Data-Efficient Histopathology Image Analysis with Deformation Representation Learning

1st Jilan Xu#
School of Computer Science
Key Laboratory of Intelligent Information Processing
Fudan University
Shanghai, China
jilanxu18@fudan.edu.cn

2nd Junlin Hou#
School of Computer Science
Key Laboratory of Intelligent Information Processing
Fudan University
Shanghai, China
jlhou18@fudan.edu.cn

3rd Yuejie Zhang*
School of Computer Science
Key Laboratory of Intelligent Information Processing
Fudan University
Shanghai, China
yjzhang@fudan.edu.cn

4th Rui Feng*
School of Computer Science
Key Laboratory of Intelligent Information Processing
Fudan University
Shanghai, China
fengrui@fudan.edu.cn

5th Chunyang Ruan
School of Economics & Finance
Shanghai International Studies University
Shanghai, China
cyruan16@fudan.edu.cn

6th Tao Zhang
School of Information Management & Engineering
Shanghai University of Finance and Economics
Shanghai, China
taozhang@mail.shufe.edu.cn

7th Weiguo Fan
Department of Business Analytics
The University of Iowa
Iowa City, USA
weiguo-fan@uiowa.edu

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313159

***Abstract*—Histopathological examination of tissue biopsies plays a fundamental role in disease assessment. Automatic histopathology image analysis requires substantial task-specific annotations, which are often expensive and laborious in real-world scenarios. This insufficient annotation of data limits the generalization ability of supervised learning models. To address this challenge, we propose a self-supervised Deformation Representation Learning (DRL) framework to learn semantic features from unlabeled data. As a novel paradigm, our approach utilizes deformation as supervisory signals based on two critical features, i.e., local structure heterogeneity and global context homogeneity. Given an original histopathology image and its deformed counterpart, there exists a moderate difference in local structures. In contrast, due to the transformation-invariance, both images share a similar global context compared with other images. Specifically, an encoder network is trained to distinguish the local inconsistency by measuring the mutual information and maintain the global consistency with noise contrastive estimation. Extensive experiments on public histopathology image datasets show that the learned representations are generalizable for various downstream tasks, such as transfer learning on segmentation and semi-supervised classification. Our approach achieves superior results over other self-supervised methods and the *ImageNet* pre-trained model, and it reveals the ability as a novel pre-training scheme in histopathology image analysis.**

***Index Terms*—Deformation Representation Learning, Global Context Homogeneity, Histopathology Image Analysis, Local Structure Heterogeneity, Self-Supervised Learning.**

I. Introduction

Recent years have seen the widespread application of deep learning in many computer vision tasks. As one of the main topics to be investigated, automatic histopathology image analysis is highly demanded to provide reliable quantitative statistics in clinical diagnosis and biomedical interventions, such as cancer grading and survival analysis. Most successful studies [1], [2] trained their models through fully supervised learning to learn meaningful feature representations, which required a substantial number of task-specific annotations. However, the performance of the supervised models in histopathology image analysis is limited to the availability of high-quality annotations [3]. The labeling process is often laborious, tedious, and time-consuming, even for professional pathologists.

Currently, unsupervised learning is under growing exploration as a vast amount of unlabeled data is readily available. A new paradigm of unsupervised learning, Self-Supervised Learning (SSL), solves a pretext task with supervisory signals that are generated automatically to learn useful representations of data. Prior pretext tasks include coloring grayscale images, filling missing holes, or predicting rotation degrees [4]–[6]. In the original SSL setting, the effectiveness of a pretext task is usually measured by the performance of downstream tasks after transferring the learned representations to them, such as image classification and semantic segmentation. Despite promising results on natural images, learning useful representations from unlabeled histopathology images remains challenging due to the problem of domain shift [7], [8].

In the field of biomedical image analysis, deformation has been widely explored in data augmentation [2] and image registration [9]. We observe that deformation is also an essential characteristic in histopathology images. For example, the glandular structure formed by abnormal cancer cells in malignant cases (e.g., poorly differentiated adenocarcinomas) is usually deformed and degenerated [1]. In this paper, we use deformation as a free supervisory signal and propose a self-supervised Deformation Representation Learning (DRL)

This work was supported by National Natural Science Foundation of China (No. 61976057 and No. 61572140), Science and Technology Development Plan of Shanghai Science and Technology Commission (No. 20511101203, No. 20511102702, No. 20511101403), Shanghai Natural Science Foundation (No. 19ZR1417200), and Humanities and Social Sciences Planning Fund of Ministry of Education of China (No. 19YJA630116).#: equal contribution.

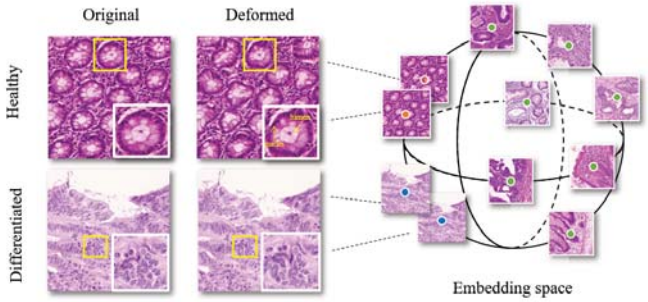

Fig. 1. An illustration of local structure heterogeneity (left) and global context homogeneity (right). The left figure shows a healthy case and a moderately-to-poorly differentiated adenocarcinomas case. The local regions in the yellow boxes are enlarged to show the difference clearly. The appearance of the glandular structures (the lumen and the nuclei) are moderately deformed. The right sphere depicts the locations of the images in the embedding space, denoted by dots with different colors.

