest self-supervised task, where the goal is to minimize reconstruction error and the proxy labels are the values of image pixels. Other self-supervised tasks in the literature are image generation [8], inpainting [27], colorizing grayscale images [28], predicting rotation [29], solving jigsaw puzzle [30], and contrastive predictive coding [31]. Perhaps the main difference between contrastive learning approaches and methods like ours is that while our method caters to a specific use case domain and the task at hand is to come up with self-supervision tasks, the contrastive learning approaches offer the advantage of a more generic framework for learning representations potentially at the cost of losing performance in a very specific use case domain (such as histopathology). Although the classical self-supervision approaches requires no additional annotations, it is also possible to leverage small amounts of labeled data within a self-supervision framework. For example, S4L [9] showed that the pretext task (e.g., rotation, self-supervised exemplar [32]) can benefit from small amount of labeled data alongside larger unlabeled data. Moreover, some works [33], [34] demonstrated the effect of self supervised tasks for domain adaptation, where in [34] the effect of various self-supervised tasks have been shown for domain alignment. Particularly, solving jigsaw puzzle [34] has been proved to be a beneficial pretext task for domain generalization.

As there is no large labeled dataset akin to ImageNet for pretraining in the pathology domain, self supervised learning offers potential to obtain pre-trained model that preserves the useful information about data in itself. Although one recent study [35] explored self-supervised similarity learning for pathology image retrieval, much of the self-supervision literature is focused on computer vision applications. A key challenge in applying self-supervision to pathology-specific applications is to define the pretext task that will be most beneficial. As such, systematic analysis and derivation of pretext tasks customized for a range of histopathology applications would be desirable.

II. Problem Formulation

We now define the problem of semi-supervised learning and domain adaptation for pathology image classification. Consider a whole slide image (WSI) that is comprised of a number of disjoint or overlapping 'patches'. We denote an input image or 'patch' as $\mathbf{x}$ and its associated class label as y.

*a) **Semi-supervised Learning**:* We consider a set of n_l limited labeled images $S_L = \{(\mathbf{x}_i^l, y_i)\}_{i=1l}^{n}$, and a set of $m_l >> n_l$ unlabeled images $S_U = \{(\mathbf{x}_i^u)\}_{i=1l}^{m}$. The semi-supervised framework seeks to leverage the large pool of unlabeled images in S_U to enhance the generalizability of learning with fewer labeled images in S_L. Generally, in the semi-supervised setting, both S_L and S_U are from the same distribution.

*b) **Domain Adaptation**:* We define a source domain S comprising a set of η_s labeled images $D_s = \{\mathbf{x}_i^s, y_i^s\}_{i=1}^{\eta_s}$. Likewise, we have a target domain T comprising a set of η_t unlabeled images $D_t = \{\mathbf{x}_i^t\}_{i=1}^{\eta_t}$. Both source and target domains have the same labels. Further, source and target domains have related task characteristics, but their data distributions are distinct.

III. Methods

Our proposed Self-Path framework is depicted in Figure 1. To address label scarcity for the main classification task (main task), Self-Path leverages self-supervision and informs the supervised learning for the main task with the self-supervised learning for pretext tasks. Further, our proposed framework employs a multi-task learning approach to learn class-discriminative and domain-invariant features that would generalize with limited annotated data. Specifically, Self-Path (a) can leverage one or more pathology-specific or pathology-agnostic pretext tasks, (b) is amenable to adversarial or non-adversarial training, and (c) allows flexibility to incorporate semi-supervised, generative learning and/or domain adaptation approaches. We now formally describe the multi-task learning objective and detail the pretext tasks that are used along with the main task.

A. Multi-task Learning

Our proposed approach trains the model using the main and pretext tasks in conjunction. The framework comprises a shared encoder which learns features that are common to both the pretext task and the main task. Each task usually has a separate head connected to the shared encoder and learning for all tasks is optimized simultaneously. Formally,

$$\begin{aligned}\underset{\theta_c,\theta_e,\,\theta_{p_1},..\theta_{p_k}}{\operatorname{argmin}} & \frac{1}{n_l}\sum_{i}^{n_l} L_c(F_c^{\theta_c}(F_e^{\theta_e}(\mathbf{x}_i^l)),\ y_i) \\ +\frac{1}{n_l}\sum_{k=1}^{K}\alpha_{p_k}\sum_{i}^{n^l} & L_{p_k}(F_{p_k}^{\theta_{p_k}}(F_e^{\theta_e}(\mathbf{x}_i^l)),\ r_{ik}^l) \\ +\frac{1}{n_u}\sum_{k=1}^{K}\alpha_{p_k}\sum_{i}^{n_u} & L_{p_k}(F_{p_k}^{\theta_{p_k}}(F_e^{\theta_e}(\mathbf{x}_i^u)),\ r_{ik}^u)\end{aligned} \quad , \tag{1}$$

where K is the number of pretext tasks, r is the label for pretext task; L_c and L_{p_k} are the losses for the main and pretext tasks, respectively; F_e, is the shared encoder, F_c is the function for main task and F_{p_k} is the function of k^{th} pretext task; θ_c, θ_e and θ_{p_n} are parameters of main task classifier, shared encoder and pretext tasks, respectively; α_{p_k} indicates weights for different tasks; and n_l and n_u indicate the number of

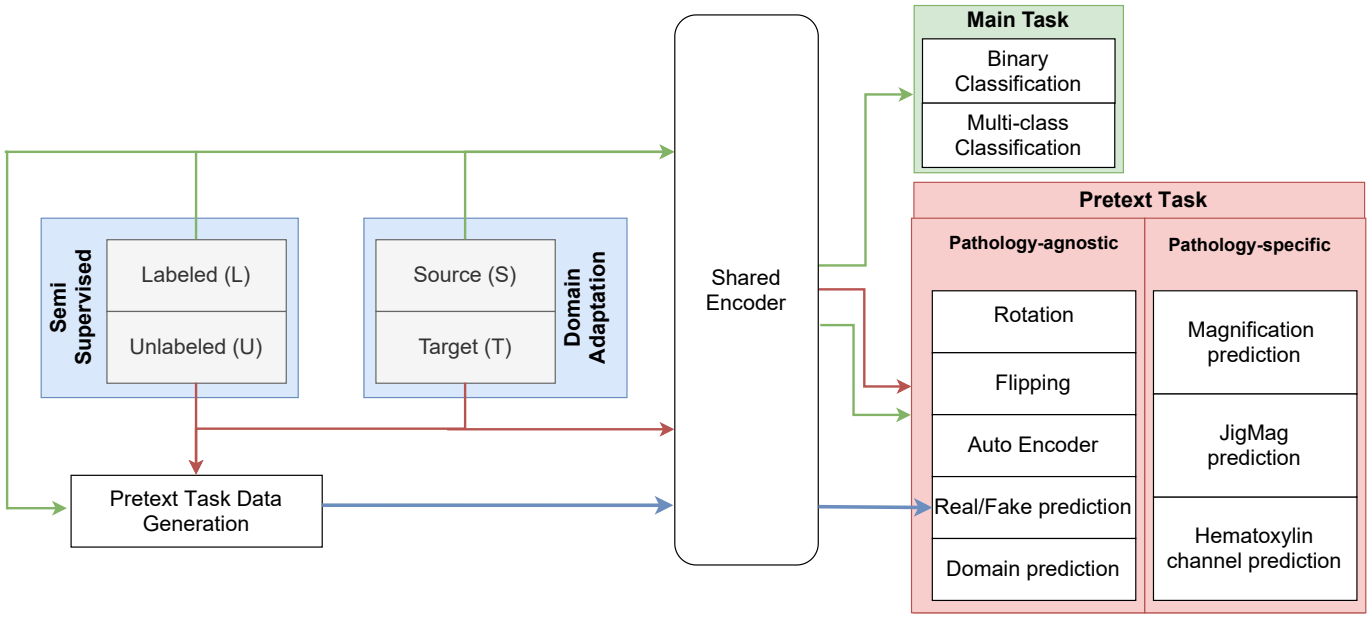

Fig. 1. Overview of Self-Path : The framework employs self-supervised pretext tasks. Pretext tasks can be added atop a shared encoder to learn useful representations and enhance semi-supervised learning or domain-adaptation. Green, red and blue lines indicate the flow of labeled, unlabeled and generated images, respectively. Generated images are used only for the generative task.

labeled and unlabeled images, respectively. When this model is used for semi-supervised learning, the labeled and unlabeled images come from the same domain. When used for domain adaptation, the labeled images come from source domain and unlabeled images come from the target domain.

