Optimal Transport and Contrastive-Based Clustering for Annotation-Free Tissue Analysis in Histopathology Images

Mohammed Aburidi
Department of Applied Mathematics
University of California Merced
Merced, USA
maburidi@ucmerced.edu

Roummel Marcia
Department of Applied Mathematics
University of California Merced
Merced, USA
rmarcia@ucmerced.edu

2023 International Conference on Machine Learning and Applications (ICMLA) | 979-8-3503-4534-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICMLA58977.2023.00049

***Abstract*—**
Training a deep learning model with a large annotated dataset is still a dominant paradigm in automatic whole slide images (WSIs) processing for digital pathology. However, obtaining manual annotations is a labor-intensive task, and an error-prone to inter and intra-observer variability. In this study, we offer an online deep learning-based clustering workflow for annotating and analysis of different types of tissues from histopathology images. Inspired by learning and optimal transport theory, our proposed model consists of two stages. In the first stage, our model learns tissue-specific discriminative representations by contrasting the features in the latent space at two levels, the instance- and the cluster-level. This is done by maximizing the similarities of the projections of positive pairs (views of the same image) while minimizing those of negative ones (views of the rest of the images). In the second stage, our framework extends the standard cross-entropy minimization to an optimal transport problem and solves it using the Sinkhorn-Knopp algorithm to produce the cluster assignments. Moreover, our proposed method enforces consistency between the produced assignments obtained from views of the same image. Our framework was evaluated on three common histopathological datasets: NCT-CRC, LC2500, and Kather_STAD. Experiments show that our proposed framework can identify different tissues in annotation-free conditions with competitive results. It achieved an accuracy of 0.9364 in human lung patched WSIs and 0.8464 in images of human colorectal tissues outperforming state of the arts contrastive-based methods.

***Index Terms*—Clustering, Optimal Transport, Contrastive Learning, Tissue Recognition, Digital Pathology, WSIs.**

I. Introduction

In the last decade, deep learning has achieved considerable progress in the field of medical image analysis and its applications [1], [2]. However, the deployment of deep learning-based methods in clinical applications is slow. One of the main challenges is the lack of high-quality annotated data required for training these models with a high degree of predictability. The manual annotation of medical data is a labor-intensive and error-prone task and relies on medical knowledge from experts.

The data annotation bottleneck is well apparent in histopathology images, one type of a widely used medical imaging modality and is considered the gold standard for cancer diagnosis [3]. With the advance in imaging techniques, digital scans generate Whole Slide Images (WSIs) from histopathology slides. WSIs are multi-giga-pixel and high-resolution images that capture the whole tissue in the slide. The common approach to utilize such huge images in a deep learining based solution is to subdivide them into small patches, where each WSI outputs thousands of patches of different tissues where each patch is processed independently in the neural network [4]. However, the large number of these patches makes the task of annotating the WSI at the patch level (i.e., local annotations) infeasible.

To solve the annotation scarcity problem, many efforts have been made by researchers to develop annotation efficient deep neural networks based training methods for WSI analysis. Current popular methods can be divided into two categories: semi-supervised, and self-supervised methods. For the semi-supervised methods [5], [6], learning is done using small amounts of labeled data, but a larger amount of unlabeled data is used to boost the ultimate performance. In self-supervised learning (SSL) [7], [8], pre-trained models are created without the need for large and annotated datasets by means of a proxy objective, for which labels are self generated. A common drawback in the two categories, is that all requires a small certain amount of manual labels.

In this paper, we extend a model that we proposed in [9], which combines contrastive learning with optimal transport (CLOT) for on-line clustering. This method was designed and evaluated on natural scene images, but here, we test its performance on a real-life and challenging problem, namely for self-generating tissue labels for hisopathology images at the patch level. The extended version is a deep learning-

This research is partially supported by NSF Grant IIS 1741490 and DMS 1840265.

2023 International Conference on Machine Learning and Applications (ICMLA) | 979-8-3503-4534-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICMLA58977.2023.00049