approach for enhancing histopathology image analysis, which aims to learn representative features benefitting the subsequent segmentation and classification tasks. We form the pretext task in two perspectives, i.e., local structure heterogeneity (LSH) and global context homogeneity (GCH). As illustrated in Fig. 1, given an original histopathology image and its deformed counterpart, they have a moderate difference in local structure. This variation in the shape and structure of the specific objects (e.g., nuclei and lumen) is embedded with different features. Specifically, we apply the Mutual Information (MI) to measure the relationship between the encoded features of the original histopathology image and the deformed one. The intuition is that the encoder network is encouraged to capture the key features about high-level factors (e.g., the existence of gland objects) instead of noises to discriminate the images. Moreover, as the structural difference exists at multiple scales in the image, a Feature Enhance Module (FEM) is further integrated into the network to improve the multi-scale feature representation ability. Contrastively, due to the transformation-invariance, the original histopathology image maintains similar global context information with its deformed counterpart compared with other images (e.g., inter-cellular patterns). As shown in the right part of Fig. 1, this global context similarity can be interpreted as a closer distance between the image and its deformed counterpart in the embedding space, which the encoder network is encouraged to learn. We propose two unsupervised approaches, namely Hard-Target GCH and Soft-Target GCH, to concentrate similar samples in the embedding space using Noise Contrastive Estimation (NCE). With the advantage of the local and global feature learning, DRL generates useful representations of the histopathology images that can be transferred to various downstream tasks. Quantitative evaluations of the learned representations on different tasks, i.e., transfer learning on segmentation and semi-supervised classification, demonstrate the generalization ability of DRL.

In summary, our contributions are:

- We propose a novel paradigm of utilizing deformation as a supervisory signal for self-supervised deformation representation learning (DRL).
- We propose a local structure heterogeneity loss by measuring the mutual information between the encoded features of an original histopathology image and its deformed counterpart. Furthermore, we build a new Feature Enhance Module (FEM) in our DRL framework to capture the structural features across multiple scales.
- We design two global context homogeneity losses by applying the deformed image representations as either the hard-target or the soft-target to concentrate the original histopathology image and its deformed counterpart in the embedding space.
- Our proposed self-supervised DRL achieves competitive performance compared with the *ImageNet* pre-trained model and other self-supervised approaches on the segmentation and classification evaluation, demonstrating the strong generalization ability of our DRL.

II. Related Work

A. Self-supervised Representation Learning

In recent years, Self-Supervised Learning (SSL) appears as a novel schema in unsupervised visual representation learning. SSL solves the pretext tasks with automatically generated supervisory signals to learn useful representations for real-world downstream tasks. Generally, these supervisory signals are generated by applying one or more data augmentations to the original image or hiding a certain part of the input [4], [5], [10], [11]. Dosovitskiy et al. [12] trained a network to classify the surrogated classes and generated the classes by using different transformations to each individual instance. Zhang et al. [13] constructed an original image and its augmented image pair to predict the corresponding transformation. Hjelm et al. [14] maximized the mutual information between input data and the learned high-level representations, highlighted the global and local information, and improved the suitability of representations.

Though these pretext tasks have proven effective in natural image understanding, few of them have been applied to biomedical images [7]. Zhou et al. [15] explored the self-supervised learning paradigm on 3D medical images using image reconstruction as a pretext task. They demonstrated superior performance on a variety of CT image classification and segmentation tasks, and their work set up a novel pre-training scheme in biomedical image analysis. Motivated by the above approaches, our proposed self-supervised DRL is tailored for histopathology image representation learning.

B. Histopathology Image Analysis

Lately, deep convolutional networks have achieved remarkable performance in histopathology image analysis. We mainly focus on recent approaches in image segmentation tasks. An array of FCN-based networks [1], [16] were proposed and achieved promising results, which trained a feature encoder to extract image features and recover high-resolution prediction from low-resolution feature maps using deconvolution or bilinear upsampling. U-Net based networks [2],