B. Self-Supervision

The self-supervision utilizes one or more pretext tasks to leverage information in the unlabeled images and improve performance for the main task. Our setup employs both pathology-specific and pathology-agnostic self-supervised tasks. Every pretext task p_k is defined by a transformation function g_k applied to input x, and an implicit label r_k for the transformed input $\tilde{x} = g_k(x)$. Then, the objective function L_{p_k} is the objective for learning the self-supervised classification task that maps $\tilde{x}$ to r_k.

C. Pathology-specific Pretext tasks for Self-supervision

Histopathology images can vary in shape, morphology and arrangement of the nuclei across tissue types and disease conditions. Learning these features or semantic representations of these features can enable generalizable classification models that can more effectively transfer knowledge across domains. Therefore, we design pathology-specific pretext tasks that cater to morphology, context and shapes of nuclei as detailed below :

1) Magnification Prediction: Histopathology images are often generated and viewed at various standard magnification levels. Considering an image of fixed size, higher magnifications provide more details but less context, whereas lower magnifications allow less details but more context of tissue region. Pathologists assessing an image tend to infer important semantic information by iterating between detail and context – i.e., by zooming in and out on WSIs or by looking at different magnification levels [1]. In other words, magnification levels are implicitly correlated with important semantic information. Therefore, to enable the classification model to learn semantic information, we set up a pretext task focused on estimating magnification level of the image. Specifically, the pretext task focuses on classifying the input image to 1 of 4 magnification levels (40×, 20×, 10× and 5×). We extract images or patches from WSIs at these magnification levels Figure 2 (A). If a magnification level is not available, we obtain the patches by (bi-linear) resizing patches from other magnification levels that are available. For example, to obtain 128 × 128 patches at 5×, we extract patches of 1024 × 1024 at 40× and downsample by factor of 4. We then feed the extracted images to the network, which learns by minimizing a cross-entropy objective function.

2) Solving Magnification Puzzle (JigMag): A basic problem in pattern recognition is the jigsaw task of retrieving an original image from its shuffled parts [36]. Convolutional neural networks (CNNs) have been employed to solve the jigsaw puzzle [7]. To solve the jigsaw puzzle, it is known that the network should learn the global semantic representation of images. This is achieved by concentrating on the differences between tiles and their positions while avoiding low level statistics [7]. In histopathology, objects are smaller compared to natural scene images, and there is no specific ordering among the objects. For example, the relative positions of different parts of dog in a natural scene image is consistent, however we do not have a similar concept in histopathology. Therefore, solving the jigsaw puzzle is by itself not sufficient

[1]Magnification levels and their corresponding resolutions vary for each scanner. However by observing one particular magnification of an image, other magnifications can be perceived easily for the same scanner.

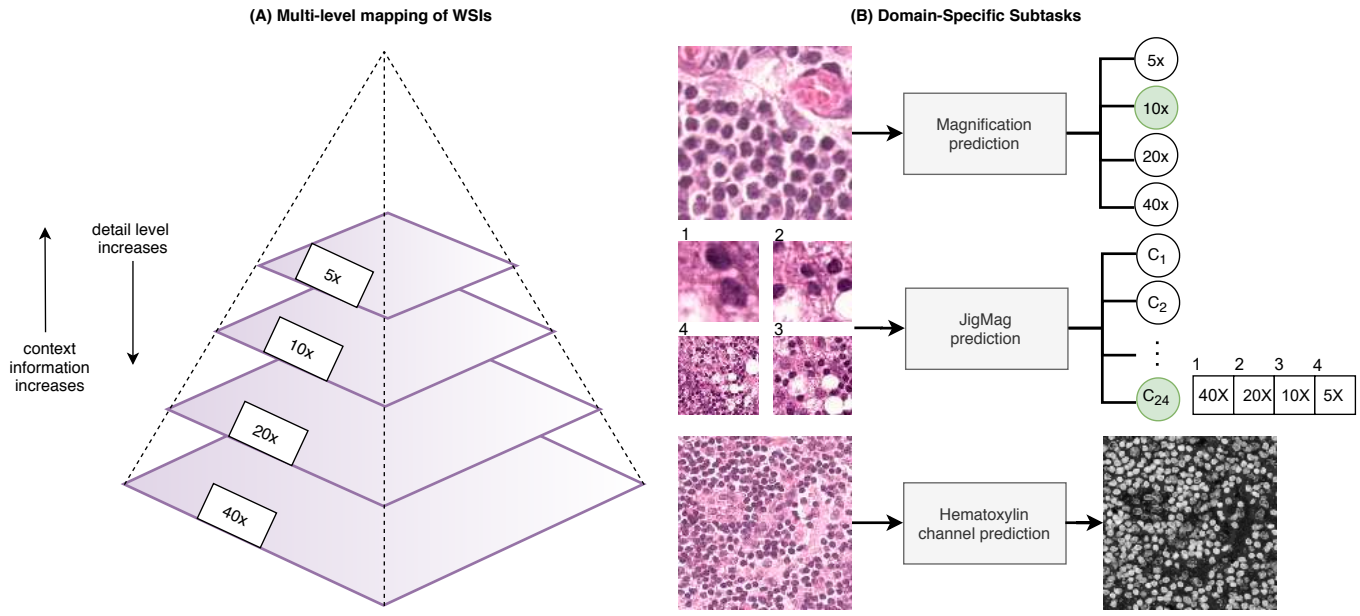

Fig. 2. (A) Whole slide images (WSI) in pathology slides organized hierarchically - each level trades-off the degree of detail against the availability of contextual information. (B) Pathology specific pretext tasks created for Self-Path.

for learning useful semantic representations in histopathology.

Instead, we propose to create a puzzle to reflect the magnification and context characteristics of histopathology images. Conceptually, classification can be enhanced by having the network implicitly learn object size and associated contextual information. Hence, we propose a pretext task focused on solving this magnification and context puzzle. In this puzzle, an image consists of image tiles with various magnifications and the network is tasked with predicting their arrangement. This set up caters also to the need to classify images containing objects with varying shapes and sizes.

Specifically, we define v as a vector of image orders in a 2×2 grid where each grid includes a specific magnification. For example $v = [0, 1, 2, 3]$ defines that image with magnification $5\times$ is on top left corner, $10\times$ is on top right and so on. We consider 24 different orders of magnification. To construct our proposed jigsaw puzzle, we first extract patches of size 512×512 at $40\times$ magnification then each part of the puzzle is constructed by down-sampling and or center-cropping to the size of 64×64, where each reflects specific context and resolution of the the original extracted patch. This pretext task employs a cross entropy loss function.

3) Hematoxylin Channel Prediction: Commonly, histopathology images are stained with Hematoxylin and Eosin (H&E). In H&E images, hematoxylin turns the palish color of nuclei to blue and eosin changes the color of other contents to pink. Color deconvolution methods have been applied to specifically identify cell nuclei in H&E images. Therefore by extracting hematoxylin channel, one can locate the nuclei and their approximate shape. Pathologists often use the location, shape and morphology of nuclei in the hematoxylin channel to diagnose or classify histopathology images (especially for malignant features).

Therefore, one way to enhance learning of useful representations is to enable the classifier to identify the nuclei and their associated characteristics. We choose to define a pretext task focused on predicting the hematoxylin channel from H&E. We use the approach in [37] to extract the hematoxylin channel in our images and define the ground truth for the self-supervision task. We scale the values of hematoxylin channel in the range [0,1] and employ a mean absolute loss for optimizing this task.

