Journal Pre-proof

Graph Temporal Ensembling based Semi-supervised Convolutional Neural Network with Noisy Labels for Histopathology Image Analysis

Xiaoshuang Shi , Hai Su , Fuyong Xing , Yun Liang , Gang Qu , Lin Yang

PII: S1361-8415(19)30160-4
DOI: https://doi.org/10.1016/j.media.2019.101624
Reference: MEDIMA 101624

To appear in: *Medical Image Analysis*

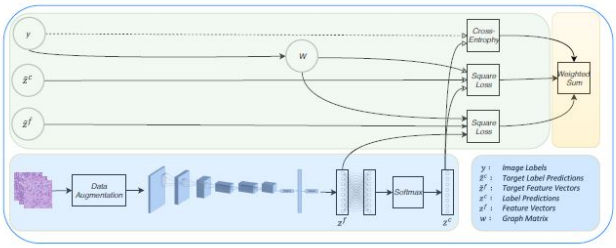
Received date: 26 December 2018
Revised date: 22 November 2019
Accepted date: 25 November 2019

Please cite this article as: Xiaoshuang Shi , Hai Su , Fuyong Xing , Yun Liang , Gang Qu , Lin Yang , Graph Temporal Ensembling based Semi-supervised Convolutional Neural Network with Noisy Labels for Histopathology Image Analysis, *Medical Image Analysis* (2019), doi: https://doi.org/10.1016/j.media.2019.101624

Research highlights

- We utilize the graph to boost the strength of ensemble predictions of training samples.
- We propose a novel loss function to form consensus predicting features and labels.
- We propose a novel self-ensembling based robust semi-supervised deep architecture for histopathology image analysis.

Graphical abstract



Graph Temporal Ensembling based Semi-supervised Convolutional Neural Network with Noisy Labels for Histopathology Image Analysis

Xiaoshuang Shi[a], Hai Su[a], Fuyong Xing[b], Yun Liang[a], Gang Qu[a], Lin Yang[a*]

[a] *J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida*
[b] *Department of Biostatistics and Informatics, University of Colorado Denver*

Abstract

Although convolutional neural networks have achieved tremendous success on histopathology image classification, they usually require large-scale clean annotated data and are sensitive to noisy labels. Unfortunately, labeling large-scale images is laborious, expensive and lowly reliable for pathologists. To address these problems, in this paper, we propose a novel self-ensembling based deep architecture to leverage the semantic information of annotated images and explore the information hidden in unlabeled data, and meanwhile being robust to noisy labels. Specifically, the proposed architecture first creates ensemble targets for feature and label predictions of training samples, by using exponential moving average (EMA) to aggregate feature and label predictions within multiple previous training epochs. Then, the ensemble targets within the same class are mapped into a cluster so that they are further enhanced. Next, a consistency cost is utilized to form consensus

*Corresponding author
Email address: lin.yang@bme.ufl.edu (Lin Yang[a])

predictions under different configurations. Finally, we validate the proposed method with extensive experiments on lung and breast cancer datasets that contain thousands of images. It can achieve 90.5% and 89.5% image classification accuracy using only 20% labeled patients on the two datasets, respectively. This performance is comparable to that of the baseline method with all labeled patients. Experiments also demonstrate its robustness to small percentage of noisy labels.

Keywords: Semi-supervised; Noisy labels; Convolutional neural network; Histopathology image classification

1. Introduction

Because of the advance in high-throughput tissue bank and archiving of digitized histological studies, histopathology images with computer aided diagnosis (CAD) using modern machine learning techniques have attracted considerable attention to facilitate disease grading and classification (Gurcan et al., 2009) (Shen et al., 2017) (Shi et al., 2017) (Xing et al., 2017). Recently, with an ever-increasing amount of images and the development of deep neural networks (Xu et al., 2018) (Xu et al., 2019), especially convolutional neural networks (CNNs), it is promising to bridge the semantic gap between images and diagnostic information (Zhang et al., 2015) (Litjens et al., 2017) (Shi et al., 2018) (Sapkota et al., 2018) (Chen et al., 2019). Most of current deep learning methods require a large amount of clean annotated data to achieve desired performance, and they are usually sensitive to noisy labels, i.e. a very small percentage of noisy labels might severely decrease the model performance (Patrini et al., 2017). Unfortunately, labeling large-scale

histopathology images is laborious, expensive and time-consuming. Additionally, the annotation process is conducted with low labeling reliability, resulting in noisy labels due to the subjective assessment of pathologists. Therefore, it is necessary to design effective and efficient deep neural networks for histopathology image analysis, that can be trained on a small amount of labeled data yet large-scale unlabeled data. Meanwhile, the network should be robust to a small percentage of noisy labels.

To leverage the semantic information of labeled data and meanwhile explore the information hidden in unlabeled data, numerous semi-supervised deep learning methods have been proposed and applied to various applications, such as image classification and retrieval (Zhang and Peng, 2017), detection (Tang et al., 2018) and segmentation (Bai et al., 2017). Among previous semi-supervised deep classification methods, self-ensembling based methods have achieved state-of-the-art accuracy on multiple benchmark image databases. This is because self-ensembling can successfully explore the semantic information hidden in unlabeled data, by creating an ensemble target for each label prediction and forming consensus predictions under different configurations, such as different epochs, dropout regularizations and input augmentations. However, previous self-ensembling based semi-supervised deep learning methods focus on natural images rather than histopathology images, and these two types of images usually have different data distributions. Additionally, most of them fail to consider the relationship among training samples (Luo et al., 2017), and this might result in suboptimal performance on histopathology images.

Due to the subjective assessment of pathologists on disease grading or

classification, it is difficult and costly to obtain large-scale clean labels for histopathology images. Recently, several methods have been developed to enhance the robustness of deep neural networks on noisy labels. Most of existing methods adopt one of the following four strategies: (i) designing robust loss functions (Ghosh et al., 2017); (ii) calculating a transformation matrix (Mnih and Hinton, 2012); (iii) reweighting examples (Ren et al., 2018); (iv) forming consensus predictions under different configurations (Reed et al., 2014). Although these strategies can alleviate the effect of noisy labels, most of them cannot simultaneously explore the semantic information in unlabeled data except several self-ensembling based algorithms, e.g. Temporal Ensembling (TE) (Laine and Aila, 2016). It creates an ensemble target for each label prediction of training samples by applying exponential moving average (EMA) to the predictions within multiple previous training epochs, and then minimizes the difference between the label prediction and its ensemble target. Therefore, in this paper, we focus on improving TE to exploit the semantic information of a small amount of labeled histopathology images, explore the information of unlabeled data, and meanwhile suppress the effect of noisy labels.