1946-0759/23/$31.00 ©2023 IEEE
DOI 10.1109/ICMLA58977.2023.00049

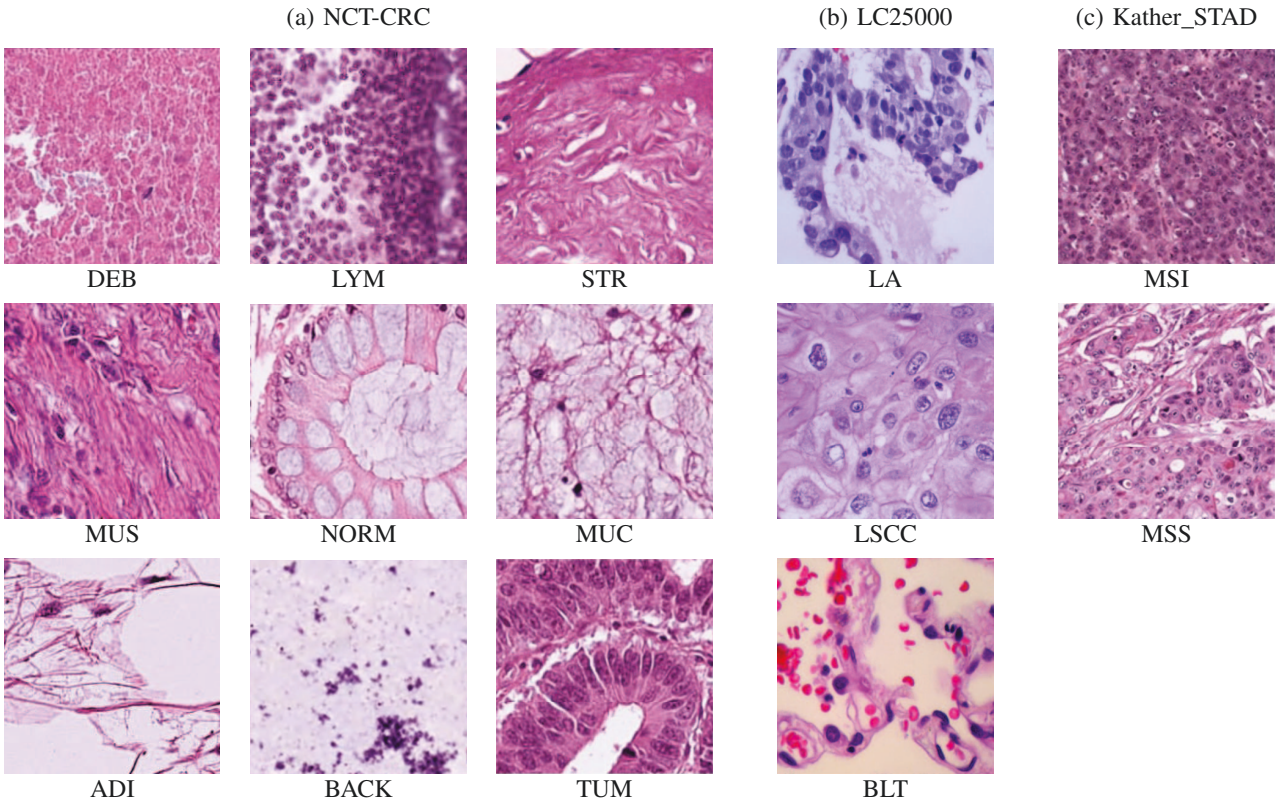


Fig. 1. Example of patches of three different datasets: (a) NCT-CRC (nine classes), (b) LC25000 (three classes), and (c) Kather_STAD (two classes). Each image belongs to a different class.

based clustering method works simultaneously and in a dual fashion. The proposed model consists of two stages. In the first stage, instance- and cluster-level representations are learned by maximizing the similarities of the projections of positive pairs while minimizing those of negative ones, thus pushing away features from different images while pulling together those from the augmented views of the same image. In the second stage, our framework extends the standard cross-entropy minimization to an optimal transport problem and solves it using a fast variant of the Sinkhorn-Knopp algorithm to produce the cluster assignments. Moreover, our framework utilizes a multi-loss objective for robust training, that compares the class assignments obtained from solving the self-labeling in an online fashion as an optimal transport, and enforce consistency between the produced assignments obtained from views of the same image.

We evaluated our framework on three common histopathology images: (1) NCT-CRC, a colorectal cancer tissue dataset; (2) LC25000, a lung histopathological dataset; and (3) Kather_STAD, which includes images of microsatellite instable (MSI) versus microsatellite stable (MSS) image patches of gastric (stomach) cancer (see Fig. 1). We describe these datasets in more detail in Sec. IV-A. Our proposed framework achieves an up-to-93% performance in terms of AUC on the LC25000 dataset, 89% on Kather_STAD, and 83% on clustering 9 tissues of the NCT-CRC dataset.

II. Related work

A. Self-supervised learning

Self-supervised learning (SSL) is a subclass of unsupervised learning has recently gained significant attention in many medical image analysis tasks. In SSL, the objective relies only on the data itself by obtaining feature-rich latent space representations without the need for manual annotations. One category of SSL is a class of discriminative methods that is proposed based on contrastive learning which learns to maximizes similarities between the latent space feature vectors of two augmented views encoded from the same image [8]. The most common contrastive based method is called SimCLR proposed by [7]. The first extension of SimCLR to histopathology was done by [10] where authors combined multiple instance with contrastive learning for weakly supervised histopathology classification. Multiple extension have been proposed afterwhile, however, very few were evaluated on histopathology datasets. A recent one is called Contrastive Clustering (CC) proposed by [11] and evaluated on histopathlogy images by [12].

B. Optimal transport

In our method we extend contrastive learning clustering proposed in [11] and conjunct it with optimal transport theory