D. Pathology-agnostic Self-supervision Tasks

The literature has investigated various pretext tasks like rotation prediction, flipping, image reconstruction [8], [29]. These were however, not tailored for pathology data. Here, we systematically study and benchmark efficacy of these pretext tasks for semi-supervised learning and domain adaptation in histopathology applications.

1) Prediction of Image Rotation: For predicting rotation, the input image is rotated with degrees of $0°$, $90°$, $180°$ and $270°$ corresponding to the labels 0, 1, 2 and 3, respectively [29].

2) Prediction of Image Flipping: The label assigned to the horizontal flipping of image is 1 and 0 if not flipped.

3) Image Reconstruction with Autoencoder: For reconstructing the image, a convolutional decoder is used on top of the feature extractor [26], similar to one for predicting hematoxylin channel however 3 channels is considered for output.

4) Real vs Fake Prediction (Generative): The generative learning literature has shown that predicting whether an image is real or fake can help to learn useful representations for classification [38]. Therefore, we introduce a generative pretext task focused on real vs. fake prediction. To learn this pretext task, we train a generative network in an adversarial fashion by using unlabeled samples. While one could use a shared encoder to extract features, we found that it is easier to employ a simpler encoder/discriminator similar to the generative adversarial network (GAN) in [38].

Formally, real images are drawn from distribution D_{real}, and the generative function learns the distribution D_{gen} where the goal is to align this two distributions ($D_{gen} \sim D_{real}$). The generator $G(.)$ takes predefined noise variables z from a uniform distribution D_{noise}. The objective function is defined as:

$$\begin{aligned} L_{dis} &= -\mathrm{E}_{x\sim\ D_{real}}[\log[1 - F_{Dis}(F_e(x))]] \\ &-\mathrm{E}_{x\sim\ D_{gen}}[\log[F_{Dis}(F_e(x))]] \\ L_{gen} &= \|\mathrm{E}_{x\sim D_{real}}[F_e(x)] - \mathrm{E}_{z\sim D_{noise}}[F_e(G(z))]\|_1 \end{aligned} \quad , \quad (2)$$

where L_{gen} and L_{dis} are the generator and discriminator losses, respectively. $F_e(x)$ is the feature from intermediate layer of feature extractor (last layer before fully connected layers) and $F_{Dis}(F_e(x))$ is the output of the discriminator (fake/real head).

5) Domain Prediction: In order to learn useful representations to facilitate domain adaptation, it is useful to have a network learn the common features between source and target domains. Therefore, we introduce a pretext task to predict if the image belongs to source or target domain, and employ it in combination with other pretext tasks for the domain adaptation experiments.

For this pretext task, we employ a domain adversarial neural network (DANN) [20]. DANN includes a minimax game where discriminator H_d (domain prediction head) is trained to distinguish between the source and target domain, and the feature extractor is simultaneously trained to confuse the discriminator. Therefore, to extract the common or domain-invariant features, the parameters of feature extractor θ_e (shared encoder in the multi-task setup) are learned by maximizing the loss of domain discriminator L_d, while parameters of the domain discriminator are learned by minimizing the loss of domain discriminator. Parameters of the main task F_c are also minimized to ensure good performance on the main task. Formally:

$$\begin{aligned} \underset{\theta_c,\theta_e}{\operatorname{argminmax}}\ \underset{\theta_d}{} \frac{1}{\eta_s}\sum_{i=0}^{\eta_s} L_c(F_c^{\theta_c}(F_e^{\theta_e}(\mathbf{x}_i^s)),\ y_i) + \\ -\frac{\alpha_d}{\eta_s+\eta_t}(\sum_{i=1}^{\eta_s+\eta_t} L_d(F_d^{\theta_d}(F_e^{\theta_e}(\mathbf{x}_i)),\ d_i) \end{aligned} \quad , \quad (3)$$

where d_i is the domain label for $\mathbf{x}_i$ and α_d is a coefficient for discriminator loss. In practice, we apply domain confusion using the Gradient Reversal Layer (GRL), where the gradients of L_d with respect to the gradients of feature extractor parameters θ_e ($\frac{\partial L_d}{\partial \theta_e}$) are reversed during back-propagation.

IV. Experiments

A. Datasets

1) Camelyon16: We used the Camelyon 16 challenge dataset [39] that contains 399 H&E stained WSIs obtained on patients with breast cancer metastasis in the lymph nodes. The WSIs were acquired from 2 different centers, namely: Radboud University Medical Center (RUMC) and University Medical Center Utrecht (UMCU). RUMC images were generated by a digital slide scanner (Pannoramic 250 Flash ; 3DHISTECH) with a 20× objective lens (0.243 μm × 0.243 μm) and UMCU images were produced using a digital slide scanner (NanoZoomer-XR Digital slide scanner C12000-01; Hamamatsu Photonics) with a 40× objective lens (0.226μm × 0.226 μm). The tumor regions are exhaustively annotated by pathologists. We used the official training and testing splits comprising 270 and 129 WSIs, respectively. We randomly sampled 34 WSIs of the training set for validation. For our experiments, we randomly extracted patches from both normal and tumor regions (Table I).

TABLE I
NUMBER OF WSIS AND PATCHES IN EACH DATASET.

		Train	Validation	Test
Camelyon16	WSIs	236	34	129
	patches	67054	15586	16562
LNM-OSCC	WSIs	100	14	103
	patches	55416	7224	14472
Kather	patches	79994	20006	7180

2) LNM-OSCC: LNM-OSCC is an in-house dataset comprising 217 H&E WSIs obtained on patients with Oral Squamous Cell Carcinoma (OSCC). Of these 217 patients, 140 have metastases in the cervical lymph nodes and 77 do not manifest metastases in the cervical lymph nodes. The WSIs were acquired from 2 hospitals using 2 different scanners – (a) 98 WSIs scanned with 40× objective lens using IntelliSite Ultra Fast Scanner (0.25 μm/pixel) at University Hospital Conventry and Warwickshire (UHCW), and (b) 119 WSIs scanned at the School of Medical Dentistry in Sheffield University by Aperio/Leica CS2 with 20× objective lens (0.2467 μm/pixel). The training set comprises 100 WSIs, the validation set 14 WSIs and testing set 103 WSIs. For those cases in the training and validation sets that have metastases, a sampling of the tumor and normal regions were delineated with bounding box annotations by pathologists. For the testing set, the tumor regions were exhaustively annotated at the pixel-level.

3) Kather: This dataset contains 107,180 image patches from H&E stained WSIs comprising human colorectal cancer (CRC) and normal tissue. For this dataset, only patches were available (no WSIs).The dataset covers 9 tissue classes: Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM). We used the official data splits comprising 100k patches for training and 7180 patches for testing. We randomly sampled 20k patches of the training set for validation.

B. Data Summary

Figure 3 shows some illustrative examples of the different datasets used in our study. The overall data statistics are shown in Table I. For Camleyon16 and LNM-OSCC datasets, we extracted patches from the WSIs, and patches are distributed equally for each class. For our main task the patch extraction size is 128 × 128 at 10×. The Kather dataset patches are sized 224 × 224 and we resized to 128 × 128 for our experiments.