Specifically, we propose a novel robust semi-supervised convolutional neural network, namely graph temporal ensembling (GTE). Inspired by TE, the proposed method creates ensemble targets for feature and label predictions of each training sample and forms consensus predictions under different configurations, in order to take advantage of the semantic information from unlabeled data and boost the model robustness to noisy labels (Laine and Aila, 2016). Because TE fails to take into account the relationship among labeled

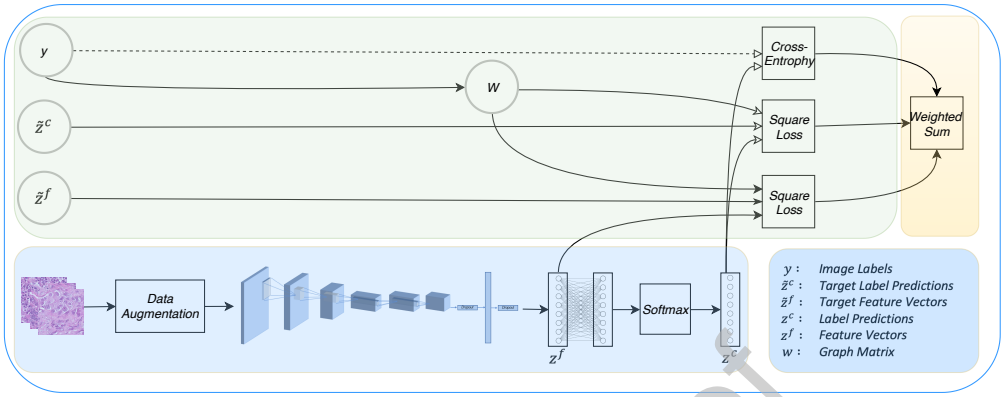


Fig 1: The proposed self-ensembling based deep architecture using AlexNet as the backbone network for histopathology image classification.

training samples, we exploit their connections by utilizing a graph to map all labeled samples from the same class into a single cluster so as to boost the strength of the ensemble targets. To the best of our knowledge, GTE is the first self-ensembling based semi-supervised deep method for histopathology image classification. The major contributions of this paper are summarized as follows:

- We present a novel self-ensembling based deep architecture to leverage the semantic information of unlabeled histopathology images, and meanwhile being robust to noisy labels. For clarity, we show the proposed deep architecture in Fig 1.

- We propose a novel loss function to form consensus feature and label predictions, i.e. creating stronger ensemble targets for feature and label predictions via the graph that maps the targets of labeled training images within the same class into a single cluster, and forming consensus predictions of all training images under different configurations.

- Extensive experiments on lung and breast cancer image datasets demonstrate that (i) The proposed method can achieve superior performance over recent state-of-the-art semi-supervised deep methods and deep neural networks with noisy labels; (ii) Using a graph-based approach (to map labeled samples of each class into a cluster) is more beneficial to semi-supervision compared to feature consistency (to form consensus predictions of feature representations), while feature consistency is more significant on the model robustness to noisy labels.

This paper is organized as follows. Section 2 briefly introduces the related work including semi-supervised deep methods and deep neural networks with noisy labels. Section 3 presents the proposed method GTE. Section 4 shows and analyzes experimental results on histopathology image classification. Section 5 concludes this paper and points out the future work.

2. Related work

In this section, we briefly review the related work: semi-supervised deep learning and deep neural networks with noisy labels.

2.1. Semi-supervised deep learning

Semi-supervised deep learning algorithms can be roughly categorized into five groups: self-training, multi-view, generative adversarial networks (GAN), graph and self-ensembling. Self-training (Yarowsky, 1995; Lee, 2013) utilizes models' own predictions of unlabeled data in order to attain additional semantic information to improve the model generalization. Multi-view based methods (Zhou and Goldman, 2004; Ruder and Plank, 2018) train different

models with different views of data based on the assumption that these views can complement each other and the models can collaborate to boost the performance of each other. GAN based semi-supervised methods (Odena, 2016; Kumar et al., 2017) usually utilize the discriminator of GAN as the classifier to obtain the output containing $K + 1$ (K real classes and one fake class) probabilities and adopt the generator to improve the performance of the discriminator. Graph based methods (Weston et al., 2012; Kipf and Welling, 2016) utilize the transduction of graph to exploit the semantic information of labeled data and meanwhile explore the underlying structure of unlabeled data, i.e. samples that are close in feature spaces should be also close in output spaces (local or global consistency) (Kamnitsas et al., 2018). This is due to the smoothness and clustering assumption. Self-ensembling (Rasmus et al., 2015; Laine and Aila, 2016) is to utilize a single model under different configurations to create a stronger prediction for each training sample, in order to form consensus predictions for boosting model robustness. Because there are numerous semi-supervised deep algorithms, we mainly review the most related ones: graph and self-ensembling based semi-supervised deep methods, several semi-supervised and weakly supervised methods for pathology image classification in the following.

Graph based methods usually construct graphs to preserve the relationship of neighbors and then utilize the transduction of the graph to simultaneously exploit the semantic information of labeled data and explore the underlying structure of unlabeled data. (Weston et al., 2012) apply "shallow" semi-supervised learning techniques to deep neural networks with adding a graph Laplacian regularizer. Diffusion-convolutional neural networks (DCNN) (At-

wood and Towsley, 2016) and graph convolutional networks (GCN) (Kipf and Welling, 2016) are proposed for graph-structured data, and they are transductive and require a pre-constructed graph. (Luo et al., 2018) incorporate one self-ensembling method, e.g. mean teacher (MT) (Tarvainen and Valpola, 2017), into GCN to further improve its performance. Instead of relying on the pre-constructed graph, (Haeusser et al., 2017) seek the association between labeled and unlabeled data using a two-step random walk in a feature space, which starts and ends at labeled samples within the same class through one intermediate unlabeled sample. Later, (Kamnitsas et al., 2018) add a regularizer to capture the global structure hidden in the data. Among graph based methods, the one most related to GTE is (Luo et al., 2018). It is suitable for graph-structured data and requires a pre-constructed graph, while GTE is applied to histopathology images and it constructs the graph based on the labels of training samples in each batch.

Self-ensembling based methods aim to create strong ensemble predictions of training samples, using the output of one single neural network under different training epochs, regularizations, input augmentation conditions, etc. The Γ-model version of ladder network (Rasmus et al., 2015) contains clean and corrupted branches, and the clean branch is to generate proxy labels of corrupted unlabeled data produced by the corrupted branch. Π-model (Laine and Aila, 2016) utilizes two corrupted branches to generate label predictions of training samples, and then applies a consistency cost to minimize the difference between the predictions. Temporal ensembling (TE) (Laine and Aila, 2016) only contains one corrupted branch, and it generates an ensemble target for each label prediction by using EMA to aggregate the predictions of

multiple previous epochs. To further smooth the model, MT (Tarvainen and Valpola, 2017) averages model weights instead of aggregating label predictions. (Su et al., 2019) embed a label propagation step into MT in order to maintain the local and global consistency of predictions. Virtual adversarial training (VAT) (Miyato et al., 2018) proposes a regularization method that utilizes a virtual adversarial loss to measure the local smoothness of the conditional label distribution. Smooth neighbors on teacher graph (SNTG) (Luo et al., 2017) constructs a graph using label predictions of the teacher model to measure the similarity of neighbors, in order to learn the representation smoothly on a low-dimensional manifold. Among the aforementioned graph and self-ensembling based methods, SNTG is the work most related to the proposed method GTE. Their major differences are: (i) SNTG samples several label predictions of the teacher model to construct a graph, while GTE constructs the graph using all given labels; (ii) SNTG only learns feature representations on a low-dimensional manifold, but GTE aims to map feature and label predictions of each class into a cluster, respectively, so that it can create stronger ensemble targets for feature and label predictions.