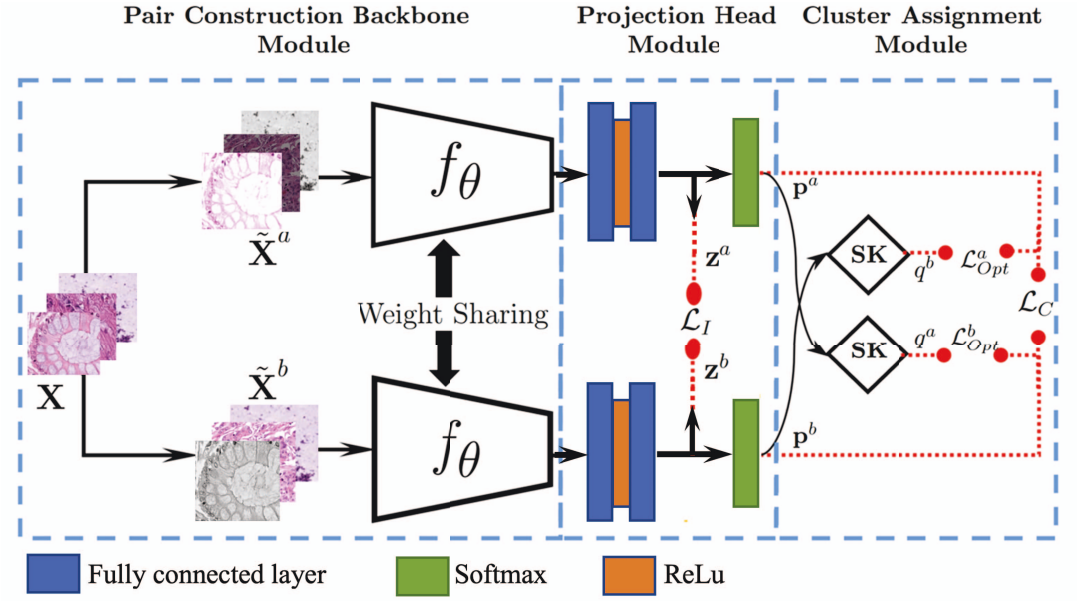


Fig. 2. CLOTpath clustering framework. In the first module, image patches are passing an augmentation module where pairs of patches $\tilde{\mathbf{X}}^a, \tilde{\mathbf{X}}^b$ are constructed using two augmentations of data $\mathbf{X}$. Then the features are extracted from the pairs using a shared encoder f_θ. In the second module, a multilayer perceptron (MLP) model is used as projection head that projects the features into latent space and outputs the feature vectors $\mathbf{z}^a$ and $\mathbf{z}^b$. The MLP is then followed by a softmax function, which outputs the probability vectors $\mathbf{p}^a$ and $\mathbf{p}^b$. Finally, the cluster assignment probabilities are used by the Sinkhorn-Knopp (SK) algorithm to generate ground-truth-like cluster assignment in a contrastive and consistent approach.

and evaluate it to self-generate labels for annotation-free tissues from histopathology images. Optimal transport [13] is a mathematical framework that defines the problem of finding the most efficient way (i.e., lowest cost) of moving an object such as probability distribution from one configuration onto another (e.g., matching two distributions or finding the similarity between two distributions). Optimal transport has been gaining in recent years increasing attention as a promising and useful tool in the machine-learning community. This success is due to its capacity to exploit the geometric property of the samples at hand. Optimal transport methods have been successfully employed in a wide variety of machine learning applications [14]–[18], computer vision [19], [20], generative adversarial networks, domain adaptation [21]. Recently, applications of optimal transport to biology have also been proposed [22]–[25].

III. Method

Fig. 2 shows an overview of the proposed method. It consists of three modules: a pair construction backbone, a projection head, and a cluster assignment module. For a given WSI tile, we first compute two different augmentations for each view (a positive pair), and then pass them through a backbone, followed by two parallel projection heads. One head is used to compute the feature vector $\mathbf{z}$, and the second is similar to the first except it projects into a subspace with a dimensionality corresponding to the number of clusters, which could be interpreted as the cluster assignment probabilities $\mathbf{p}$ (i.e., instance soft labels). The probabilities are then used as an input to the Sinkhorn-Knopp algorithm to find the cluster assignments by solving the problem as an optimal transport. In this section, we explain the core units of the models. We start by explaining the cluster assignment module, and how labels are self-generated using optimal transport. The, the second unit is illustrated, in which contrastive learning at the two mentioned levels are merged into the model to strengthen it.

A. Online Computing of Cluster Assignments

Building upon [26], we encode the cluster labels as posterior distributions $q(y = k|x_i)$, and we formulate the problem of finding optimal assignments as an optimal transport optimization problem.

Consider a given mini-batch $\mathbf{X}$ of N images $\{x_1, ..., x_N\}$, we compute their predicted cluster assignment probabilities or the cluster-level representations $\mathbf{p} \in \mathbb{R}^{N\times K}$ using an encoder network f_θ, and a projection head $g_1(\cdot)$ (two stacked nonlinear multilayer perceptron (MLP) layers followed by a softmax), where K is the number of clusters. We compare the class label predictions $\mathbf{p}$ with the label assignments obtained when solving the optimization problem.