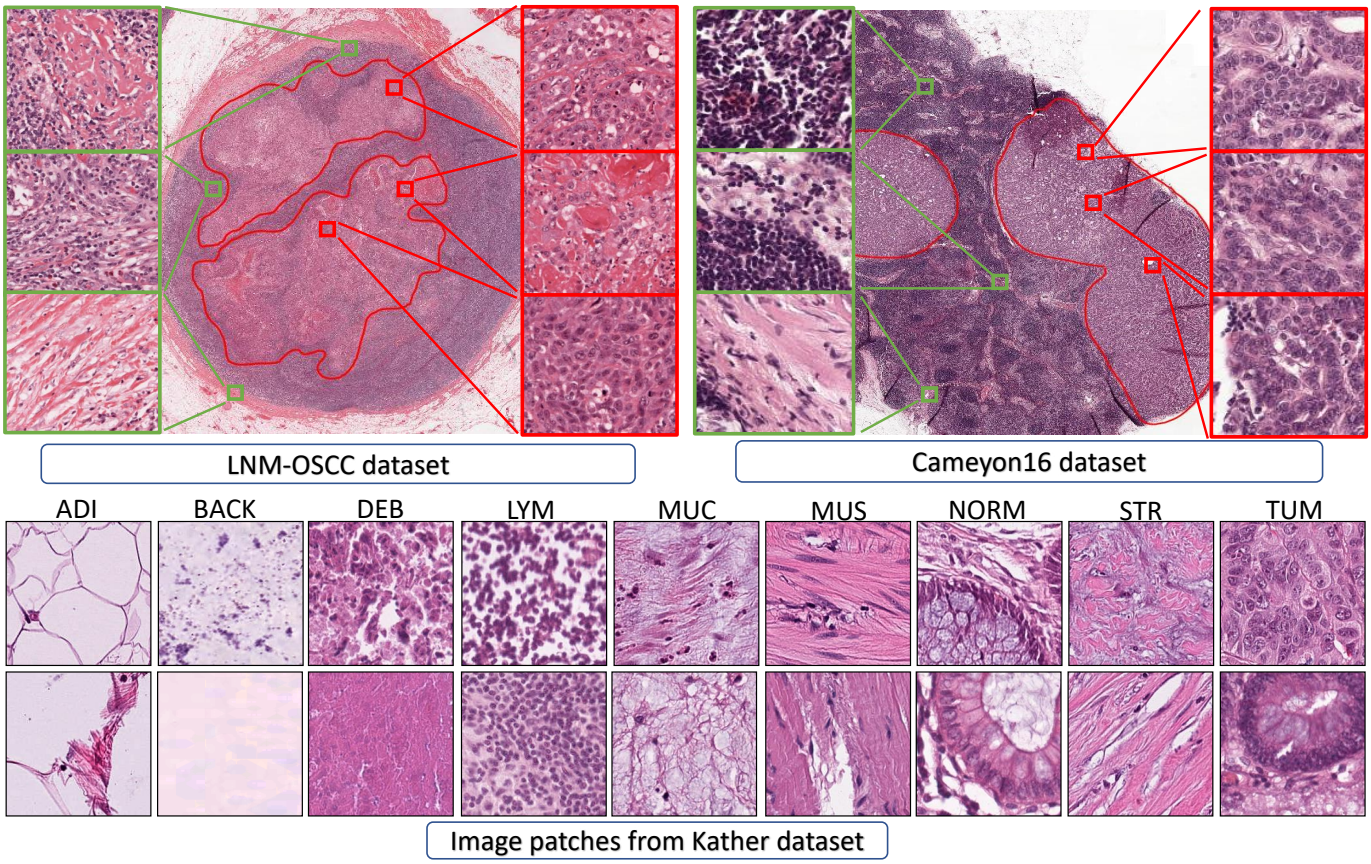

Fig. 3. Exemplar images of different datasets that are used in this study. Red and green boxes denote the tumor and normal image patches.

C. Experimental Setup

1) Networks: We chose Resnet50 [40] as the feature extraction backbone for all our experiments. The classifier head consists of adaptive average pooling which is followed by fully connected layer and softmax. The decoder head for reconstructing image and predicting hematoxylin channel is similar to the UNet decoder [41] (Supplementary Material) without using any skip connections. While using the real vs fake pretext task for image generation, we utilize the architecture presented in [38] (Supplementary Material) and find that this simpler feature extractor allows easy and robust convergence for the image generator.

2) Implementation Details: When Resnet50 is used as the shared encoder, we trained the network for 200 epochs. Our experiments used batch size 64, Adam optimizer, and learning rate of 10^{-3}. We fed batches of labeled and unlabeled images to the network separately. Therefore an epoch is defined as one full step through all the unlabeled images. Since our self-supervised experiments utilize fewer labeled images than unlabeled images, the labeled images are repeated in an epoch. Experiments related to real vs fake prediction used number of epochs and batch size of 500 and 32, respectively; and employed Adam optimizer with learning rate of 3×10^{-4}. For training model in multitask setup, we separately input batches of images for each task to the network and then sum their losses with their corresponding weights. Finally we back-propagate the whole loss through the network.

D. Results of Semi-Supervised Experiments

Here, we compare the effect of different self-supervision tasks for semi-supervised learning. We compare our models against the popular semi-supervised benchmarks, namely Mean Teacher [2] and VAT [3]. We also compare with teacher-student chain [17] (TSchain). TSchain is a recent semi-supervised approach for histopathology domain, that predicts the pseudo-labels for the unlabeled data and then uses all images for iteratively retraining the model. For performance evaluations, we follow the typical protocol of varying the annotation budget for the training set while maintaining a fixed validation set, and reporting AUCs (average across 3 seeds) on the test set.

1) Results for LNM-OSCC Dataset: We report performance of each of the self-supervised tasks on LNM-OSCC dataset in Table II. We have evaluated the model performance in terms of AUROC (Area Under the Receiver Operating Characteristic) for different annotation budgets (1%, 4%, 5%, 10% and 20% of the available WSIs). The semi-supervised approaches train on a combination of the labeled and unlabeled WSIs. The supervised baseline is only trained on labeled images without utilizing any unlabeled images.

We observe from Table II that at very low annotation budgets, pathology specific self-supervised tasks outperform the baselines and the pathology agnostic self-supervised tasks. For instance, at annotation budgets of 1% (1 labeled WSI, 134 labeled patches) and 4% (4 labeled WSIs, 1120 labeled

TABLE II

LNM-OSCC RESULTS FOR DIFFERENT ANNOTATION BUDGETS. ANNOTATION BUDGET IS DEFINED AS THE PERCENTAGE OF AVAILABLE WSIS THAT ARE LABELED. THE NUMBER OF PATCHES ASSOCIATED WITH EACH BUDGET ARE INDICATED IN THE PARENTHESES. THE SUPERVISED UPPER BOUND PERFORMANCE WHEN USING ALL LABELED DATA IS 98.4%.

% Labeled WSIs (No. Patches)	1%(134)	2%(1024)	5%(1880)	10%(3334)	20%(7558)
	AUROC(%)	AUROC(%)	AUROC(%)	AUROC(%)	AUROC(%)
	Baselines				
supervised baseline	73.4 ± 2.0	76.1 ± 5.3	85.3 ± 6.3	86.3 ± 2.7	96.3 ± 0.3
mean teacher [2]	75.1 ± 4.5	78.4 ± 5.6	86.2 ± 7.6	91.4 ± 1.2	**97.4 ± 0.3**
VAT [3]	74.5 ± 5.6	77.4 ± 3.3	85.3 ± 4.3	92.1 ± 1.2	96.5 ± 0.9
TS chain [17]	75.3 ± 2.4	79.3 ± 2.5	85.2 ± 3.1	94.1 ± 1.7	97.2 ± 0.2
	Pathology-Agnostic Self-supervised Tasks				
rotation	74.5 ± 5.6	76.3 ± 4.2	88.4 ± 1.5	93.2 ± 0.3	96.2 ± 0.1
flipping	74.6 ± 4.0	74.2 ± 5.3	85.3 ± 4.1	91.4 ± 0.4	94.2 ± 0.4
autoencoder	73.0 ± 6.5	75.1 ± 3.5	84.2 ± 3.3	90.3 ± 1.5	94.3 ± 0.2
generative	73.4 ± 7.1	79.3 ± 4.1	**90.3 ± 2.4**	**95.4 ± 0.2**	97.1 ± 0.3
	Pathology-Specific Self-supervised Tasks				
magnification	76.3 ± 4.0	76.6 ± 3.6	87.4 ± 2.3	92.5 ± 0.2	94.1 ± 0.4
JigMag	**80.6 ± 3.5**	**81.8 ± 5.3**	89.5 ± 5.4	92.4 ± 0.5	96.5 ± 0.2
hematoxylin	75.3 ± 7.6	80.2 ± 5.3	87.5 ± 1.2	94.4 ± 1.3	**97.4 ± 0.5**
Best self-supervised	**80.6 ± 3.5**	**81.8 ± 5.3**	**90.3 ± 2.4**	**95.4 ± 0.2**	**97.4 ± 0.5**

patches), JigMag task has the best performance. At annotation budgets of 1% and 2%, Hematoxylin and magnification tasks outperform pathology agnostic tasks and generative tasks. When annotation budget increases to 10%, we observe that the generative task performs much better (AUC 95.4%), suggesting that the generated images can help the classifier to boost the performance. Overall, our LNM-OSCC experiments suggest that for limited annotation budgets, pathology specific pretext tasks are helpful for enhancing the model performance, with JigMag outperforming other approaches.