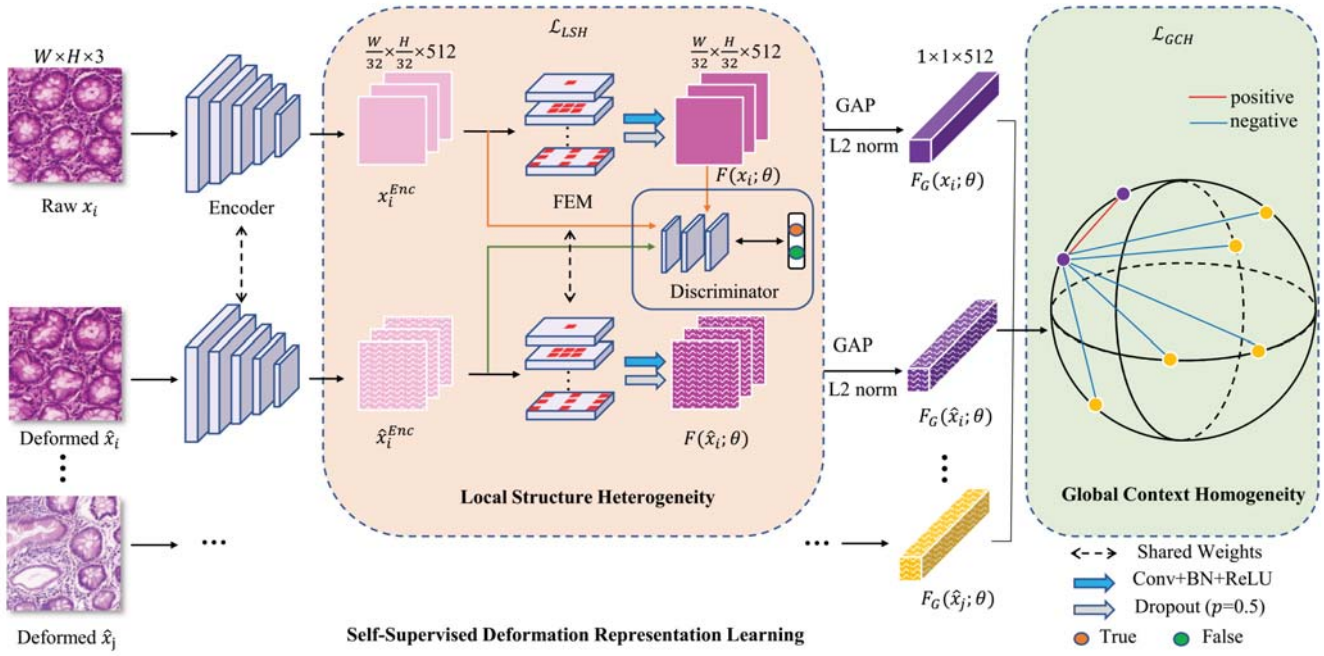

Fig. 2. An overview of our proposed DRL framework. The local structure heterogeneity and global context homogeneity are proposed to solve the self-supervised pretext task. After solving the pretext task, the encoder network can be transferred to various downstream tasks. The discriminator is introduced to distinguish the positive and the negative pairs.

[17] used multiple skip connections between the encoder and the decoder to fuse mid-level features and relieved the problem of vanishing gradient in deep networks. Especially, Chen et al. [1] formulated a two-branch end-to-end network to segment gland objects and contour simultaneously, laying the foundation for histopathology image segmentation. As data acquisition becomes challenging, a wide range of not-so-supervised approaches have been proposed to enable data-efficient biomedical image analysis [3]. Jia et al. [18] developed a weakly-supervised learning algorithm based on multiple instance learning and introduced a novel constrained deep weak supervision approach, which demonstrated strong performance on large-scale histopathology image datasets. Taylan et al. [19] performed salient subregion identification, characterization, and tissue image classification in an unsupervised feature extraction framework.

III. Methodology

A. Overview of the Framework

Deformation Representation Learning (DRL) is designed as the pretext task in the self-supervised learning paradigm. The goal of DRL is to learn good feature representations for histopathology images. The overall framework of our DRL is illustrated in Fig. 2. Let $X = \{x_1, x_2, ..., x_N\}$ represents the input images. A convolutional neural network F parameterized by θ, is usually used as a feature encoder to generate the representations $Y = F(X;\theta)$. The objective functions of our DRL are constructed based on the properties (i.e., local structure heterogeneity and global context homogeneity) of the original image and the deformed counterpart. This self-supervised learning process makes full use of the features in training data and does not include manual labels for supervision. After training, the encoder network is transferred to various downstream tasks to validate its performance.

B. Elastic Deformation

Elastic deformation was initially proposed to augment handwritten digits [20] and biomedical images [2]. First, the random displacement fields of an image are generated for x-axis and y-axis at each location (i, j) with a uniform distribution, i.e., $\Delta \mathrm{d}_{ij}^x = \mathrm{rand}(-1, 1)$, $\Delta \mathrm{d}_{ij}^y = \mathrm{rand}(-1, 1)$. A 2D Gaussian filter G of zero mean and standard deviation σ is then denoted as:

$$G(i,j) = \frac{1}{2\pi\sigma^2} e^{-\frac{(i^2+j^2)}{2\sigma^2}} \quad (1)$$

We generate the smoothed displacement fields by applying the convolution with zero padding to the original fields with the Gaussian filter G, and then multiply the displacement fields with a factor α that controls the scale of the displacement. Having obtained the new displacement fields, we map the pixels from the original location (i, j) to new coordinates $(i+\Delta \mathrm{d}_{ij}^x, j+\Delta \mathrm{d}_{ij}^y)$ by interpolation. In the family of geometric transformation, rigid transformation preserves the length and angle, and affine transformation preserves parallel lines in an image. As a form of non-rigid transformation, elastic deformation enables flexible mapping from lines to curves, which reduces the potential trivial clues in the following representation learning framework.

C. Local Structure Heterogeneity

First, we introduce the general form of our objective function. Given an original histopathology image x, we first generate the high-level representation $y = F(x;\theta)$ with the feature encoder network. The relationship between the input image and the representation, or more specifically, how well $F(x;\theta)$ is representative for x, can be measured by Mutual Information (MI), which is defined as:

$$\begin{aligned} I(X,Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \\ &= KL(p(x,y)||p(x)p(y)), \end{aligned} \tag{2}$$

where $p(x,y)$ is the joint distribution of X and Y; $p(x)$ and $p(y)$ are the marginal distributions of X and Y. The defination is equivalent to the Kullback-Leibler (KL) divergence between the joint distribution and the product of the marginal distributions. The training objective is to maximize the mutual information between the input images and the generated representations, which is formulated as:

$$\theta^* = \arg\max_{\theta} I(X, F(X;\theta)) \tag{3}$$

Deformation is introduced to construct the local structure heterogeneity. Let $\hat{X}$ be the corresponding deformed images produced by elastic deformation. As illustrated in Fig. 1, for each image and its deformed counterpart $(x_i, \hat{x}_i)$, the moderate difference in the local structures is termed as Local Structural Heterogeneity (LSH). Meanwhile, the corresponding representations in the feature space $(F(x_i;\theta), F(\hat{x}_i;\theta))$ are expected to embed this structural difference. Therefore, the network maximizes the mutual information of the positive pair $I(x_i, F(x_i;\theta))$ because x_i and $F(x_i;\theta)$ carry the information of the same image. In contrast, the network minimizes the mutual information of the negative pair $I(\hat{x}_i, F(x_i;\theta))$. The advantage of LSH originates from the objective of mutual information. It is assumed that the encoded feature $F(x_i;\theta)$ has limited capacity of representations, so the network is encouraged to encode the distinctive features such as structural or morphological patterns that share across the input image to maximize the mutual information. The computation of mutual information has been challenging as it is hard to estimate the marginal distributions in high dimensions. Motivated by the research work in [14], [21], [22], we use a Jensen-Shannon Divergence (JSD) to estimate the mutual information because JSD shares an approximately monotonic relationship with KL divergence, which is formulated as:

$$\begin{aligned} I^{JSD}(X,Y;\theta,\phi) = {} & \mathbb{E}_X[-sp(-D(x,F(x;\theta);\phi))] \\ & - \mathbb{E}_{X\hat{X}}[sp(D(\hat{x},F(x;\theta);\phi))], \end{aligned} \tag{4}$$

where $sp(x) = \log(1+e^x)$ is the softplus function. In particular, a fully convolutional discriminator $D(X, F(X;\theta);\phi) \in [0,1]$, parameterized by ϕ, is applied to distinguish the positive and the negative pairs. It consists of three convolutional blocks with 512, 256, and 256 channels, respectively. In practice, the original input X is replaced with mid-level feature maps $X^{Enc} \in R^{M\times M\times C_1}$ with the spatial resolution M and channel C_1. X^{Enc} can be interpreted as the local features where each value corresponds to a specific region in the original image; $F(X;\theta) \in R^{M\times M\times C_2}$ is generated by applying a Feature Enhance Module (FEM) on X^{Enc}. As illustrated in Fig. 3, in a few cases, some gland objects are seriously degenerated or they have large white lumen area expand across the image. To address this issue, FEM is proposed to capture the structural difference that varies across multiple scales and locations. Inspired by ASPP [23] which contains four 3×3 dilated convolutions with the dilation rates 1, 6, 12, and 18, we build FEM by extending the number of filters dynamically based on the size of the feature map. Specifically, we define $k = \lceil \log_2 M \rceil$ convolutional filters with the kernel size s and the dilation rates $r_0 = 1, r_1 = 2, r_2 = 6, \ldots, r_{k-1} = 6\times 2^{k-3}$. For each convolutional filter, the actual kernel size $\hat{s}_i$ with respect to s and r_i is calculated as:

$$\hat{s}_i = s + (s-1)\times(r_i - 1), \tag{5}$$

where $0 \le i \le k-1$. Following the definition of the dilation rates, suppose the original kernel size is $s = 3$, it can be inferred that the largest actual kernel size $\hat{s}_{k-1} = 2r_{k-1}+1 > M$, which guarantees that the kernels can cover the entire feature map for a long-range context. This condition is also satisfied when larger filters are applied (e.g., 5×5). The learned features are then concatenated to form uniform multi-scale feature maps $F(X;\theta)$. FEM is only integrated into the self-supervised framework, and will not be transferred to downstream tasks for fair evaluations. Having obtained X^{Enc} and $F(X;\theta)$, we concatenate them as the input of the discriminator with the positive label. The output of $D(X^{Enc}, F(X;\theta);\phi)$ is then averaged after a sigmoid function to calculate the local structural heterogeneity loss $\mathcal{L}_{LSH}$. Our local structure heterogeneity loss for the feature encoder and the discriminator is defined as:

$$\begin{aligned} \mathcal{L}_{LSH}(X,\hat{X};\theta,\phi) = {} & \frac{1}{N}\sum_{i=1}^{N}\log(sp(-D(x_i^{Enc}, F(x_i;\theta);\phi)) \\ & + \frac{1}{N}\sum_{i=1}^{N}\log(sp(D(\hat{x}_i^{Enc}, F(x_i;\theta);\phi)), \end{aligned} \tag{6}$$

where N is the total number of samples. With the local structure heterogeneity loss, the network can serve as a self-supervised classifier because the frequently used softmax with cross-entropy loss can also be viewed as the mutual information estimator [24].

D. Global Context Homogeneity

Global context information is usually crucial for analyzing the image-level features, such as benign and malignant in cancer classification. Since elastic deformation only changes the relative locations of the pixels randomly, x and $\hat{x}$ still preserve similar global context information, which is termed as Global Context Homogeneity (GCH). For each image x_i, we use the Gaussian function with dot product to measure the

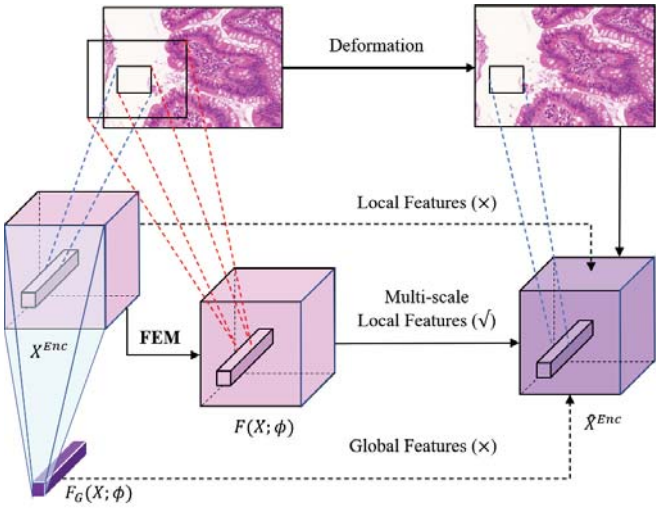

Fig. 3. An illustration of the local structure heterogeneity with FEM. Neither local features or global features have enough ability to capture structural difference at multiple scales.