Semi-supervised deep learning methods have been widely studied in the fields of natural image recognition, structural data and natural language processing (NLP). However, few efforts are devoted to digit pathology image analysis. A cluster-then-label semi-supervised method (Peikari et al., 2018) identifies high-density regions in the data space to help support vector machine (SVM) find the decision boundary. (Sparks and Madabhushi, 2016) present an image query method based on semi-supervised manifold learning. (Kapil et al., 2018) utilize class auxiliary generative adversarial networks

(AC-GANs) for programmed death ligand 1 (PD-L1) scoring on needle biopsies. In addition to semi-supervised algorithms, multiple instance learning (MIL) and transfer learning (Quellec et al., 2017; Cheplygina et al., 2019) are also two popular strategies to reduce the workload of label annotations for pathology images. (Xu et al., 2014) propose an effective MIL method, multiple clustered instance learning, to simultaneously perform image-level classification, medical image segmentation and patch-level clustering. (Kraus et al., 2016) combine deep learning and MIL for medical image classification using only whole image labels. (Chang et al., 2017) propose an unsupervised transfer learning method using multi-scale convolutional sparse coding to learn transferable base knowledge for medical image classification.

2.2. Deep neural networks with noisy labels

Although several methods have been proposed to improve the robustness of deep neural networks to noisy labels, very limited efforts have been devoted to boosting the model robustness to noisy labels for pathology image classification. (Mnih and Hinton, 2012) propose two robust loss functions based on noise distribution to deal with omission and registration noise in aerial images. (Ghosh et al., 2017) demonstrate that the ℓ_1-norm based loss function is more robust than cross-entropy and least-squares loss functions. (Veit et al., 2017) utilize the ℓ_1-norm to clean labels. Unfortunately, the ℓ_1-norm based loss function is much more difficult to converge. (Natarajan et al., 2013; Sukhbaatar and Fergus, 2014; Xiao et al., 2015; Patrini et al., 2017; Vahdat, 2017) model the relationship between images, class labels and label noise with probability graphs and then integrate them into end-to-end deep neural networks. (Ren et al., 2018) assign weights to training samples based

on their gradient directions to decrease the effect of samples with noisy labels. (Reed et al., 2014; Laine and Aila, 2016; Li et al., 2017) aim to form consensus label predictions under different epochs to boost the model robustness. Compared to previous methods only forming consensus label predictions, the proposed GTE aims to form both consensus label and feature predictions to further improve the consistency and smoothness of predictions.

3. Methods

In this section, we first provide a brief introduction of the TE method (Laine and Aila, 2016) and then present the proposed method, GTE.

3.1. Temporal ensembling based convolutional neural network

Given a set of labeled and unlabeled images $\mathbf{X} = [\mathbf{X}_u; \mathbf{X}_l] = \{\mathbf{x}_i\}_{i=1}^{N}$, $\mathbf{y} = \{y_i\}_{i=1}^{n}$ ($y_i \in \{0, 1, \cdots, K-1\}$) denotes the labels of labeled images, where $\mathbf{X}_u$ and $\mathbf{X}_l$ represent the labeled and unlabeled images, respectively, $\mathbf{x}_i$ denotes the i-th training image and K is the number of classes. Let $\mathbf{z}_i^c \in \mathbb{R}^K$ be the label prediction (which is a predicted class probability vector) of $\mathbf{x}_i$ and $\tilde{\mathbf{z}}_i^c \in \mathbb{R}^K$ represent the ensemble target for label prediction $\mathbf{z}_i^c$. $\tilde{\mathbf{z}}_i^c$ is obtained by applying EMA to label predictions within multiple previous training epochs. Specifically, in each training epoch, $\mathbf{z}_i^c$ is firstly accumulated into an ensemble vector $\tilde{\mathbf{z}}_i^{ce} \in \mathbb{R}^K$, i.e. $\mathbf{z}_i^{ce} = \alpha \mathbf{z}_i^{ce} + (1-\alpha)\mathbf{z}_i^c$, and then $\tilde{\mathbf{z}}_i^c$ is computed by $\tilde{\mathbf{z}}_i^c = \mathbf{z}_i^{ce}/(1-\alpha^t)$, where α is a momentum term to control how far the ensemble $\mathbf{z}_i^{ce}$ reaches into the training history, and t is the current number of training epochs.

Let L represent the index set of labeled images in $\mathbf{X}$ and B denote the index set of selected images from $\mathbf{X}$. The loss function of TE is (Laine and

Aila, 2016):

$$J_{TE} = -\frac{1}{|B|} \sum_{i \in (B \cap L)} log\ \mathbf{z}_i^c[y_i] + \frac{\tau(t)}{K\,|B|} \sum_{i \in B} \|\mathbf{z}_i^c - \tilde{\mathbf{z}}_i^c\|_F^2\,, \tag{1}$$

where the first term is a cross-entropy loss function for labeled images, $|B|$ indicates the number of selected images, and $\tau(t)$ indicates a time-dependent weighting function to gradually enhance the weight of the consistency cost, e.g. $\sum_{i \in B} \|\mathbf{z}_i^c - \tilde{\mathbf{z}}_i^c\|_F^2$. This term is to explore the semantic information hidden in unlabeled data and meanwhile smooth the model. Eq. (1) suggests that TE, which extrapolates the labels of unlabeled images by aggregating label predictions within multiple previous epochs, can effectively explore the semantic information in unlabeled data.

3.2. Graph temporal ensembling based convolutional neural network

TE has obtained promising classification performance on natural images and handwritten digits. However, it fails to take into account the connection among labeled images and the consistency of predicting features, which can further boost model smoothness. Considering these two factors, we construct a graph to connect feature and label predictions among labeled images, so that images of each class can be mapped into a cluster, which can create more robust and stronger ensemble feature and label predictions.

First, we define a graph to preserve the relationship among training samples. Suppose that B represents the index set of images in each batch, and a symmetric matrix $\mathbf{S} \in \mathbb{R}^{|B| \times |B|}$ maintains their relations. For any two samples $\mathbf{x}_i$ and $\mathbf{x}_j$, their relations are described by:

$$s_{ij} = \begin{cases} 1 & if\ \mathbf{x}_i\ and\ \mathbf{x}_j\ belong\ to\ the\ same\ class \\ 0 & otherwise, \end{cases} \tag{2}$$

where $s_{ij} \in \mathbf{S}$ $(i, j \in B)$. Note that for unlabeled data $\mathbf{x}_i$, only $s_{ii} = 1$ and $s_{ij} = 0$ for $i \neq j$.