2) Results for Camelyon16 Dataset: We report performance of each of the self-supervised tasks on Camelyon16 dataset in Table III. We have evaluated the model performance in terms of AUROC (Area Under the Receiver Operating Characteristic) for different annotation budgets (1%, 2%, 5%, 10% and 20% of the available WSIs). The semi-supervised approaches train on a combination of the labeled and unlabeled WSIs. The supervised baseline is only trained on labeled images without utilizing any unlabeled images.

Similar to LNM-OSCC dataset, pathology specific tasks outperform other semi supervised methods. In particular, the JigMag task improves the performance over the supervised baseline by 13.4%, 11.8% and 6.2% at 1% (2 WSIs), 2% (4 WSIs) and 5% (8 WSIs) annotation budgets, respectively. At 1% annotation budget, only magnification and JigMag outperform mean teacher and supervised baseline. Unlike LNM-OSCC, the generative model cannot achieve highest AUROC for any annotation budget, but it's performance is competitive with mean teacher and VAT. Similar to LNM-OSCC, JigMag could achieve highest performance overall, and the main boost is obtained at very low annotation budgets.

3) Results for Kather Dataset: We report performance of each of the self-supervised tasks on Kather dataset in Table IV. Since there are 9 classes in the Kather dataset, Macro AUROC is used for evaluation of classification performance. Unlike the other 2 datasets, only patches were available for this dataset, therefore the annotation budget only reflects the proportion of the overall patches that is labeled. Further, we observe that at 2% annotation budget, the performance of supervised baseline is still high (Macro AUC of 98%). Hence using semi-supervised approaches would not add much benefit. Hence, we focus on the very low annotation budget regime where some degradation of Macro-AUC can be observed for supervised model – i.e., annotation budgets of 0.1%(100 labeled) and at 1% (800 labeled images). Moreover, as this dataset does not include WSIs, we were unable to extract large patches or patches at different magnificationsand hence could not evaluate JigMag and magnification self-supervised tasks on this dataset.

From Table IV, we observe that at 0.1% annotation budget, predicting hematoxylin channel as a self-supervised task improves the performance by 2.8% and 1.2% compared to the baseline and mean teacher, respectively. At 1% annotation budget, we see that the various self-supervised tasks can again improve performance compared to the baseline. Predicting hematoxylin channel can also give the superior performance, suggesting that the prediction of rough nuclear segmentations can be helpful for semi-supervised learning.

E. Domain Adaptation Experiments

We conduct two domain transfer experiments, (i) Camelyon16 to LNM-OSCC (Cam16→LNM-OSCC) and (ii) LNM-OSCC to Camelyon16 (LNM-OSCC→Cam16). In both cases, we do unsupervised domain transfer, where the source is the labeled set and the target set is completely unlabeled.

We evaluate our approach against the naive supervised baseline, and two other domain adaptation methods WDGRL [21] and DANN [20]. The supervised baseline employs Resnet50 and is trained with source domain data only. WDGRL trains a domain critic network to estimate the Wasserstein distance between the source and target feature representations. The feature extractor network will then be optimized to minimize the estimated Wasserstein distance in an adversarial manner. By iterative adversarial training, WDGRL learns feature representations invariant to the covariate shift between domains. DANN is a domain prediction approach based on the GRL unit and was mentioned in Section III-D.

We report the results obtained with Self-Path (using different pretext tasks) and the comparisons with the supervised and

TABLE III

CAMELYON16 RESULTS FOR DIFFERENT ANNOTATION BUDGETS. ANNOTATION BUDGET IS DEFINED AS THE PERCENTAGE OF AVAILABLE WSIS THAT ARE LABELED. THE NUMBER OF PATCHES ASSOCIATED WITH EACH BUDGET ARE INDICATED IN THE PARENTHESES. THE SUPERVISED UPPER BOUND PERFORMANCE WHEN USING ALL LABELED DATA IS 94.2%.

Labeled WSIs (No. Patches)	1%(600)	2%(1000)	5%(2600)	10%(6400)	20%(13540)
	AUROC(%)	AUROC(%)	AUROC(%)	AUROC(%)	AUROC(%)
	Baselines				
supervised baseline	68.3 ± 5.1	74.5 ± 5.8	81.2 ± 2.5	88.4 ± 2.3	92.1 ± 0.5
Mean Teacher [2]	73.7 ± 3.8	78.5 ± 2.6	84.5 ± 2.4	92.7 ± 1.9	93.1 ± 0.9
VAT [3]	70.9 ± 5.8	77.4 ± 3.3	81.3 ± 5.2	90.3 ± 2.3	92.8 ± 1.5
TS chain [17]	74.9 ± 6.9	76.9 ± 3.2	83.8 ± 2.1	93.1 ±2.5	**93.9 ± 1.3**
	Pathology-Agnostic Self-supervised Tasks				
rotation	69.8 ± 4.8	74.5 ± 3.1	80.4 ± 2.5	90.1 ± 2.0	92.4 ± 2.5
flipping	70.2 ± 6.2	75.4 ± 3.5	81.6 ± 5.1	89.4 ± 0.6	92.3 ± 1.6
autoencoder	70.1 ± 2.4	75.6 ± 4.1	82.3 ± 4.5	90.5 ± 2.3	92.4 ± 1.1
generative	72.5 ± 5.5	77.6 ± 5.4	82.4 ± 7.2	92.6 ± 3.2	93.6 ± 1.5
	Pathology-Specific Self-Supervised Tasks				
magnification	77.5 ± 3.1	84.6 ± 5.2	85.1 ± 3.6	**93.2 ± 3.4**	93.4 ± 2.5
JigMag	**81.7 ± 3.8**	**86.3 ± 5.2**	**87.4 ± 4.5**	90.6 ± 4.6	92.8 ± 2.4
hematoxylin	72.8 ± 4.6	78.3 ± 4.5	84.6 ± 3.4	92.3 ± 4.1	**93.7 ± 2.5**
Best Self-supervised	**81.7 ± 3.8**	**86.3 ± 5.2**	**87.4 ± 4.5**	**93.2 ± 3.4**	**93.7 ± 2.5**

TABLE IV

KATHER RESULTS FOR DIFFERENT ANNOTATION BUDGETS. ANNOTATION BUDGET IS DEFINED AS THE PERCENTAGE OF AVAILABLE WSIS THAT ARE LABELED. THE NUMBER OF PATCHES ASSOCIATED WITH EACH BUDGET ARE INDICATED IN THE PARENTHESES. THE SUPERVISED UPPER BOUND PERFORMANCE WHEN USING ALL LABELED DATA IS 99.4%.