To formulate the problem of finding the labels mathematically, we encode the labels as posterior distributions in the average cross-entropy objective [27], [28]. In this case, our loss will be

$$\mathcal{L}_{Opt}(p, q) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} q(y = k|x_i)\log p(y = k|x_i) \quad (1)$$

where the values in the vector $p(y|x_i) \in \{p_i^a, p_i^b\}$, and $q(y|x_i) \in \{q_i^a, q_i^b\}$. Optimizing q is the same as reassigning

the labels, which leads to a degenerate solution, i.e., (1) can be trivially minimized by assigning all data points to a single and arbitrary class label. A common way to avoid this is by adding a constraint that enforces an equally-sized partition [28]. The learning objective objective is thus

$$\begin{aligned} \underset{q}{\text{minimize}} \quad & \mathcal{L}_{Opt}(p,q) \\ \text{subject to} \quad & \sum_{i=1}^{N} q(y=k|x_i) = \frac{N}{K}, \quad q(y=k|x_i) \in \{0,1\}. \end{aligned} \tag{2}$$

At this step, we only optimize the labels, keeping the predictions p fixed, given a batch of images. The constraints means that each data point x_i is assigned to exactly one class label and the N data points are split equally among the K classes. By reforming it as an optimal transport using the notations in [26], let $P_{y,i} = p(y|x_i)$ be the $K \times N$ matrix of joint probabilities which is estimated by the model, and $Q_{y,i} = q(y|x_i)/N$ be the $K \times N$ matrix of assigned joint probabilities. Using the notation of [26], we restrict the matrix Q to the transportation polytope $\mathbf{Q} = \{Q \in \mathbb{R}^{K \times N} \mid Q\mathbf{1}_N = \frac{1}{K}\mathbf{1}_K, Q^T\mathbf{1}_K = \frac{1}{N}\mathbf{1}_N\}$, where $\mathbf{1}_N$ denotes the vector of ones in dimension N. The constraints enforce that the matrix Q splits the data uniformly. We then can rewrite the optimization problem (2) as

$$\underset{Q \in \mathbf{Q}}{\text{minimize}} \langle Q, -\log P \rangle \tag{3}$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius dot-product of two matrices. This optimization problem is linear optimization, and we would solve it using the last version of Sinkhorn-Knopp algorithm [26], which amounts by introducing a regularization term

$$\underset{Q \in \mathbf{Q}}{\text{minimize}} \quad \langle Q, -\log P \rangle - \frac{1}{\lambda} S(Q) \tag{4}$$

where $S(Q) = -\sum_{i=1}^{N}\sum_{j=1}^{K} q_{ij} \log q_{ij}$ is the entropy. This problem can be solved using the Lagrange multiplier for the entropy constraint of Sinkhorn distances [26], and its minimizer can be written as

$$Q = \text{Diag}(u)\, P^{\lambda}\, \text{Diag}(v), \tag{5}$$

where u and v are normalization vectors chosen such that the resulting matrix Q is also a probability matrix (see [26] for a derivation). Once Q is found, we optimize the overall objective defined next section to find the optimal P (i.e., the model parameters).

B. Enforcing Consistancy

In order to build an image-transformation invariant model, we use augmentations to generate two stochastic-based views $\tilde{\mathbf{X}}^a$ and $\tilde{\mathbf{X}}^b$ of the given mini-batch $\mathbf{X}$. Passing it through the model (f_θ, and $g_1(\cdot)$), we obtain the predicted cluster assignment probabilities or the cluster-level representations $\mathbf{p}^a$, $\mathbf{p}^b \in \mathbb{R}^{N \times K}$.

Because these two cluster assignment probabilities come form the same image, it capture the same information. Therefore, to enforce consistancy, we compare the class label predictions obtained from the first augmented view $\mathbf{p}^a$ with the label assignments Q^b obtained when solving the optimization problem, and the predictions obtained from the second augmented view $\mathbf{p}^b$ with the label assignments Q^a.

C. Contrastive Learning

Contrastive learning maximizes the similarities of positive pairs (i.e. the transformed views of the same image) while minimizing those of negative ones by pushing away features from different images while pulling together those from the augmented views of the same image.

The idea of contrastive learning is to compute the latent space feature matrices $\mathbf{z}^a, \mathbf{z}^b \in \mathbb{R}^{N \times D}$, where the rows in these matrixes are the feature vectors of the two augmented views obtained using an encoder network f_θ, and a projection head $g_2(\cdot)$ (two stacked nonlinear MLP layers but without the softmax layer). For a specific sample x_i^a, there are $2N-1$ pairs in total, among which we choose its corresponding augmented sample x_i^b to construct the positive pair $\{x_i^a, x_i^b\}$, and leave the rest $2N-2$ to be negative. The features $\mathbf{z}^a, \mathbf{z}^b$ in this case are the instance representations. Because these two latent space feature vectors come form the same image, it should capture the same information. Therefore, we apply instance-level contrastive loss used in [11] to contrast them and assure they are the same which is of the form

$$\mathcal{L}_{I,i}^a = -\log \frac{\exp\left(\frac{s(z_i^a, z_i^b)}{\tau_I}\right)}{\sum_{j=1}^{N}\left\{\exp\left(\frac{s(z_i^a, z_j^a)}{\tau_I}\right) + \exp\left(\frac{s(z_i^a, z_i^b)}{\tau_I}\right)\right\}}, \tag{6}$$

where $s(\cdot,\cdot)$ is the pair-wise cosine distance, and z_i^a and z_i^b are two corresponding rows from the feature matrices $\mathbf{z}^a$ and $\mathbf{z}^b$, respectively. Here, τ_I is the instance-level temperature parameter [29] that is used to control the "softness" of this loss function.