Because each class consists of different numbers of samples in one batch, we normalize the matrix $\mathbf{S}$ as follows:

$$\mathbf{W} = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}, \tag{3}$$

where $\mathbf{D} \in \mathbb{R}^{|B|\times|B|}$ is a positive diagonal matrix with the i-th element being $d_{ii} = \sum_{j=1}^{|B|} s_{ij}$. Thus, the sum of each row of $\mathbf{W}$ is one.

Next, we present the loss function of GTE by using the matrix $\mathbf{W}$. In addition to the label prediction $\mathbf{z}_i^c$ of the image $\mathbf{x}_i$, we adopt its feature representation $\mathbf{z}_i^f$ to enhance the relationship among training samples. For samples in each batch, to map feature and label predictions of labeled images within the same class into a cluster, we propose the following loss:

$$J_1 = \frac{1}{K\,|B|}\sum_{i\in B}\left\|\mathbf{z}_i^f - \sum_{j=1}^{|B|} w_{ij}\mathbf{z}_j^f\right\|_F^2 + \left\|\mathbf{z}_i^c - \sum_{j=1}^{|B|} w_{ij}\mathbf{z}_j^c\right\|_F^2. \tag{4}$$

For any unlabeled image $\mathbf{x}_i$, we have $s_{ii} = 1$ so that $\mathbf{w}_{ii} = 1$ and $\mathbf{w}_{ij} = 0$ for $i \neq j$, Eq. (4) will remove the effect of unlabeled samples and only map the labeled images of each class into a cluster.

Let $\tilde{\mathbf{z}}_i^f$ denote the ensemble target of the feature representation $\mathbf{z}_i^f$. Similar to $\tilde{\mathbf{z}}_i^c$, $\tilde{\mathbf{z}}_i^f$ is obtained by applying EMA to feature representations within multiple previous training epochs, i.e. $\mathbf{z}_i^{fe} = \alpha\mathbf{z}_i^{fe} + (1-\alpha)\mathbf{z}_i^f$, and $\tilde{\mathbf{z}}_i^f = \mathbf{z}_i^{fe}/(1-\alpha^t)$, where $\mathbf{z}_i^{fe}$ is an ensemble feature vector over multiple previous feature representations. $\tilde{\mathbf{z}}_i^f$ and $\tilde{\mathbf{z}}_i^c$ usually have better prediction quality than $\mathbf{z}_i^f$ and $\mathbf{z}_i^c$, respectively. Additionally, Eq. (4) does not take advantage of unlabeled data. Therefore, to map labeled images of each class into a

cluster and explore the semantic information in unlabeled images, we utilize $\sum_{j=1}^{|B|} w_{ij}\tilde{\mathbf{z}}_j^f$ and $\sum_{j=1}^{|B|} w_{ij}\tilde{\mathbf{z}}_j^c$ to replace $\sum_{j=1}^{|B|} w_{ij}\mathbf{z}_i^f$ and $\sum_{j=1}^{|B|} w_{ij}\mathbf{z}_i^c$ in Eq. (4). It becomes:

$$J_2 = \frac{1}{K|B|}\sum_{i\in B}\left\|\mathbf{z}_i^f - \sum_{j=1}^{|B|} w_{ij}\tilde{\mathbf{z}}_j^f\right\|_F^2 + \left\|\mathbf{z}_i^c - \sum_{j=1}^{|B|} w_{ij}\tilde{\mathbf{z}}_j^c\right\|_F^2. \tag{5}$$

It is worth noting when $\mathbf{x}_i$ is one unlabeled sample, Eq. (5) explores its information via $\left\|\mathbf{z}_i^f - \tilde{\mathbf{z}}_i^f\right\|_F^2 + \|\mathbf{z}_i^c - \tilde{\mathbf{z}}_i^c\|_F^2$, because of $w_{ii} = 1$ and $w_{ij} = 0$ for $i \neq j$.

Each image should be more similar to itself than the others under different configurations, and thus we redefine the matrix $\mathbf{S}$ by using $\mathbf{S} \leftarrow \mathbf{S} + \gamma\mathbf{I}_{n^b}$ to enhance the consistency of predictions, where $\mathbf{I}_{n^b} \in \mathbb{R}^{n^b\times n^b}$ is an identity matrix and γ is a non-negative constant to adjust the weight between cluster mapping and the consistency. The diagonal matrix $\mathbf{D}$ is calculated by $d_{ii} = \sum_{j=1}^{n^b} s_{ij}$ and then $\mathbf{W}$ is computed based on Eq. (3).

By introducing the time-dependent weighting function $\tau(t)$ to Eq. (5) and then combining it with the cross-entropy function on labeled images, we obtain the loss function of GTE as follows:

$$\begin{aligned} J = &-\frac{1}{|B|}\sum_{i\in(B\cap L)} log\ \mathbf{z}_i^c[y_i] \\ &+\frac{\tau(t)}{K|B|}\sum_{i,j\in(B)}(\lambda_1\left\|\mathbf{z}_i^f - \sum_{j=1}^{|B|} w_{ij}\tilde{\mathbf{z}}_j^f\right\|_F^2 + \lambda_2\left\|\mathbf{z}_i^c - \sum_{j=1}^{|B|} w_{ij}\tilde{\mathbf{z}}_j^c\right\|_F^2), \end{aligned} \tag{6}$$

where λ_1 and λ_2 are non-negative constants to balance the three terms. Let $f_\theta(\cdot)$ represent a stochastic neural network with parameters θ. Based on the loss function Eq. (6), the parameters θ can be updated by any optimizer, e.g. Adam (Kingma and Ba, 2014). For clarity, we present the detailed procedure using Eq. (6) to learn model parameters in Algorithm 1.

Algorithm 1: GTE

Input: Training images $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, label index set L, labels $\mathbf{y} = \{y_i\}_{i=1}^{n}$, weighting function $\tau(t)$, parameters $\gamma, \lambda_1, \lambda_2$, ensembling momentum α, stochastic neural network with parameters θ: $f_\theta(x)$, stochastic input augmentation function: $g(x)$

Output: Parameters θ

1. **Initialization:**
   $\mathbf{Z}^{fe} \leftarrow \mathbf{0}_{N\times K}$, ▷ ensemble previous feature vectors
   $\tilde{\mathbf{Z}}^{f} \leftarrow \mathbf{0}_{N\times K}$, ▷ target feature vectors
   $\mathbf{Z}^{ce} \leftarrow \mathbf{0}_{N\times K}$, ▷ ensemble previous label predictions
   $\tilde{\mathbf{Z}}^{c} \leftarrow \mathbf{0}_{N\times K}$, ▷ target label predictions
2. **for** t **in** $[1, num\_epochs]$ **do**
3. **for** each minibatch B **do**
4. Constructing $\mathbf{S}$ via Eq. (2) and then calculating $\mathbf{S} \leftarrow \mathbf{S} + \gamma\mathbf{I}_{n^b}$;
5. Calculating $\mathbf{W}$ via Eq. (3);
6. $\mathbf{z}^{c}_{i\in B}, \mathbf{z}^{f}_{i\in B} \leftarrow f_\theta(g(\mathbf{x}_{i\in B}))$;
7. loss $\leftarrow$ Eq. (6);
8. updating θ using optimizers, e.g. Adam (Kingma and Ba, 2014);
9. **end for**
10. $\mathbf{Z}^{fe} \leftarrow \alpha\mathbf{Z}^{fe} + (1-\alpha)\mathbf{Z}^{f}$
11. $\tilde{\mathbf{Z}}^{f} \leftarrow \mathbf{Z}^{fe}/(1-\alpha^t)$
12. $\mathbf{Z}^{ce} \leftarrow \alpha\mathbf{Z}^{ce} + (1-\alpha)\mathbf{Z}^{c}$
13. $\tilde{\mathbf{Z}}^{c} \leftarrow \mathbf{Z}^{c}/(1-\alpha^t)$
14. **end for**