similarity to another image x_j in the feature space, which can be formulated as:

$$\mathcal{K}(x_i, x_j) = e^{F_G(x_i;\theta)^T F_G(x_j;\theta)}, \tag{7}$$

where $F_G(x_i;\theta)$ is the global feature descriptor obtained by applying global average pooling and L2 normalization to $F(x_i;\theta)$. Here we propose two alternative ways to implement the global context homogeneity loss $\mathcal{L}_{GCH}$ in the embedding space, namely Hard-Target GCH ($\mathcal{L}^H_{GCH}$) and Soft-Target GCH ($\mathcal{L}^S_{GCH}$)

1) Hard-Target Global Context Homogeneity: Intuitively, it is expected that the network embeds the similarity between x_i and $\hat{x}_i$. In other words, x_i is supposed to be closer to $\hat{x}_i$ in the feature space compared with other deformed instances x_j. To achieve this goal, we use the Noise Contrastive Estimation (NCE) [25] which has proven successful in contrastive learning. It forms a set of binary classification problems with the aim of distinguishing positive data samples from negative noise samples. NCE combines softmax with the negative log-likelihood to obtain the objective function. We specify $(x_i, \hat{x}_i)$ as the positive pair and $(x_i, \hat{x}_j)$ as negative pairs, where $j \in \{1, 2, .., N\}$ and $j \neq i$. In this way, $\hat{x}_i$ is considered as the pseudo hard target. Our hard-target global context homogeneity loss is formulated as:

$$\mathcal{L}^H_{GCH}(X;\theta) = -\sum_{i=1}^{N} \log \frac{\mathcal{K}(x_i, \hat{x}_i)}{\sum_{j=1}^{N} \mathcal{K}(x_i, \hat{x}_j)} \tag{8}$$

2) Soft-Target Global Context Homogeneity: In the aforementioned formulation of NCE, all the negative pairs contribute equally to Hard-Target GCH. However, in a few cases, the images cropped from the same whole-slide histopathology image also share similar context information. Therefore, the corresponding global feature representations of these images and their deformed counterparts should be concentrated in the embedding space. Compared with one hard target $\hat{x}_i$, our goal here is to find a section of near neighbors of x_i in $\hat{X}$. We define a soft deformed target $\tilde{x}_i$ of each x_i by using the weighted combination on the deformed instances:

$$\tilde{x}_i = \frac{1}{Z} \sum_{j=1}^{N} \mathcal{K}(x_i, \hat{x}_j) \hat{x}_j \tag{9}$$

where $Z = \Sigma_j \mathcal{K}(x_i, \hat{x}_j)$ is the normalization factor. Hence, the deformed images that are similar to x_i make higher contributions to form the soft target $\tilde{x}_i$. We then follow the process in Hard-Target GCH and perform an inverse mapping of the soft target back to x_i by using NCE. The Soft-Target GCH loss is defined as:

$$\mathcal{L}^S_{GCH}(X;\theta) = -\sum_{i=1}^{N} \log \frac{\mathcal{K}(x_i, \tilde{x}_i)}{\sum_{j=1}^{N} \mathcal{K}(x_j, \tilde{x}_i)} \tag{10}$$

Both $\mathcal{L}^H_{GCH}$ and $\mathcal{L}^S_{GCH}$ incorporate the high-level context information of the images to realize the global feature learning in DRL.

By setting λ as the balancing parameter, we arrive at our objective function, which is formulated as:

$$\mathcal{L} = \mathcal{L}_{LSH} + \lambda \mathcal{L}_{GCH} \tag{11}$$

Our proposed local structure heterogeneity and global context homogeneity only employ deformation as a supervisory signal, so that the entire DRL framework can be learned in an unsupervised manner.

IV. Experiments

In order to evaluate the generalization ability of the learned representations, we exploit different downstream tasks, i.e., transfer learning on segmentation and semi-supervised classification. In this section, we introduce the detailed configuration in our experiments, and mainly describe the quantitative results on the histopathology image datasets across different tasks.

A. Datasets

We evaluate the effectiveness of DRL on two public histopathology image datasets, i.e., the MICCAI 2015 Gland Segmentation Challenge (GLaS) dataset [26], and the Patch Camelyon (PCam) image classification dataset [27].

1) GLaS Dataset: The *GLaS* dataset contains 85 training images (37 benign and 48 malignant) and 80 (37 benign and 43 malignant) test images, all obtained from 16 H&E stained histological sections. Most of the images are of 775 × 522 pixels. The ground truth segmentation masks mark the existence of glands in the histopathology images. Besides, each image owns a patient ID as well as a coarse grade (benign or malignant) that indicates the malignancy of the image.

2) PCam Dataset: The Patch Camelyon dataset is composed of 327,680 patches cropped from 400 H&E stained Whole Slide Images (WSIs) of sentinel lymph node sections. The images are of 96×96 pixels. Each image contains a binary label, indicating the presence of metastatic tissue. The whole dataset is split into the training set, validation set, and test set with 70%, 15%, and 15% of total images, respectively.