To fully utilize contrastive learning, we further contrast not only the feature vectors but also the columns of the predicted probability vectors (i.e., cluster-level representations) in the matrices $\mathbf{p}^a$, $\mathbf{p}^b \in$ obtained using the first projection head $g_1(\cdot)$. Similarly, the cluster-level representation loss is utilized to distinguish cluster-level representations of positive pairs from the rest as follows

$$\mathcal{L}_{C,i}^a = -\log \frac{\exp\left(\frac{s(p_i^a, p_i^b)}{\tau_c}\right)}{\sum_{j=1}^{K}\left\{\exp\left(\frac{s(p_i^a, p_i^a)}{\tau_c}\right) + \exp\left(\frac{s(p_i^a, p_i^b)}{\tau_c}\right)\right\}}, \tag{7}$$

where p_i^a and p_i^b are two corresponding columns from the probability matrices $\mathbf{p}^a$ and $\mathbf{p}^b$, respectively, that come from the first projection head. Here τ_c is the cluster-level temperature parameter. To include every possible positive pair across

the dataset, the instance-level contrastive loss and the cluster-level contrastive loss are as follows:

$$\mathcal{L}_I = \frac{1}{2N}\sum_{i=1}^{N}(\mathcal{L}^a_{I,i}+\mathcal{L}^b_{I,i}),$$

$$\mathcal{L}_C = \frac{1}{2K}\sum_{i=1}^{K}(\mathcal{L}^a_{C,i}+\mathcal{L}^b_{C,i}) - S(\mathbf{p}),$$

where $S(\mathbf{p}) = -\sum_{i=1}^{K}[p^a_i \log p^a_i + p^b_i \log p^b_i]$ is the entropy of cluster assignment probabilities added to prevent assigning all instances within the mini-batch to the same cluster [39]. The functions $\mathcal{L}^b_{I,i}$ and $\mathcal{L}^b_{C,i}$ are defined similarly as in (6) and (7), respectively.

D. Objective Function

In our method, the optimization is done in an end-to-end process. The parameters θ of the backbone and the two heads are simultaneously optimized. Thus, the overall objective function consists of (1) the instance-level contrastive loss, (2) the cluster-level contrastive loss, and (3) the two cross-entropy loss functions that enforce the consistency:

$$\mathcal{L}(z,p) = \mathcal{L}_I + \mathcal{L}_C + \mathcal{L}^a_{Opt} + \mathcal{L}^b_{Opt} \quad (8)$$

Our objective enables a robust training at both the latent feature and the code assignment levels. In general, we solve two optimization problems: the first is to find the labels and the second is to find the predictions of the model (i.e. the model parameters). We do so, by first initializing the mode parameters randomly and then by alternating between following two steps:

1) Given the current model's parameters θ, we first compute the log probabilities P, then, we find Q using (5).
2) Given the current label assignments Q, we optimize the model parameters θ by minimizing (8). This step is the same as training the model but with a multi-loss function.

IV. Experiments and results

The proposed method was evaluated on three publicly available WSI datasets: 1) the NCT-CRC [30], a colorectal cancer tissue dataset, 2) Kather_STAD [31], which has histological images of gastric (stomach)) cancer patients whom their tumor shows microsatellite stablity (MSS) versus patients whom their tumor shows microsatellite instablity (MSI) 3) LC25000 [32], a lung histopathological dataset. The three datasets were pre-processed into WSI patches.

A. Datasets

1) NCT-CRC: NCT-CRC consists 100000 non-overlapping 224 $\times$ 224 pixels image patches extracted at 0.5 microns per pixel from hematoxylin and eosin (H&E) stained histological images of human colorectal cancer and normal colon tissue. Each patch is assigned a single label and classified into one of 9 classes of tissues by pathologists including: Adipose (ADI), Cancer associated Stroma (STR), Debris (DEB), Mucus (MUC), smooth Muscle (MUS), Normal Colon Mucosa (NORM), Lymphocytes (LYM), Colorectal Adenocarcinoma Epithelium (TUM), Background (BACK).

2) LC25000: LC25000 contains 15000 patches of size 768 $\times$ 768 pixels. All patches are assigned a single label out of 3 possible classes: lung adenocarcinomas (LA) lung squamous cell carcinomas (LSCC) and benign lung tissues (BLT).

3) Kather_STAD: Kather_STAD contains 100570 patches belongs to 315 WSIs extracted from TCGA [33]. Each patch is assgined to one of the two classes: microsatellite stable tumor (MSS), and microsatellite instable tumor (MSI). Microsatellite instability determines whether patients with gastric cancer respond exceptionally well to immunotherapy. Therefore, its important for a patient to be tested for microsatellite instability, which is not always available in clinical practice because it requires additional genetic tests.

B. Implementation Details

We implement ResNet34 as an encoder backbone architecture [8] and use the Adam optimizer [34] to simultaneously optimize the two projection heads and the backbone network, with cosine learning rate scheduler [35]. The weight decay is set to 0.0001. ResNet is designed for images of size 224×224, so we resize all input images to this size. Both projection heads consists two-layer nonlinear MLP. ReLU activation was used in between the two layers. Softmax activation was used in the in the cluster-level contrastive projection head to produce soft labels as in [11]. Following [7] we set the dimension of the latent vector to 128 and the temperatures parameters to 0.5. The batch size is set to 256 due to the memory limitation. All the models are trained from scratch for 1000 epochs. The training is carried out on UC Merced Pinnacles Cluster using one 2x NVIDIA Tesla A100 PCIe v4 40GB HBM2 Single GPU.