3.3. Implementation details

We implement GTE with the PyTorch framework and adopt AlexNet (Krizhevsky et al., 2012) as our backbone network, which is pre-trained on the ImageNet database (Deng et al., 2009). The maximum learning rate η is 0.00005 and the Adam momentum parameters are $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For GTE, we empirically set the parameters $\alpha = 0.6$, $\lambda_1 = \lambda_2 = 0.1$ and $\gamma = 1$. We totally run 100 epochs for training with mini-batch size of 40. The time-dependent weighting function $\tau(t)$ ramps up from 0 to 1 during the first 40 epochs. $\tau(t)$ is a Gaussian ramp-up curve $e^{-5\|1-T\|^2}$, where T advances linearly from 0 to 1 during the first 40 epochs. The learning rate η also gradually increases to the maximum using $\tau(t)$ during the first 40 epochs and then keeps unchanged in the following 30 epochs, but decreases to 0 by using a Gaussian ramp-down curve during the last 30 epochs. The ramp-down curve is similar to the ramp-up curve while using a scaling constant 12.5, and linearly decreasing T from 1 to 0 during the last 30 epochs. Additionally, the Adam momentum parameter β_1 ramps down to 0.5 during the last 30 epochs.

4. Experimental Results and Analysis

To evaluate the proposed method, GTE, we conduct experiments on two histopathology image datasets consisting of lung cancer and breast cancer images, respectively. All images are selected from The Cancer Genome Altas (TCGA) and each of them is stained with Hematoxylin and eosin (H&E). Both lung and breast cancer datasets contain images with both 10x and 20x magnifications.

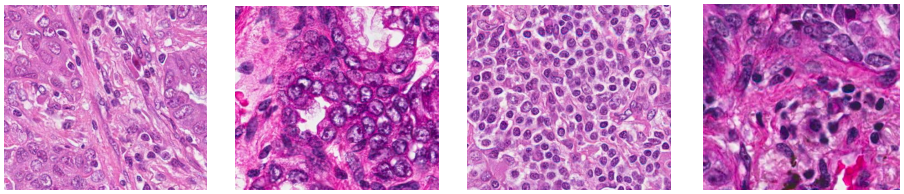
The lung cancer dataset contains two types of diseases: adenocarcinoma (AC) and squamous cell carcinoma (SC). We crop and select 2,904 (1,456 AC and 1,448 SC) lung cancer image patches from whole slide images of 42 patients. The size of each cropped image patch is 500×500. Each patient is composed of a set of image patches, and the number of patches varies from 24 to 196. We randomly select 30 patients containing 2104 image patches (1,008 AC and 1,096 SC) to construct a training set, and the remaining 12 patients with 800 image patches (448 AC and 352 SC) are used for testing.

The breast cancer dataset consists of 1,763 image patches selected from whole slide images of 193 patients in TCGA. The nuclear pleomorphism score of each whole slide image is annotated by pathologists. Then we select and crop image patches corresponding to the nuclear pleomorphism score of whole slide images. Each image patch is cropped with the size of 1000×1000. When one slide includes examples of several grades, we only select the patches corresponding to the highest grade. The number of image patches for each patient ranges from 1 to 15. These patches are classified into three groups by grading the nuclear pleomorphism based on the nucleus, shape and size of cells. The grading nuclear pleomorphism score is judged by the following rules (Galea et al., 1992):

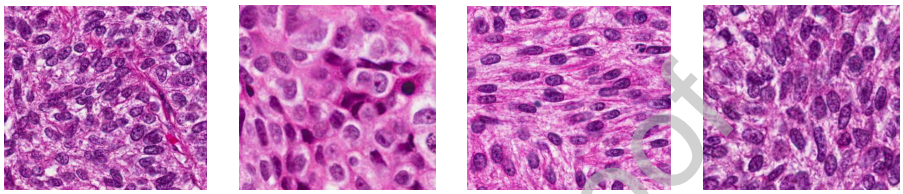
Score 1: Cells are uniform in size and shape compared to healthy cells (breast epithelial cells);

Score 2: Cells might be more extensive and slightly variational in size and shape compared to healthy cells;

Score 3: Cells have remarkable variation in size and shape compared to

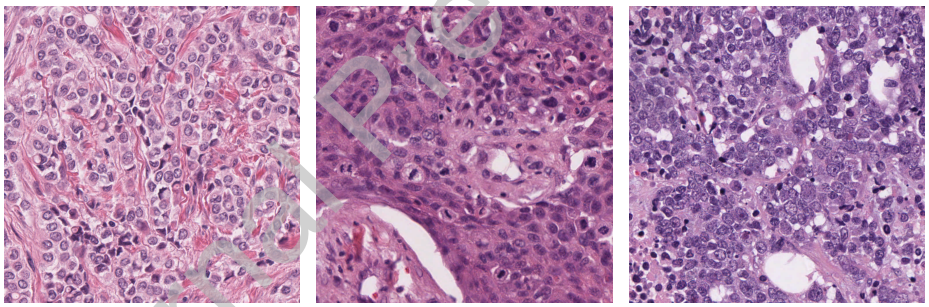


(a) AC



(b) SC

Fig 2: Examples of lung cancer images.



(a) Score 1 (b) Score 2 (c) Score 3

Fig 3: Examples of breast cancer images with different nuclear pleomorphism scores.

healthy cells;

Fig 2 and Fig. 3 show several example lung cancer images with two types of diseases and breast cancer images with the nuclear pleomorphism graded by scores 1-3, respectively. Because the images with nuclear pleomorphism scores 2 and 3 are very difficult to distinguish even for pathologists, we combine the images with nuclear pleomorphism scores 2 and 3 into one class. As

a result, the selected breast cancer images contain two classes. We name nuclear pleomorphism score 1 as class 1 and label nuclear pleomorphism scores 2-3 as class 2. We randomly select 1,421 images (476 class 1 and 945 class 2) of 152 patients for training, and the remaining 342 images (88 class 1 and 254 class 2) of 41 patients are utilized for testing. All labels and nuclear pleomorphism scores of image patches are confirmed by pathologists.