Labeled WSIs (No. Patches)	0.1%(100)	1%(800)
	AUROC(%)	AUROC(%)
	Baselines	
supervised baseline	87.5 ± 2.0	92.5 ± 1.2
mean teacher [2]	89.1 ± 1.5	93.9 ± 0.3
VAT [3]	88.5 ± 1.4	92.6 ± 0.4
TS chain [17]	88.9 ± 0.3	93.5 ± 0.2
	Self-supervised tasks	
generative	88.4 ± 3.5	92.3 ± 2.6
rotation	87.4 ± 1.6	93.3 ± 0.4
flipping	88.6 ± 0.8	93.0 ± 0.9
autoencoder	89.3 ± 1.3	94.3 ± 1.2
hematoxylin	90.3 ± 0.7	95.1 ± 0.5
Best self-supervised	**90.3 ± 0.7**	**95.1 ± 0.5**

TABLE V

AUROC RESULTS FOR DOMAIN ADAPTATION

	Cam16→LNM-OSCC	LNM-OSCC→Cam16
	Baselines	
supervised baseline	79.53 ± 0.2	63.73 ± 0.5
DANN	89.23 ± 1.5	71.15 ± 0.6
WDGRL	89.64 ± 2.6	72.65 ± 2.2
	Pathology-Agnostic Self-supervised Tasks	
rotation	86.14 ± 3.4	66.91 ± 4.1
flipping	82.14 ± 3.6	65.95 ± 4.4
autoencoder	89.90 ± 2.8	71.62 ± 2.6
generative	91.54 ± 3.5	74.14 ± 2.7
	Pathology-Specific Self-supervised Tasks	
magnification	89.69 ± 3.6	73.62 ± 4.1
JigMag	92.34 ± 4.4	74.51 ± 3.6
hematoxylin	90.47 ± 4.5	73.24 ± 3.8
mag+hem+JigMag	92.85 ± 3.6	74.95 ± 3.5

domain adaptation baselines in Table V. We observe that the pathology-specific pretext tasks can help the model outperform the baseline by a large margin. For Cam16→LNM-OSCC, the pathology-specific pretext tasks provide more than 10% boost in AUROC over the supervised baseline. The combination of all pathology specific pretext tasks achieves the best performance. Amongst the individual pretext tasks, JigMag achieves the best performance (∼2% better than DANN and WDGRL). Further, we note that the pathology agnostic generative model also performs well – with 1.9% higher AUROC than WDGRL and 11% higher AUROC over the supervised baseline. This suggests that the images from the generator can contribute to learning useful domain-invariant features as well. We see similar trends for LNM-OSCC→Cam16 – where again combining pathology specific tasks has the best performance and JigMag provides the second best performance. We highlight that we have used domain prediction with GRL layer in all non-generative methods as it improves the performance. Generative models, owing to adversarial training can still achieve very high performance, even without GRL.

1) WSI Analysis: While the results thus far are reported at the patch level, it is also useful to consider the WSI-level performance. For the Cam16→LNM-OSCC domain adaptation task, we now report the WSI-level results for the top two best performing Self-Path settings i.e., combination of all pathology specific pretext tasks and JigMag pretext task. We also provide comparisons with the supervised baseline (source only), WDGRL, and the pathology agnostic generative pretext task.

In order to quantify WSI-level performance, we aggregate patches belonging to a WSI and construct a WSI-level heat map based on the patch level predictions. For heat map generation, there are two steps. First, we extract patches of 128 × 128 at 10× magnification with overlap of 50% from tissue regions of WSIs. Second, we aggregate the prediction of each patch together to build the final heat map of WSIs. We then post-process these heat maps to obtain the WSI-level prediction. The post-processing steps are uniform for all models in this section, and as follows: we extract 10 morphological and geometrical features from objects within binarized heat map at three thresholds of 0.25, 0.5 and 0.9. Then we calculate the mean, stddev, minimum and maximum of object features for each WSI. Therefore, in total we use 120 features for constructing feature vectors. Afterwards, we

TABLE VI
CAM16 → LNM-OSCC DOMAIN ADAPTATION RESULTS ON THE WSI-LEVEL. THE UPPER BOUND PERFORMANCE USING ALL LABELS FOR TARGET DOMAIN IN SUPERVISED FASHION IS 93.3%.

	AUROC(%)	Average Precision(%)
supervised baseline (source only)	75.2	81.7
WDGRL	85.8	91.6
generative	90.4	95.2
JigMag	91.6	96.7
mag+JigMag+hem	91.6	96.3

employ the random forest algorithm for classification of the features. Finally, we evaluate the model on the test set of LNM-OSCC.

The results are shown in Table VI. The supervised baseline has WSI-level AUROC of 75.2% whereas Self-Path with JigMag pretext task and Self-Path with the combination of all pathology specific pretext tasks each improve the performance by 16.4%. Further, we note that Self-Path with JigMag improves performance over WDGRL by 2% at the patch-level and a ∼6% improvement at the WSI-level. This suggests that the magnification puzzle and the pretext tasks that can help learn from various image resolutions in a self-supervised manner enable strong performance boost at WSI-level (beyond patch-level).

These improvements are also evident in the WSIs overlaid with the heatmaps, as visualized in Figure 4. This figure shows that the supervised baseline (source only) model (middle column) has many false negatives and often misses tumor regions. However, WDGRL, Self-Path with JigMag, and Self-Path with generative pretext task can all increase true positives while decreasing false negatives. We note that WDGRL and Self-Path with generative pretext task do not perform as well as Self-Path with JigMag - mainly because they suffer larger number of false positives at the patch-level classification.

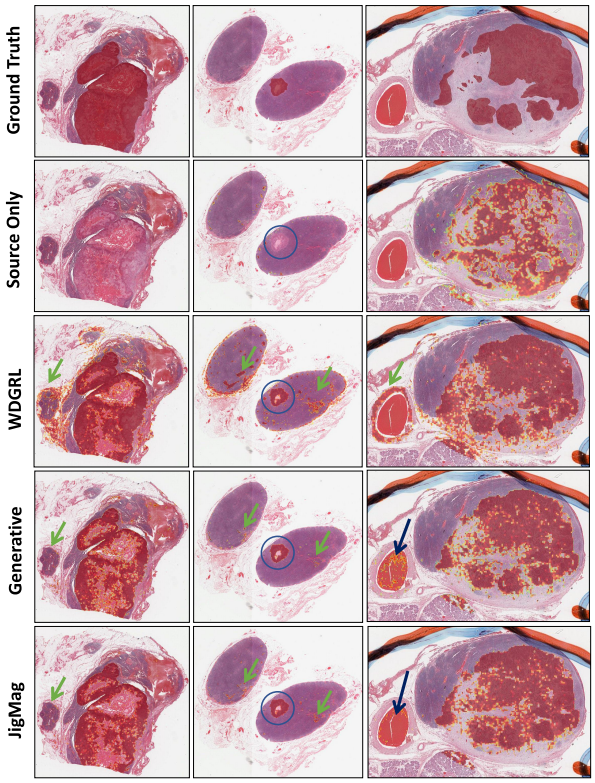

Fig. 4. Three WSI samples and their overlaid heatmaps. from top to bottom, first row: the overlaid ground-truth mask, second row: overlaid heat map of model predictions when it is trained using only Camlelyon16 data, third row: Overlaid heatmap of WDGRL predictions, fourth row depicts the overlaid predictions of Self-path using generative task and the last row shows the heatmaps generated Self-path using JigMag task. The circle indicates a region which is missed using the supervised baseline (source only) model and green arrows point to the false positive regions generated by WDGRL where using generative task and JigMag task eliminate those regions. Black arrow also shows regions that are misclassified by generative model but are correctly classified as normal regions by Jig-Mag. (Best viewed in color, zoom in to see more details)

V. DISCUSSION

In this section we describe sensitivity analyses and discuss the model performance by changing the values of loss weights, decreasing the annotation budget and combing all pathology specific tasks. Moreover, we conduct an experiment to show the usefulness of transfer learning using our proposed self-supervised tasks. For following experiments, we choose Camelyon dataset. Since the variation of hyperparameters are studied, it is expected that these trends will be similar on other dataset.

A. Effect of Loss Weight for Each Task

We consider the task of training with 1% of annotation budget on Camelyon16 dataset. To understand the effect of loss weights for each pretext task, we experiment with different values of α and show the results in Table VII. Overall, assigning more weights on each task shows better performance. More precisely, when α is set to 1, maximum value of AUROC is obtained. Therefore we can conclude when we are using only one pretext task, the pretext task and the main task should have similar weight to be effective for semi-supervised learning. The optimum value of α may change when we use all tasks together which we investigate in the next section. In here, by choosing the alpha values greater than one, the pretext task will be dominant. Therefore the main task does not learn discriminant features for separating the classes. Moreover, we are interested to see the values of alpha up to one (when it is similar to the main task).