B. Experimental Setup

1) Pre-training Stage: Our framework is implemented on two NVIDIA GeForce GTX TITAN Xp GPUs. To train the DRL model, we adopt Adam optimizer with a learning rate of 0.0001 and set the batch size as 8. Note that the prior loss term is added as introduced in [14] for a fair comparison, and λ=0.01 and λ=0.5 are chosen in the segmentation task and classfication task by grid search, respectively. To fully utilize the unlabeled data, we expand the scale of *GLaS* by applying sliding window crop strategy on the training and validation set with the size=256 and the stride=64. We use VGG-19 [28] as the backbone network. We re-implement the other SSL approaches by changing the backbone network to VGG and use the same hyper-parameter configurations as DRL. We perform a grid search on the standard deviation σ of the convolution kernel and the scale factor α as mentioned in Section 3.2. We specify $\sigma \in [0.03S, 0.05S]$ and $\alpha = 2S$ where $S = \min(W, H)$ is the image size.

2) Downstream Tasks: In each downstream task, we choose the initial hyperparameters as suggested in [1] and [27], respectively, and adjust the optimal hyperparameters with grid search. The datasets are augmented with random flip, random resized crop (with the size of 256 on *GLaS* and 96 on *PCam*), and color jittering. We use F1 Score, Object-level Dice Score and Object-level Hausdorff Distance as the evalution metrics on the segmentation task and accuracy on the classification task. On the segmentation task, we initialize the encoder network with the pre-trained weights on different pretext tasks, and then fine-tune the entire network end-to-end with the SGD optimizer. The learning rate is 0.005 and the batch size is set as 2. On the classification task, we use the Adam optimizer with a learning rate of 0.001 and a batch size of 128. Note that we do not transfer the FEM into the downstream tasks, and all the networks have the same number of parameters for fair evaluations.

C. Experimental Results

1) Transfer Learning on Gland Segmentation: We evaluate the learned representations by transferring them on the gland segmentation task. Table I shows the quantitative comparison results. We train the network with random initialization as the baseline method, and compare our DRL with (1) the approaches that have dense prediction pretext tasks [4], [5]; (2) the reconstruction-based approaches [15], [29] because mutual information is closely related to the reconstruction error [14], [22]; and (3) other related self-supervised approaches [6], [12], [14]. Our $\mathcal{L}^{H}_{GCH}$ is similar to [10], [30], but they are not fairly comparable as we neither use the memory bank nor train different key and query encoders. DRL achieves the superior segmentation performance over the other approaches on all three metrics. We observe that the performance of the reconstruction-based approaches in their pretext tasks are sensitive to the image size, as high-resolution image generation remains challenging. Our self-supervised DRL is generally more robust to image resolution. Moreover, DRL outperforms the *ImageNet* pre-trained model (denoted as ImageNet), indicating that DRL can serve as a novel pre-training paradigm on histopathology images to avoid the domain gap effectively.

TABLE I
THE TRANSFER LEARNING EVALUATION RESULTS ON THE *GLaS* SEGMENTATION DATASET. IMAGENET STAND FOR *ImageNet* PRE-TRAINED WEIGHTS. THE p-VALUES ARE CALCULATED BETWEEN OUR DRL AND DIM. †: THIS MARK REPRESENTS USING ELASTIC DEFORMATION AS DATA AUGMENTATION.

Methods	**F1 Score**	**Obj. Dice**	**Obj. Hausdorff**
Supervised			
UNet† [2]	0.870	0.876	57.09
Un-/Self-supervised			
VAE [29]	0.859±0.012	0.867±0.011	60.00±3.97
MG [15]	0.862±0.003	0.856±0.009	70.13±2.25
Rotation [6]	0.838±0.015	0.850±0.009	65.31±4.93
Colorization [4]	0.863±0.004	0.863±0.005	64.85±3.02
CENet [5]	0.868±0.006	0.879±0.004	57.13±1.85
Exemplar [12]	0.876±0.004	0.879±0.006	59.71±2.56
DIM [14]	0.878±0.005	0.866±0.008	63.13±3.73
Random init.	0.853±0.009	0.861±0.006	63.49±4.40
ImageNet	0.883±0.003	0.887±0.004	53.50±0.83
DRL	**0.900±0.006**	**0.896±0.004**	**50.55±2.64**
p-value	0.0055	0.0011	0.0516

TABLE II
THE OBJECT DICE SCORES FOR TRANSFERRING THE MODEL TO THE SEGMENTATION TASK ON *GLaS*. THE COMPARISONS ARE MADE ON DIFFERENT MID-LEVEL FEATURES WITH THE SPATIAL RESOLUTION M.

Methods	M**=64**	M**=32**	M**=16**	M**=8**
DIM [14]	0.850±0.008	0.876±0.003	0.865±0.004	0.866±0.003
DRL	0.858±0.004	0.882±0.005	0.881±0.004	0.873±0.008

Besides, we make a comprehensive comparison with the closely related work (i.e., DIM [14]) in the evaluation. Note that our approach mainly differs from DIM in the following aspects. (1) We introduce LSH and GCH using deformation as supervisory signals to encode representative features in histopathology images with different objective functions. They introduced local mutual information to improve the representation for classification. (2) Our deformed pairs in the local objective preserve the spatial alignment, while DIM randomly samples instances or different patches within an instance. Also, the generated deformed pairs are much more difficult to be discriminated than randomly sampled pairs. Table II validates the effectiveness of deformation compared with randomly sampled pairs in DIM. Also, the p-values in Table I demonstrate that DRL consistently outperforms DIM on the segmentation performance (p = 0.0011).

2) Semi-Supervised Classification on PCam: **Low-data classification** First, the pretext task is trained on the entire unlabeled dataset. In the low-data regime, the entire network is fine-tuned on a randomly selected subset of the labeled data. The proportion of labeled data ranges from 0.01% to 100%. The network is trained from scratch in a fully-supervised manner to form the baseline result. As shown in Table III,

TABLE III

THE LOW-DATA CLASSIFICATION ACCURACY ON *PCam*. WE INITIALIZE THE NETWORK WITH THE PRE-TRAINED WEIGHTS AND FINE-TUNE THE ENTIRE NETWORK ON AN INCREASING PROPORTION OF LABELED DATA. THE BASELINE METHOD IS INITIALIZED WITH RANDOM WEIGHTS. THE p-VALUES ARE CALCULATED BETWEEN OUR DRL AND DIM.