10x and 20x image patches are fed into a single network together, in order to enhance the model robustness to the scale of cellular information. All of them are resized to 224×224 as the given input and training images are augmented with random translations ($\{\triangle x, \triangle y\} \sim [-32, 32]$), horizontal flips ($p = 0.5$), Gaussian noise ($\sigma = 0.15$) and color normalization.

4.1. Experiments for semi-supervision

We compare GTE against six popular methods, including one supervised method: AlexNet with the cross-entropy loss function (Baseline), and five semi-supervised methods: Π-model and TE (Laine and Aila, 2016), MT (Tarvainen and Valpola, 2017), VAT (Miyato et al., 2018) and SNTG (Luo et al., 2017). Here, SNTG utilizes TE as the base. For fairness, all the five semi-supervised methods adopt the pre-trained AlexNet as the backbone network. Additionally, we set their maximum learning rate to be 0.00005, upon which they can obtain the best or suboptimal performance on the pre-trained network. We randomly select 20%, 35% and 50% patients of each class from the lung or breast cancer training set to construct a labeled set, and the remaining patients are used as unlabeled ones. We repeat this process ten times and report the average image classification accuracy, recall, F-score of various methods. Additionally, we also present their performance when all

Table 1: Results for semi-supervision on the lung cancer dataset. We bold the best results and highlight the second best results via underlines.

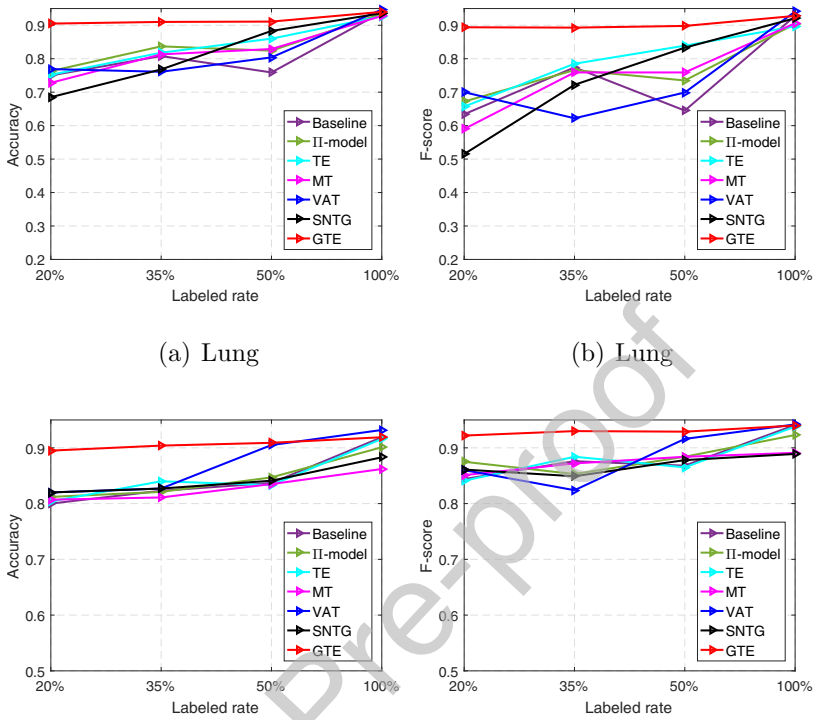
Method	Accuracy			Recall	F-Score
	AC	SC	Avg		
20% labeled patients					
Baseline	0.875 ± 0.099	0.591 ± 0.300	0.750 ± 0.106	0.591 ± 0.230	0.634 ± 0.248
Π-model	0.832 ± 0.201	0.676 ± 0.166	0.763 ± 0.130	0.636 ± 0.188	0.672 ± 0.150
TE	0.855 ± 0.121	0.620 ± 0.284	0.752 ± 0.119	0.620 ± 0.284	0.657 ± 0.234
MT	0.855 ± 0.099	0.567 ± 0.276	0.728 ± 0.091	0.543 ± 0.248	0.590 ± 0.211
VAT	0.827 ± 0.224	0.696 ± 0.259	$\underline{0.769} \pm 0.115$	$\underline{0.692} \pm 0.259$	$\underline{0.700} \pm 0.197$
SNTG	0.838 ± 0.135	0.491 ± 0.268	0.685 ± 0.089	0.460 ± 0.242	0.515 ± 0.217
GTE	0.921 ± 0.046	0.886 ± 0.101	$\mathbf{0.905} \pm 0.041$	$\mathbf{0.886} \pm 0.101$	$\mathbf{0.894} \pm 0.071$
35% labeled patients					
Baseline	0.842 ± 0.149	0.767 ± 0.230	0.808 ± 0.147	0.761 ± 0.231	$\underline{0.774} \pm 0.204$
Π-model	0.893 ± 0.061	0.765 ± 0.124	$\underline{0.837} \pm 0.060$	0.730 ± 0.155	0.766 ± 0.119
TE	0.855 ± 0.114	0.772 ± 0.197	0.818 ± 0.081	$\underline{0.772} \pm 0.197$	0.785 ± 0.147
MT	0.862 ± 0.098	0.750 ± 0.196	0.813 ± 0.100	0.738 ± 0.193	0.759 ± 0.158
VAT	0.902 ± 0.065	0.581 ± 0.310	0.761 ± 0.122	0.571 ± 0.311	0.622 ± 0.258
SNTG	0.782 ± 0.210	0.751 ± 0.210	0.768 ± 0.152	0.716 ± 0.207	0.722 ± 0.181
GTE	0.936 ± 0.026	0.878 ± 0.059	$\mathbf{0.910} \pm 0.026$	$\mathbf{0.878} \pm 0.059$	$\mathbf{0.893} \pm 0.042$
50% labeled patients					
Baseline	0.897 ± 0.087	0.584 ± 0.225	0.759 ± 0.090	0.584 ± 0.225	0.645 ± 0.188
Π-model	0.925 ± 0.044	0.696 ± 0.239	0.824 ± 0.104	0.688 ± 0.232	0.735 ± 0.191
TE	0.891 ± 0.060	0.821 ± 0.067	0.860 ± 0.045	$\underline{0.821} \pm 0.067$	$\underline{0.839} \pm 0.052$
MT	0.905 ± 0.060	0.733 ± 0.267	0.829 ± 0.129	0.733 ± 0.264	0.759 ± 0.241
VAT	0.921 ± 0.076	0.654 ± 0.266	0.804 ± 0.125	0.644 ± 0.262	0.699 ± 0.222
SNTG	0.933 ± 0.032	0.820 ± 0.169	$\underline{0.883} \pm 0.081$	0.802 ± 0.183	0.833 ± 0.145
GTE	0.933 ± 0.034	0.884 ± 0.063	$\mathbf{0.911} \pm 0.019$	$\mathbf{0.884} \pm 0.063$	$\mathbf{0.898} \pm 0.039$
All labeled patients					
Baseline	0.953 ± 0.028	0.922 ± 0.031	0.939 ± 0.011	0.917 ± 0.025	$\underline{0.928} \pm 0.017$
Π-model	0.965 ± 0.013	0.897 ± 0.036	0.935 ± 0.010	0.884 ± 0.048	0.904 ± 0.032
TE	0.951 ± 0.024	0.897 ± 0.055	0.927 ± 0.024	0.871 ± 0.055	0.897 ± 0.039
MT	0.935 ± 0.030	0.921 ± 0.019	0.928 ± 0.009	0.896 ± 0.027	0.906 ± 0.017
VAT	0.950 ± 0.023	0.942 ± 0.026	$\mathbf{0.947} \pm 0.006$	$\mathbf{0.939} \pm 0.026$	$\mathbf{0.941} \pm 0.014$
SNTG	0.951 ± 0.019	0.916 ± 0.020	0.936 ± 0.006	0.910 ± 0.019	0.921 ± 0.011
GTE	0.954 ± 0.013	0.923 ± 0.035	$\underline{0.940} \pm 0.012$	$\underline{0.919} \pm 0.035$	$\underline{0.928} \pm 0.025$