B. Combining tasks

We now evaluate the effect of the loss weights (α's) when combining all pathology specific tasks. We consider the task of training with 1% and 2% of annotation budget on Camelyon16 dataset, and experiment with different combinations of loss coefficients. The results, in Table VIII, suggest that assigning high weights (similar to main task) to all pretext tasks can degrade the performance. For example, if all tasks are given $\alpha = 1$, overall the weights for pretext tasks would be $3\times$ more than the main task which would cause drop in performance. However by assigning smaller weight values for each task, we can achieve better performance. Particularly, best performance

TABLE VII
AUROC PERFORMANCE OF PATHOLOGY SPECIFIC TASKS WITH DIFFERENT VALUES OF α ON CAMELYON16 DATASET.

α	magnification	JigMag	hematoxylin
1	77.5 ± 3.1	81.7 ± 3.8	72.8 ± 4.6
0.8	77.1 ± 2.8	81.5 ± 3.4	71.3 ± 2.4
0.6	76.4 ± 4.0	78.8 ± 2.6	70.2 ± 3.5
0.5	74.6 ± 3.4	78.4 ± 2.4	70.3 ± 4.6
0.2	72.5 ± 3.7	74.1 ± 4.6	69.5 ± 4.4

TABLE VIII
USING ALL PATHOLOGY SPECIFIC TASKS FOR SEMI-SUPERVISED LEARNING ON CAMELYON16 DATASET. α_{mag}, α_{JigMag} AND α_{hem} INDICATE THE LOSS COEFFICIENT FOR MAGNIFICATION, JIGMAG AND HEMATOXYLIN TASKS, RESPECTIVELY.

α_{mag}	α_{JigMag}	α_{hem}	1%	2%
1	1	1	79.1 ± 4.5	83.5 ± 5.1
0.25	0.5	0.25	83.2 ± 4.3	86.3 ± 5.3
0.5	0.25	0.25	80.2 ± 2.5	85.4 ± 3.1
0.25	0.25	0.5	79.6 ± 2.7	84.3 ± 5.5
0.25	0.25	0.25	80.3 ± 3.4	85.5 ± 1.8

is obtained when more weight is assigned to JigMag task and lower weights to Hematoxylin and magnification tasks. This is in line with previous experiments which showed that JigMag had better performance as compared to other tasks. We can, therefore, recommend that a good strategy can be to start with heavy weight to JigMag for computational pathology tasks before combining it with other self-supervision tasks.

C. Performance at Very Low Annotation Budget

In section IV-D, we evaluated the performance of self-supervised tasks with different annotation budgets. we observed, despite high boost in performance by applying self-supervised tasks, the supervised baseline also gives reasonable results (e.g., 73.4% on LNM-OSCC for 134 patches). To assess performance at even lower annotation budget, we further decreased number of patches annotated (while maintaining the same number of WSIs) to 50 for LNM-OSCC and Camelyon datasets. As shown in Table IX, Self-Path with pathology-specific pretext tasks can improve the AUC by about 10% over the supervised baseline. Again, the JigMag pretext task is the best performing pretext task. Moreover, we also note that combining all pathology specific tasks (with loss weights 0.25, 0.25 and 0.5 for hematoxylin, magnification and JigMag respectively) can result in even better performance.

D. Transfer Learning

We finally investigate the usefulness of the representations learned by Self-Path for related tasks. For this, we conduct a transfer learning experiment using Camelyon16 dataset. We first train Self-Path with each self-supervised pretext task on the entire dataset, and then fine-tune the backbone (the model excluding the final linear layer/decoder) for the main task. We compare the performance against the naive method of training the network from scratch with random weight initializations (Scratch). The results for different pretext tasks at varying annotation budgets are shown in Table X. We can see that the representations learned by Self-Path with transfer

TABLE IX
AUROC RESULTS FOR VERY LOW BUDGET OF ANNOTATION:HERE ONLY 25 IMAGE PATCHES ARE USED IN EACH CLASS

	Camelyon16	LNM-OSCC
	Baselines	
supervised baseline	55.3 ± 5.1	54.8 ± 8.1
mean Teacher	65.4 ± 4.8	60.4 ± 5.4
VAT	64.3 ± 6.4	58.6 ± 6.5
TS chain	62.4 ± 10.6	59.4 ± 7.7
	Pathology-Agnostic Self-supervised Tasks	
rotation	62.6 ± 4.6	58.7 ± 4.6
flipping	65.7 ± 9.3	58.9 ± 5.3
autoencoder	65.1 ± 6.4	59.6 ± 4.3
generative	64.2 ± 5.7	60.1 ± 10.3
	Pathology-Specific Self-supervised Tasks	
magnification	65.3 ± 7.5	62.2 ± 6.7
JigMag	**66.2 ± 6.4**	**63.5 ± 7.9**
hematoxylin	64.2 ± 7.4	62.4 ± 4.6
mag+hem+JigMag	**66.5 ± 5.5**	**64.1 ± 5.5**

TABLE X
RESULTS OF TRANSFER LEARNING OF SELF-SUPERVISED TASKS WITH DIFFERENT BUDGET OF ANNOTATIONS USING CAMELYON16 DATASET.

	1%	2%	5%	10%	20%
Scratch	68.3	74.5	81.2	88.4	92.1
magnification	72.6	77.4	84.8	89.9	92.2
JigMag	73.3	79.4	85.8	90.4	92.7
hematoxylin	72.9	79.5	85.9	88.6	92.3

learning enable performance improvement over 'Scratch' in each case. Again, Self-Path with JigMag achieves the best performance. The improvements with fine-tuning is largest in the low annotation regime, and drops off when more annotated data are available. These results suggest that the pretext tasks in Self-Path enable learning of useful representations. Overall, with annotation budget of over 20%, fine-tuning gives the same result as training from scratch. Therefore multi-task approach where self-supervision task and main task are trained together leads to better results than fine-tuning.Therefore multi-task approach where self-supervision task and main task are trained together leads to better results than fine-tuning. This phenomenon is also shown by [42].

VI. CONCLUSIONS

In this paper, we proposed Self-Path – a generic framework based on self-supervision tasks for histopathology image classification – to address the challenge of limited annotations in the area of computational pathology. We introduced 3 novel self-supervision tasks to cater to the contextual, multi-resolution and semantic features in pathology images. We showed that such pathology specific self-supervision tasks can improve the classification performance for both semi-supervised learning and domain adaptation. Moreover, we thoroughly investigated general self-supervised approaches such as generative models within this pipeline and showed that using the pathology-specific tasks, despite being simple and easy to implement, can improve performance over generic self-supervision in many scenarios involving limited annotation budget or domain shift. In particular, we note that the JigMag self-supervision can be extremely helpful when the amount of labeled data is very small. Unlike baseline methods that are highly dependent on hyperparameters values, our method can

achieve good performance without exhaustive hyperparameter tuning. Self-Path can be applied to other problems in computational pathology, where annotation budget is often limited or large amounts of unlabeled image data are available. In our sensitivity analyses, we considered only domain specific tasks and showed that their combination leads to better performance compared to using only one pretext task in the multitask setup. Using all domain agnostic task as pretext task can also potentially increase the performance and requires further exploration. Other future directions include employing other self-supervision tasks (such as predicting the Eosin channel or a combination of Hematoxylin and Eosin after estimating the two channels, rather than keeping them fixed), increasing the number of magnification levels, increasing the JigMag grids to incorporate wider and more complex puzzles for the network to solve, exploring different variations of orders for JigMag (here all 24 orders were used) and a deeper investigation into other domain adaptation tasks.

REFERENCES

[1] O.-J. Skrede, S. De Raedt, A. Kleppe, T. S. Hveem, K. Liestøl, J. Maddison, H. A. Askautrud, M. Pradhan, J. A. Nesheim, F. Albregtsen, *et al.*, "Deep learning for prediction of colorectal cancer outcome: a discovery and validation study," *The Lancet*, vol. 395, no. 10221, pp. 350–360, 2020.

[2] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, pp. 1195–1204, 2017.

[3] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[4] M. Peikari, S. Salama, S. Nofech-Mozes, and A. L. Martel, "A cluster-then-label semi-supervised learning approach for pathology image classification," *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.