C. Data Augmentations

Following [7], [11] we use random cropping, color jittering, grayscale transformation, horizontal flipping, and Gaussian blurring for augmentation. Each transformation is applied with a certain probability.

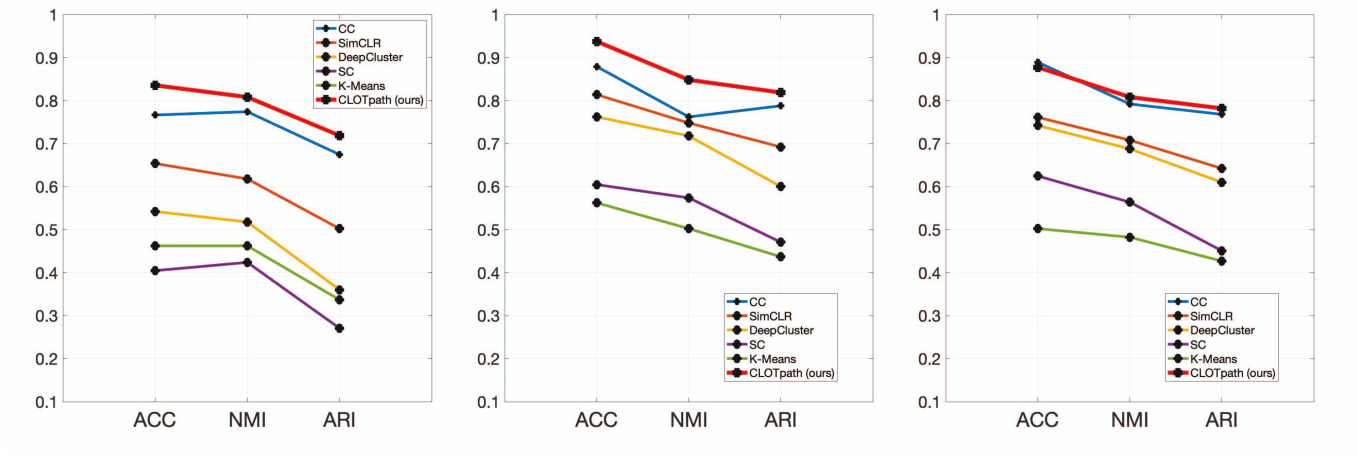
D. Evaluation Metrics

We utilize three common clustering evaluation metrics including Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) to evaluate our model and compare it with baselines. Higher values indicate better performance. We also used one-vs-rest multiclass receiver operating characteristic (ROC) curve and we report the area under curve (AUC) as an image-level diagnosis evaluation metric. These metrics were calculated between clustering predicted labels and ground-truth labels.

E. Comparison Study

Figure 3 shows the comparisons between our method with five state-of-the-art clustering methods, including K-means [36], Spectral Clustering (SC) [37], Contrastive Clustering (CC) [11], SimCLR [7], and DeepCluster [38].

Results shown in Figure 3 demonstrate the clustering ability of CLOTpath, which outperforms the baselines by a large



(a) NCT-CRC (b) LC25000 (c) Kather_STAD

Fig. 3. The clustering performance on our three histopathology tissue image benchmarks: (a) NCT-CRC, (b) LC25000, and (c) Kather_STAD.

margin on all of the three datasets. Specifically, CLOTpath outperforms the closest competitor (CC) on the three datasets in terms of the three evaluation measures. The results demonstrate that this robustness is a result of combining both contrastive learning and contrasting cluster assignments obtained by solving the labeling problem as an optimal transport.

To more specifically investigate the well clustered tissues, we present the confusion matrix (Figure 4) of the model's predictions. In NTC-CRC, five out of nine tissues were clustered with an accuracy of 85% and more and a sixth tissue was clustered with a moderate accuracy of 76%. In contrast, the MUS, DEB, and STR were more mislabeled with each other due to their image structural similarities (see Fig. 1 for details). Likewise, in the LSC25000 dataset, LA was mislabeled more with LSCC because of their common structural and coloration features.

F. Ablation Study

We carried out an ablation study to further understand the contribution of each module in our model. To prove the effectiveness of the projection head module (PHM) and the cluster assignment module (CAM), we conduct ablation studies by removing one of the two modules. Thus, the model parameters are not updated based on the losses $\mathcal{L}_I$ and $\mathcal{L}_C$. The results obtained using NCT-CRC and LC25000 datasets are particularly strong, which we present in Table I. It shows that PHM modules contribute more to the performance of both datasets. The results show performance improvement when both modules are combined.