Table 2: Results for semi-supervision on the breast cancer dataset. We bold the best results and highlight the second best results via underlines.

Method	Accuracy			Recall	F-Score
	Class 1	Class 2	Avg		
20% labeled patients					
Baseline	0.523 ± 0.289	0.896 ± 0.051	0.800 ± 0.058	0.896 ± 0.051	0.844 ± 0.037
Π-model	0.375 ± 0.367	0.963 ± 0.040	0.812 ± 0.063	$\mathbf{0.958} \pm 0.036$	$\underline{0.875} \pm 0.043$
TE	0.572 ± 0.246	0.884 ± 0.092	0.804 ± 0.057	0.884 ± 0.092	0.841 ± 0.064
MT	0.401 ± 0.273	0.938 ± 0.060	0.807 ± 0.068	0.938 ± 0.111	0.851 ± 0.100
VAT	0.480 ± 0.196	0.937 ± 0.068	$\underline{0.820} \pm 0.064$	0.922 ± 0.076	0.860 ± 0.066
SNTG	0.488 ± 0.285	0.935 ± 0.065	$\underline{0.820} \pm 0.060$	0.922 ± 0.068	0.861 ± 0.059
GTE	0.738 ± 0.111	0.949 ± 0.016	$\mathbf{0.895} \pm 0.023$	$\underline{0.951} \pm 0.016$	$\mathbf{0.922} \pm 0.014$
35% labeled patients					
Baseline	0.489 ± 0.235	0.939 ± 0.049	0.823 ± 0.040	0.939 ± 0.047	0.876 ± 0.027
Π-model	0.600 ± 0.192	0.898 ± 0.053	0.821 ± 0.054	0.899 ± 0.053	0.853 ± 0.044
TE	0.564 ± 0.199	0.935 ± 0.045	$\underline{0.840} \pm 0.035$	0.935 ± 0.049	$\underline{0.884} \pm 0.029$
MT	0.393 ± 0.229	0.956 ± 0.033	0.811 ± 0.056	$\mathbf{0.970} \pm 0.027$	0.872 ± 0.023
VAT	0.696 ± 0.258	0.873 ± 0.077	0.827 ± 0.093	0.845 ± 0.098	0.824 ± 0.096
SNTG	0.581 ± 0.279	0.913 ± 0.076	0.827 ± 0.066	0.893 ± 0.075	0.849 ± 0.059
GTE	0.751 ± 0.119	0.957 ± 0.025	$\mathbf{0.904} \pm 0.026$	$\underline{0.959} \pm 0.025$	$\mathbf{0.930} \pm 0.019$
50% labeled patients					
Baseline	0.642 ± 0.183	0.905 ± 0.085	0.837 ± 0.060	0.905 ± 0.085	0.868 ± 0.008
Π-model	0.599 ± 0.254	0.933 ± 0.070	0.847 ± 0.060	0.932 ± 0.069	0.884 ± 0.055
TE	0.638 ± 0.186	0.901 ± 0.055	0.833 ± 0.050	0.901 ± 0.055	0.865 ± 0.044
MT	0.581 ± 0.224	0.955 ± 0.042	0.835 ± 0.040	$\mathbf{0.958} \pm 0.040$	0.884 ± 0.021
VAT	0.758 ± 0.143	0.956 ± 0.053	$\underline{0.905} \pm 0.045$	0.945 ± 0.055	$\underline{0.916} \pm 0.043$
SNTG	0.514 ± 0.273	0.954 ± 0.024	0.841 ± 0.073	0.937 ± 0.049	0.878 ± 0.067
GTE	0.791 ± 0.053	0.950 ± 0.024	$\mathbf{0.909} \pm 0.023$	$\underline{0.950} \pm 0.024$	$\mathbf{0.929} \pm 0.023$
All labeled patients					
Baseline	0.825 ± 0.026	0.965 ± 0.012	$\underline{0.919} \pm 0.007$	$\mathbf{0.965} \pm 0.012$	$\underline{0.941} \pm 0.008$
Π-model	0.768 ± 0.083	0.947 ± 0.029	0.901 ± 0.019	0.954 ± 0.023	0.923 ± 0.014
TE	0.779 ± 0.044	0.963 ± 0.016	0.916 ± 0.011	0.960 ± 0.015	0.938 ± 0.011
MT	0.719 ± 0.080	0.911 ± 0.033	0.862 ± 0.021	0.931 ± 0.032	0.891 ± 0.023
VAT	0.848 ± 0.013	0.961 ± 0.015	$\mathbf{0.932} \pm 0.010$	$\underline{0.963} \pm 0.018$	$\mathbf{0.942} \pm 0.015$
SNTG	0.753 ± 0.059	0.928 ± 0.019	0.883 ± 0.017	0.914 ± 0.024	0.889 ± 0.020
GTE	0.798 ± 0.039	0.961 ± 0.018	$\underline{0.919} \pm 0.008$	$\underline{0.963} \pm 0.016$	0.940 ± 0.011

training images are labeled.

Table 1 presents the image classification accuracy, recall and F-score of various methods on the lung cancer image dataset. GTE can achieve superior performance over the others when 20%, 35% and 50% patients are labeled. For example, the gain of GTE in terms of the average accuracy ranges from 3.2% to 17.7% over the best competitors on 20%, 35% and 50% labeled patients. Additionally, GTE obtains the best recall and F-score among all methods on the three cases. When all patients are labeled, GTE attains inferior performance to VAT, while it achieves the similar performance to Baseline and outperforms the other four methods. Table 2 shows the image classification results on the breast cancer image dataset. GTE can also obtain superior performance over the other six methods when 20%, 35% and 50% patients are labeled. For instance, the gain of GTE in accuracy is from 0.4% to 9.2% over the best competitors on 20%, 35% and 50% labeled patients. When using all labeled patients, GTE exhibits similar performance to Baseline and outperforms the other methods except VAT. As shown in Tables 1 and 2, some methods with less labeled patients obtain better performance than that with more labeled patients, like VAT using 20% and 35% labeled patients in Table 1, and TE using 35% and 50% labeled patients in Table 2. *This might be because each patient in training and test sets contains different number of images, leading to distinct significance of patients for model training.* To more clearly compare their classification performance, we show the average accuracy and F-score of seven methods on lung and breast cancer databases in Fig 4.