[5] J. Li, W. Speier, K. C. Ho, K. V. Sarma, A. Gertych, B. S. Knudsen, and C. W. Arnold, "An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies," *Computerized Medical Imaging and Graphics*, vol. 69, pp. 125–133, 2018.

[6] R. Sparks and A. Madabhushi, "Out-of-sample extrapolation utilizing semi-supervised manifold learning (ose-ssl): content based image retrieval for histopathology images," *Scientific reports*, vol. 6, p. 27306, 2016.

[7] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*, pp. 69–84, Springer, 2016.

[8] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[9] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4L: Self-supervised semi-supervised learning," in *Proceedings of the IEEE international conference on computer vision*, pp. 1476–1485, 2019.

[10] S. Graham, D. Epstein, and N. Rajpoot, "Dense steerable filter cnns for exploiting rotational symmetry in histology images," *arXiv preprint arXiv:2004.03037*, 2020.

[11] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[12] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, 2013.

[13] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, pp. 5049–5059, 2019.

[14] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[15] A. K. Jaiswal and et al., "Semi-supervised learning for cancer detection of lymph node metastases," *arXiv preprint arXiv:1906.09587*, 2019.

[16] H. Su and et al., "Local and global consistency regularized mean teacher for semi-supervised nuclei classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 559–567, Springer, 2019.

[17] S. Shaw, M. Pajak, A. Lisowska, S. A. Tsaftaris, and A. Q. O'Neil, "Teacher-student chain for efficient semi-supervised histology image classification," *arXiv preprint arXiv:2003.08797*, 2020.

[18] M. Y. Lu and et al., "Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding," *arXiv preprint arXiv:1910.10825*, 2019.

[19] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

[20] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[21] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[22] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3722–3731, 2017.

[23] J. Ren, I. Hacihaliloglu, E. A. Singer, D. J. Foran, and X. Qi, "Adversarial domain adaptation for classification of prostate histopathology whole-slide images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 201–209, Springer, 2018.

[24] F. Xing, T. Bennett, and D. Ghosh, "Adversarial domain adaptation and pseudo-labeling for cross-modality microscopy image quantification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 740–749, Springer, 2019.

[25] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström, "A closer look at domain shift for deep learning in histopathology," *arXiv preprint arXiv:1909.11575*, 2019.

[26] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[27] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

[28] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6874–6883, 2017.

[29] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[30] M. Noroozi, H. Pirsiavash, and P. Favaro, "Representation learning by learning to count," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5898–5906, 2017.

[31] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. Eslami, and A. v. d. Oord, "Data-efficient image recognition with contrastive predictive coding," *arXiv preprint arXiv:1905.09272*, 2019.

[32] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with convolutional neural networks," in *Advances in neural information processing systems*, pp. 766–774, 2014.

[33] Y. Sun, E. Tzeng, T. Darrell, and A. A. Efros, "Unsupervised domain adaptation through self-supervision," *arXiv preprint arXiv:1909.11825*, 2019.

[34] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2229–2238, 2019.

[35] J. Gildenblat and E. Klaiman, "Self-supervised similarity learning for digital pathology," *arXiv preprint arXiv:1905.08139*, 2019.

[36] H. Freeman and L. Garder, "Apictorial jigsaw puzzles: The computer solution of a problem in pattern recognition," *IEEE Transactions on Electronic Computers*, no. 2, pp. 118–127, 1964.

[37] A. C. Ruifrok, D. A. Johnston, *et al.*, "Quantification of histochemical staining by color deconvolution," *Analytical and quantitative cytology and histology*, vol. 23, no. 4, pp. 291–299, 2001.

[38] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, pp. 2234–2242, 2016.

[39] B. E. Bejnordi, M. Veta, and V. Diest, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
[41] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.
[42] A. Newell and J. Deng, "How useful is self-supervised pretraining for visual tasks?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7345–7354, 2020.

IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. XX, NO. XX, XXXX 2020 1

Supplementary Material for Self-Path: Self-supervision for Classification of Pathology Images with Limited Annotations

TABLE I
PERFORMANCE OF DIFFERENT BASELINE MODELS ON THE THREE DATASETS. THE EVALUATION WAS DONE USING ONLY THE SUPERVISED LOSS AND KEEPING THE LABELING BUDGET AT ONE PERCENT.

	Kather	Camleyon16	LNM-OSCC
Labeled patches	800	600	134
Resnet50	0.9137	0.6467	0.7387
Resnet101	0.9015	0.6515	0.7314
Densenet121	0.9014	0.6514	0.7265
InceptionV3	0.8914	0.6618	0.7264

TABLE II
NETWORK ARCHITECTURE WHILE USING THE GENERATIVE REAL VS FAKE SUBTASK. CONV.T STANDS FOR TRANSPOSED CONVOLUTION.

Generator
latent space (100)
dense 4×4×512 batchnorm ReLU
5×5 Conv.T 512 batchnorm ReLU stride=2
5×5 Conv.T 256 batchnorm ReLU stride=2
5×5 Conv.T 128 batchnorm ReLU stride=2
5×5 Conv.T 128 batchnorm ReLU stride=2
5×5 Conv.T 3 weightnorm Tanh stride=2
Discriminator
128×128×3 images
dropout, $p = 0.2$
3×3 conv. weightnorm 96 lReLU
3×3 conv. weightnorm 96 lReLU
3×3 conv. weightnorm lReLU stride=2
dropout, $p = 0.5$
3×3 conv. weightnorm 128 lReLU
3×3 conv. weightnorm 128 lReLU
3×3 conv. weightnorm 128 lReLU stride=2
dropout, $p = 0.5$
3×3 conv. weightnorm 192 lReLU
3×3 conv. weightnorm 192 lReLU
3×3 conv. weightnorm 192 lReLU stride=2
dropout, $p = 0.5$
3×3 conv. weightnorm 192 lReLU
3×3 conv. weightnorm 192 lReLU
3×3 conv. weightnorm 192 lReLU
Adaptive maxpool
weightnorm dense 2

I. NETWORK ARCHITECTURE

The performance on classification tasks was evaluated using supervised learning. ResNet50 was chosen since it has overall good performance while having lower number of parameters. The AUC-ROC performances can be seen in table I. ResNet50 was used as the backbone architecture in all the self-supervision experiments except when the generative real vs fake prediction had to be used. While using the real vs fake auxiliary task for image generation, we utilize the architecture presented in table II and find that this simpler feature extractor allows easy and robust convergence for the image generator.

TABLE III
NETWORK ARCHITECTURE FOR HEMATOXYLIN/DECODER TASKS

Decoder
Resnet50 backbone
1×1 Conv.T 512 ReLU stride=1
BilinearUpsample scale_factor=2
3×3 Conv.T 512 ReLU stride=1
BilinearUpsample scale_factor=2
3×3 Conv.T 256 ReLU stride=1
BilinearUpsample scale_factor=2
3×3 Conv.T 256 ReLU stride=1
BilinearUpsample scale_factor=2
3×3 Conv.T 128 ReLU stride=1
BilinearUpsample scale_factor=2
3×3 Conv.T 65 ReLU stride=1
1×1 Conv.T **Number of classes** stride=1

TABLE IV
HYPER-PARAMETERS OF MODEL WHEN RESNET 50 IS USED AS FEATURE EXTRACTOR

Hyperparameters	Values
Batch size	64
Epoch	200
Optimizer	ADAM ($\alpha = 3 * 10^{-3}$, $\beta_1 = 0.9$)

II. HYPER-PARAMETERS

The hyper-parameters when using the various network architectures for training are shown in table IV and table V. table IV is the hyper-parameter setting when using ResNet50 as the backbone and table V are the settings used when the generative real vs fake sub-task is used.

TABLE V
HYPER-PARAMETERS FOR REAL VS FAKE PREDICTION SUBTASK

Hyperparameters	Values
Batch size	32
Epoch	500
Leaky ReLU slope	0.2
Exp. moving average decay	0.999
Optimizer	ADAM ($\alpha = 3 * 10^{-4}$, $\beta_1 = 0.5$)
Weight initialization	Isotropic gaussian ($\mu = 0$, $\sigma = 0.05$)
Bias initialization	Constant (0)