Methods	**0.01%**	**1%**	**10%**	**50%**	**100%**
VAE [29]	64.33±1.02	79.52±1.22	80.78±2.02	87.39±0.43	88.14±1.34
MG [15]	62.58±2.45	78.59±0.73	80.99±0.59	86.31±0.68	88.30±1.09
Rotation [6]	62.69±1.84	80.57±1.31	83.26±1.62	86.30±0.53	88.44±1.06
Colorization [4]	65.92±0.46	77.44±1.14	80.46±2.24	83.32±1.10	85.34±1.61
CENet [5]	62.79±0.78	79.93±0.81	83.35±1.36	87.60±1.24	88.85±2.06
Exemplar [12]	66.08±1.83	81.30±0.62	82.97±2.08	86.51±0.79	85.93±0.92
DIM [14]	63.53±0.08	79.22±0.31	81.26±1.28	85.88±0.54	89.18±0.70
Random init.	53.09±2.89	64.24±1.11	77.71±2.71	80.64±0.93	83.14±1.30
ImageNet	63.66±1.38	**83.58±1.90**	84.53±1.17	87.66±0.97	88.69±1.24
DRL	**66.18±0.93**	79.73±0.24	**85.48±0.88**	**88.47±0.52**	**89.35±0.08**
p-value	0.0056	0.2219	0.0467	0.0445	0.0311

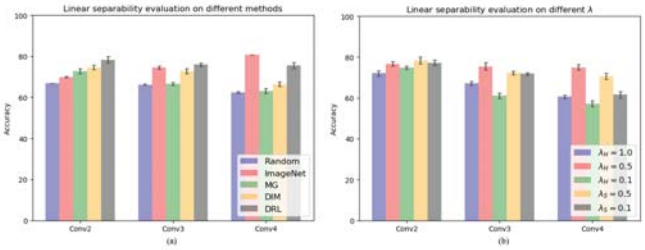

Fig. 4. The linear separability evaluation on different methods (a) and the balancing parameter λ (b). λ_H and λ_S denote hard-target GCH and soft-target GCH, respectively. The error bars are generated based on five individual runs.

when the amount of labeled data reduces from 10% to 0.01%, the performance of the baseline model (Random init.) decreases significantly (from 77.71% to 53.09%), indicating that it suffers from the potential overfitting. In contrast, our model consistently retains the strong performance in both low-data and high-data regime. To the best of our knowledge, our DRL with the plain VGG backbone also performs comparably with the state-of-the-art fully supervised method with specialized architecture and data augmentation techniques [31] (i.e., 89.35% vs. 91.87%).

Linear separability To evaluate the linear separability, we fix the feature extractor and train a linear classifier on the top of each convolution block (from Conv2 to Conv4). We do not include the first convolution block for comparison as the network is too shallow to encode enough image-level features. As can be seen from Fig. 4(a), our self-supervised DRL achieves superior results over other methods on Conv2 and Conv3. The mid-level features that DRL learns are valuable for downstream classification tasks, which is in accordance with the conclusions in prior work [6]. The p-values of Conv2, Conv3, and Conv4 are 0.0031, 0.0113, and 0.0176, respectively, which exhibits a significant improvement of DRL over DIM (p<0.05). Both evaluations on semi-supervised learning prove that our learned features are adaptive to the classification task.

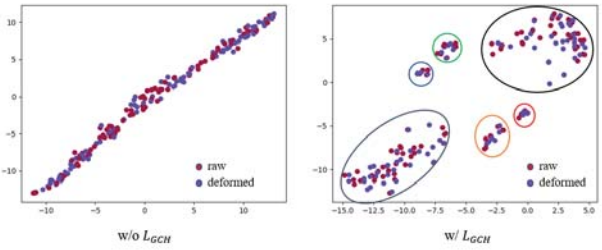

Fig. 5. The t-SNE visualizations of the raw images and their deformed counterparts on *GLaS* w/o or w/ $\mathcal{L}_{GCH}$.

D. Ablation Study

The ablation study of each component in our DRL is shown in Tables II and IV. First, we replace the random batch-wise sampling strategy used in DIM [14] with our constructed deformed pair. As shown in Table II, the introduction of deformation leads to superior performance, regardless of which mid-layer the representations are transferred from. In Table IV, incorporating DRL($\mathcal{L}_{LSH}$) only into the pretext task outperforms the baseline network by a large margin, which validates that deformation, as a supervisory signal, brings more contrastive information that facilitates the learned representations. Table IV also indicates that integrating either Hard-Target or Soft-Target global context information brings better segmentation performance. We find that $\mathcal{L}^S_{GCH}$ and $\mathcal{L}^H_{GCH}$ perform comparably on F1 Score and Object Dice while $\mathcal{L}^S_{GCH}$ outperforms $\mathcal{L}^H_{GCH}$ on Object Hausdorff (p=0.0011). Similar results can also be observed in Fig. 4(b), where λ=0.5 performs better on both $\mathcal{L}^H_{GCH}$ and $\mathcal{L}^S_{GCH}$. We adopt $\mathcal{L}^H_{GCH}$ as the default global context homogeneity loss as it is computationally efficient. Additionally, we visualize the distribution of the raw images and their deformed counterparts in the embedding space using t-SNE, as shown in Fig. 5. By introducing $\mathcal{L}_{GCH}$, similar images and their deformed counterparts are gathered into several clusters (marked by different ovals), indicating the network learns a useful embedding space. We also validate FEM by discarding it and replacing $F(x;\theta)$ with the global feature $F_G(x;\theta)$, which is expanded to the feature size to build $\mathcal{L}_{LSH}$. The last row in Table IV signifies the multi-scale feature representation ability of FEM on both the segmentation and classification task.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a self-supervised Deformation Representation Learning (DRL) approach based on local structure heterogeneity and global context homogeneity for enhancing data-efficient histopathology image analysis. In particular, we train a network to distinguish the difference between an original histopathology image and its deformed counterpart in local structures by leveraging the mutual information. Meanwhile, we maintain a consistent global context of the image pair with noise contrastive estimation. Extensive experiments