TABLE I
EFFECTS OF MODEL COMPONENTS

Dataset	Model Component	ACC	NMI	ARI
NCT-CRC	PHM + CAM	**0.8357**	**0.7883**	**0.7189**
	PHM only	0.7666	0.7643	0.6746
	CAM only	0.7411	0.7303	0.6071
LC25000	PHM + CAM	**0.9314**	**0.8401**	**0.8131**
	PHM only	0.8792	0.7622	0.7881
	CAM only	0.8323	0.7270	0.6987
Kather_STAD	PHM + CAM	**0.8862**	**0.8094**	**0.7824**
	PHM only	0.8787	0.7926	0.7802
	CAM only	0.8367	0.7191	0.7047

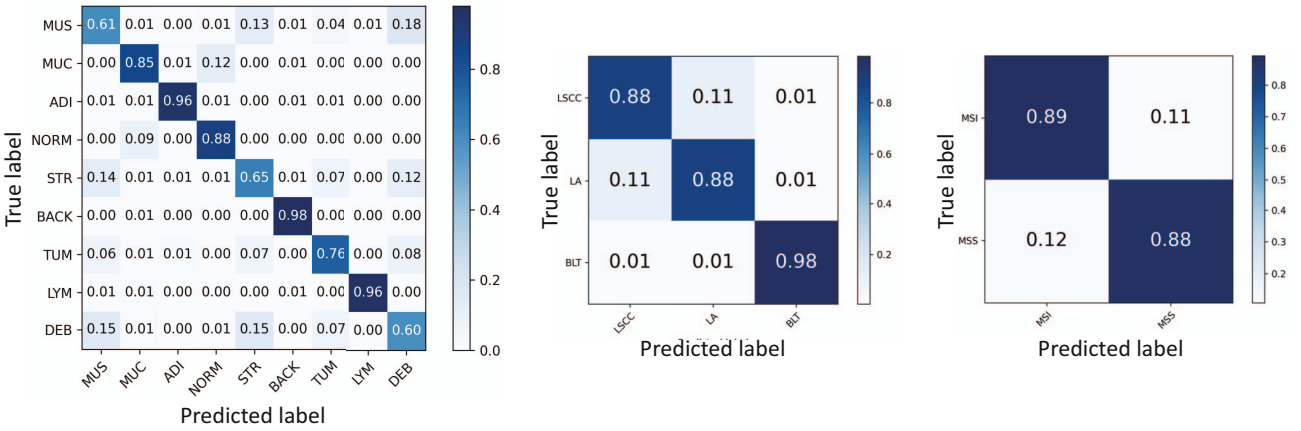
PHM: Projection Head Module, CAM: Cluster Assignment Module

V. CONCLUSION

We present an online deep learning based clustering framework for analyzing and annotating tissues at the patch level in whole slide pathology images. Our proposed model is based on contrastive feature representation learning and contrasting cluster assignments. It also handles the assignment problem as an optimal transport and solves it using Sinkhorn-Knopp algorithm to self-generate the labels. In contrast to other methods, ours optimizes three objectives during feature learning and during clustering, thus providing a robust training setting. Moreover, the tissue discriminative features are learned in two levels, at an instance and cluster level, and more importantly, the objective enforces consistency on the generated labels. Compared to existing state-of-the-art methods, the proposed CLOTpath shows promising performance in clustering on three challenging datasets, and thus, it could be considered as a powerful tool to self generate artificial labels for non-annotated data.

REFERENCES

[1] Litjens, Geert, et al. "A Survey on Deep Learning in Medical Image Analysis." CoRR, vol. abs/1702.05747, 2017, http://arxiv.org/abs/1702.05747.
[2] Skrede, Ole-Johan, et al. "Deep Learning for Prediction of Colorectal Cancer Outcome: A Discovery and Validation Study." The Lancet, vol. 395, no. 10221, 2020, pp. 350–60.
[3] Hamilton, Stanley R., et al. Pathology and Genetics of Tumours of the Digestive System. IARC press Lyon:, 2000.
[4] Tizhoosh, Hamid Reza, and Liron Pantanowitz. "Artificial Intelligence and Digital Pathology: Challenges and Opportunities." Journal of Pathology Informatics, vol. 9, no. 1, 2018, p. 38.



(a) NCT-CRC (b) LC25000 (c) Kather_STAD

Fig. 4. Confusion matrices for the three datasets ((a) NCT-CRC, (b) LC25000, and (c) Kather_STAD) in our numerical experiments.

[5] Li, Jiayun, et al. "An EM-Based Semi-Supervised Deep Learning Approach for Semantic Segmentation of Histopathological Images from Radical Prostatectomies." Computerized Medical Imaging and Graphics, vol. 69, 2018, pp. 125–33.

[6] Sparks, Rachel, and Anant Madabhushi. "Out-of-Sample Extrapolation Utilizing Semi-Supervised Manifold Learning (Ose-Ssl): Content Based Image Retrieval for Histopathology Images." Scientific Reports, vol. 6, no. 1, 2016, pp. 1–15.

[7] Chen, Ting, et al. "A Simple Framework for Contrastive Learning of Visual Representations." International Conference on Machine Learning, PMLR, 2020, pp. 1597–607.

[8] He, Kaiming, et al. "Momentum Contrast for Unsupervised Visual Representation Learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–38.

[9] Aburidi, Mohammed, and Roummel Marcia. "CLOT: Contrastive Learning-Driven and Optimal Transport-Based Training for Simultaneous Clustering." 2023 IEEE International Conference on Image Processing (ICIP). IEEE, 2023.

[10] Lu, Ming Y., et al. "Semi-Supervised Histology Classification Using Deep Multiple Instance Learning and Contrastive Predictive Coding." arXiv Preprint arXiv:1910.10825, 2019.

[11] Li, Yunfan, et al. "Contrastive Clustering." CoRR, vol. abs/2009.09687, 2020, https://arxiv.org/abs/2009.09687.

[12] Yan, Jiangpeng, et al. "Deep Contrastive Learning Based Tissue Clustering for Annotation-Free Histopathology Image Analysis." Computerized Medical Imaging and Graphics, vol. 97, 2022, p. 102053.

[13] Villani, Cédric. Optimal Transport: Old and New. 2008.

[14] Alvarez-Melis, David, et al. Structured Optimal Transport. 2017.

[15] Cañas, Guillermo D., and Lorenzo Rosasco. "Learning Probability Measures with Respect to Optimal Transport Metrics." CoRR, vol. abs/1209.1077, 2012, http://arxiv.org/abs/1209.1077.

[16] Peyré, Gabriel, and Marco Cuturi. Computational Optimal Transport. 2020.

[17] Arjovsky, Martin, et al. Wasserstein GAN. 2017.

[18] Flamary, Rémi, et al. "Wasserstein Discriminant Analysis." Machine Learning, vol. 107, no. 12, May 2018, pp. 1923–45, https://doi.org/10.1007/s10994-018-5717-1.

[19] Ferradans, Sira, et al. "Regularized Discrete Optimal Transport." CoRR, vol. abs/1307.5551, 2013, http://arxiv.org/abs/1307.5551.

[20] Su, Zhengyu, et al. "Optimal Mass Transport for Shape Matching and Comparison." IEEE Trans Pattern Anal Mach Intell, vol. 37, no. 11, Nov. 2015, pp. 2246–59.

[21] Courty, Nicolas, et al. "Optimal Transport for Domain Adaptation." CoRR, vol. abs/1507.00504, 2015, http://arxiv.org/abs/1507.00504.

[22] Bellazzi, Riccardo, et al. The Gene Mover's Distance: Single-Cell Similarity via Optimal Transport. 2021.

[23] Cao, Kai, et al. "Manifold Alignment for Heterogeneous Single-Cell Multi-Omics Data Integration Using Pamona." Bioinformatics, vol. 38, no. 1, Dec. 2021, pp. 211–19.

[24] Demetci, Pinar, et al. "Gromov-Wasserstein Optimal Transport to Align Single-Cell Multi-Omics Data." bioRxiv, 2020, https://doi.org/10.1101/2020.04.28.066787.

[25] Huizing, Geert-Jan, et al. Unsupervised Ground Metric Learning Using Wasserstein Singular Vectors. 2022.

[26] Cuturi, Marco. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport." Advances in Neural Information Processing Systems, edited by C. J. Burges et al., vol. 26, Curran Associates, Inc., 2013.

[27] Asano, Yuki Markus, et al. "Self-Labelling via Simultaneous Clustering and Representation Learning." CoRR, vol. abs/1911.05371, 2019, http://arxiv.org/abs/1911.05371.

[28] Caron, Mathilde, et al. "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments." CoRR, vol. abs/2006.09882, 2020, https://arxiv.org/abs/2006.09882.

[29] Wu, Zhirong, et al. "Unsupervised Feature Learning via Non-Parametric Instance-Level Discrimination." CoRR, vol. abs/1805.01978, 2018, http://arxiv.org/abs/1805.01978.

[30] Kather, Jakob Nikolas, et al. 100,000 Histological Images Of Human Colorectal Cancer And Healthy Tissue. Zenodo, 7 Apr. 2018. DOI.org (Datacite), https://doi.org/10.5281/ZENODO.1214456.

[31] Kather, Jakob Nikolas. Histological Images for MSI vs. MSS Classification in Gastrointestinal Cancer, FFPE Samples. Zenodo, 7 Feb. 2019. DOI.org (Datacite), https://doi.org/10.5281/ZENODO.2530835.

[32] Borkowski, Andrew A., et al. Lung and Colon Cancer Histopathological Image Dataset (LC25000). 2019.

[33] Raju 13, Brigham &. Women's Hospital &. Harvard Medical School Chin Lynda 9. 11 Park Peter J. 12 Kucherlapati, et al. "Comprehensive Molecular Portraits of Human Breast Tumours." Nature, vol. 490, no. 7418, 2012, pp. 61–70.

[34] Kingma, Diederik, and Jimmy Ba. "Adam: A Method for Stochastic Optimization." International Conference on Learning Representations, Dec. 2014.

[35] Loshchilov, Ilya, and Frank Hutter. "SGDR: Stochastic Gradient Descent with Restarts." CoRR, vol. abs/1608.03983, 2016, http://arxiv.org/abs/1608.03983.

[36] MacQueen, J. "Some Methods for Classification and Analysis of Multivariate Observations." Proc. 5th Berkeley Symposium on Math., Stat., and Prob, 1967.

[37] Liu, Jialu, and Jiawei Han. "Spectral Clustering." Data Clustering, Chapman and Hall/CRC, 2018, pp. 177–200.

[38] Caron, Mathilde, et al. "Deep Clustering for Unsupervised Learning of Visual Features." CoRR, vol. abs/1807.05520, 2018, http://arxiv.org/abs/1807.05520.

[39] Hu, Weihua, et al. Learning Discrete Representations via Information Maximizing Self-Augmented Training. arXiv, 2017, https://doi.org/10.48550/ARXIV.1702.08720.