TABLE IV
THE ABLATION EXPERIMENTS OF FINE-TUNING DRL ON *GLaS* AND *PCam*. THE p-VALUES ARE CALCULATED BETWEEN EACH METHOD AND THE BASELINE NETWORK.

Methods					Segmentation						Classification	
Baseline	$\mathcal{L}_{LSH}$	$\mathcal{L}^{H}_{GCH}$	$\mathcal{L}^{S}_{GCH}$	FEM	F1	p-value	Dice	p-value	Hausdorff	p-value	Accuracy	p-value
✓					0.853±0.009	-	0.861±0.006	-	63.49±4.40	-	83.14±1.30	-
✓	✓				0.883±0.004	0.0027	0.873±0.003	0.0054	64.99±0.94	0.0831	87.23±0.32	0.0114
✓	✓	✓			0.886±0.004	0.0023	0.878±0.002	0.0256	53.69±2.31	0.3321	88.28±0.44	0.0080
✓	✓		✓		0.884±0.004	0.0061	0.884±0.004	0.0033	55.71±1.92	0.0031	88.02±0.11	0.0081
✓	✓	✓		✓	**0.900±0.006**	0.0002	**0.896±0.004**	0.0002	**50.55±2.65**	0.0035	**89.35±0.08**	0.0006

demonstrate that the learned high-level representations are generalizable for various downstream tasks with a portion of labeled data. The advantage of DRL lies in its encoding ability of the informative features of histopathology images in the local and global manner. Our future work will mainly focus on adapting our DRL framework to multimodal biomedical data over various diseases and organs.

REFERENCES

[1] H. Chen, X. Qi, L. Yu, and P.-A. Heng, "Dcan: deep contour-aware networks for accurate gland segmentation," in *CVPR*, 2016, pp. 2487–2496.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[3] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020.

[4] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *ECCV*, 2016, pp. 649–666.

[5] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016, pp. 2536–2544.

[6] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *ICLR*, 2018.

[7] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *NeurIPS*, 2019, pp. 3342–3352.

[8] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[9] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. V. Guttag, and A. V. Dalca, "An unsupervised learning model for deformable medical image registration," in *CVPR*, 2018, pp. 9252–9260.

[10] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR 2020*, 2020, pp. 9729–9738.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML 2020*, 2020.

[12] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. A. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *TPAMI*, vol. 38, no. 9, pp. 1734–1747, 2016.

[13] L. Zhang, G.-J. Qi, L. Wang, and J. Luo, "Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data," in *CVPR*, 2019, pp. 2547–2555.

[14] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *ICLR*, 2019.

[15] Z. Zhou, V. Sodha, M. M. R. Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, "Models genesis: Generic autodidactic models for 3d medical image analysis," in *MICCAI*, 2019, pp. 384–393.

[16] S. Graham, H. Chen, J. Gamper, Q. Dou, P. Heng, D. R. J. Snead, Y. Tsang, and N. M. Rajpoot, "Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images," *Medical Image Analysis*, vol. 52, pp. 199–211, 2019.

[17] Z. Zhou, M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation." *IEEE Transactions on Medical Imaging*, pp. 1–1, 2019.

[18] Z. Jia, X. Huang, E. I.-C. Chang, and Y. Xu, "Constrained deep weak supervision for histopathology image segmentation," *IEEE Transactions on Medical Imaging*, vol. 36, no. 11, pp. 2376–2388, 2017.

[19] C. T. Sari and C. Gunduz-Demir, "Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images," *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1139–1149, 2019.

[20] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *ICDAR*, 2003, pp. 958–962.

[21] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, "Mine: mutual information neural estimation," *arXiv preprint arXiv:1801.04062*, 2018.

[22] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *NIPS*, 2016, pp. 2172–2180.

[23] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[24] Z. Qin and D. Kim, "Rethinking softmax with cross-entropy: Neural network classifier as mutual information estimator," *CoRR*, vol. abs/1911.10688, 2019.

[25] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *NIPS*, 2016, pp. 1857–1865.

[26] K. Sirinukunwattana, J. P. W. Pluim, H. Chen, X. Qi, P. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, A. Böhm, O. Ronneberger, B. B. Cheikh, D. Racoceanu, P. Kainz, M. Pfeiffer, M. Urschler, D. R. J. Snead, and N. M. Rajpoot, "Gland segmentation in colon histology images: The glas challenge contest," *Medical Image Analysis*, vol. 35, pp. 489–502, 2017.

[27] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant cnns for digital pathology," in *MICCAI*, 2018, pp. 210–218.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[29] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[30] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *CVPR*, 2019, pp. 6210–6219.

[31] Y. Huang and A. C. S. Chung, "Evidence localization for pathology images using weakly supervised learning," in *MICCAI*, ser. Lecture Notes in Computer Science, vol. 11764, 2019, pp. 613–621.