(a) Lung (b) Lung

(c) Breast (d) Breast

Fig 4: The average accuracy and F-score of seven methods with different rates of labeled patients for semi-supervised learning on lung and breast cancer databases.

4.2. Experiments for noisy labels

We compare GTE against six popular methods, including one baseline, AlexNet with the cross-entropy loss function, and five state-of-the-art approaches, such as Bootstrap (Reed et al., 2014), Forward and Backward (Patrini et al., 2017), TE (Laine and Aila, 2016) and SNTG (Luo et al., 2017). These approaches also utilize the pre-trained AlexNet as the backbone network and their maximum learning rate is 0.00005. We generate symmetric label noise by randomly flipping 10% and 20% training patients of each class

into the other class. We repeat this process ten times and present their average image classification accuracy, recall and F-score.

Tables 3 illustrates that GTE can obtain competitive or even better average image classification accuracy and F-score than the others on lung and breast cancer datasets. For example, GTE only achieves 0.2% higher average accuracy than Backward when using 10% patients with noisy labels on the lung caner dataset, but it obtains 4.2% higher accuracy than Backward when 20% patients are with noisy labels. Therefore, GTE is more robust to noisy labels than the other methods.

4.3. Experiments for semi-supervision and noisy labels

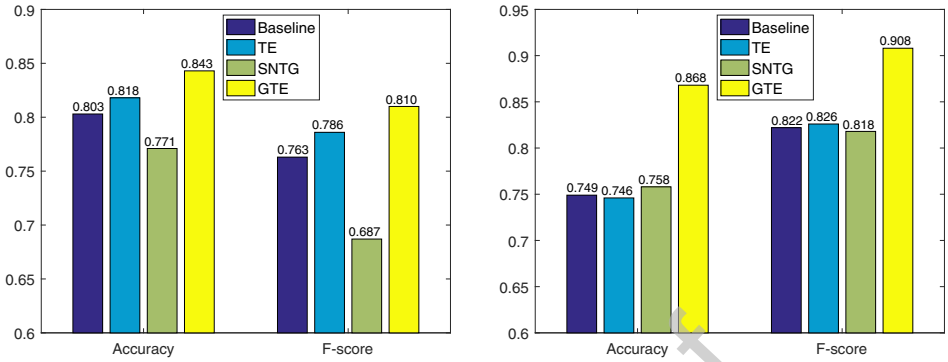
To better demonstrate the strength of GTE, we conduct experiments on the training set with noisy labeled images and unlabeled ones. Specifically, we uniformly select 50% patients from the lung or breast cancer training set to construct a labeled set and utilize the remaining patients as unlabeled ones. Then we randomly flip 20% labeled patients of each class into the other class. We run the experiment ten times and present the average image classification accuracy, recall and F-score of four methods: Baseline, TE, SNTG and GTE.

Fig 5 shows their results on lung and breast cancer datasets. It demonstrates that GTE outperforms the other three methods when using noisy labeled and unlabeled data for lung and breast cancer image classification. Specifically, the gain of GTE in terms of average accuracy is 3.1% and 14.5% over the best competitors on lung and breast cancer datasets, respectively. GTE also achieves better F-score than the other three methods.

We further utilize a popular dimensionality reduction method, t-SNE

Table 3: Results for noisy labels on the lung cancer dataset. We bold the best results and highlight the second best results via underlines.

Method	Accuracy			Recall	F-Score
	AC	SC	Avg		
Lung cancer					
10% patients with noisy labels					
Baseline	0.905 ± 0.033	0.908 ± 0.035	0.906 ± 0.029	$\mathbf{0.898} \pm 0.035$	0.896 ± 0.032
Bootstrap	0.916 ± 0.076	0.857 ± 0.135	0.890 ± 0.101	0.857 ± 0.135	0.871 ± 0.124
Forward	0.902 ± 0.071	0.834 ± 0.189	0.872 ± 0.121	0.832 ± 0.188	0.844 ± 0.160
Backward	0.938 ± 0.021	0.889 ± 0.049	$\underline{0.916} \pm 0.019$	0.884 ± 0.049	0.899 ± 0.024
TE	0.928 ± 0.030	0.887 ± 0.033	0.910 ± 0.020	$\underline{0.887} \pm 0.033$	$\underline{0.899} \pm 0.022$
SNTG	0.941 ± 0.038	0.828 ± 0.161	0.891 ± 0.084	0.824 ± 0.154	0.860 ± 0.124
GTE	0.946 ± 0.025	0.883 ± 0.035	$\mathbf{0.918} \pm 0.022$	0.883 ± 0.035	$\mathbf{0.903} \pm 0.036$
20% patients with noisy labels					
Baseline	0.860 ± 0.071	0.778 ± 0.207	0.824 ± 0.100	0.778 ± 0.207	0.783 ± 0.148
Bootstrap	0.869 ± 0.090	0.756 ± 0.175	0.819 ± 0.074	0.756 ± 0.175	0.777 ± 0.122
Forward	0.884 ± 0.074	0.845 ± 0.076	$\underline{0.867} \pm 0.065$	0.845 ± 0.076	$\underline{0.848} \pm 0.072$
Backward	0.822 ± 0.132	0.847 ± 0.107	0.833 ± 0.105	0.847 ± 0.107	0.819 ± 0.108
TE	0.864 ± 0.055	0.861 ± 0.089	0.863 ± 0.064	$\mathbf{0.861} \pm 0.089$	0.847 ± 0.073
SNTG	0.891 ± 0.068	0.697 ± 0.201	0.806 ± 0.125	0.695 ± 0.184	0.735 ± 0.172
GTE	0.895 ± 0.059	0.850 ± 0.097	$\mathbf{0.875} \pm 0.060$	$\underline{0.850} \pm 0.097$	$\mathbf{0.855} \pm 0.075$
Breast cancer					
10% patients with noisy labels					
Baseline	0.749 ± 0.098	0.886 ± 0.046	0.851 ± 0.046	0.886 ± 0.046	0.892 ± 0.032
Bootstrap	0.708 ± 0.096	0.919 ± 0.046	$\underline{0.865} \pm 0.035$	$\mathbf{0.919} \pm 0.046$	$\underline{0.910} \pm 0.025$
Forward	0.753 ± 0.063	0.876 ± 0.066	0.844 ± 0.043	0.876 ± 0.066	0.892 ± 0.034
Backward	0.780 ± 0.082	0.878 ± 0.065	0.853 ± 0.054	0.878 ± 0.065	0.898 ± 0.041
TE	0.709 ± 0.252	0.897 ± 0.071	0.849 ± 0.067	0.897 ± 0.071	0.898 ± 0.044
SNTG	0.765 ± 0.095	0.700 ± 0.830	0.854 ± 0.053	0.870 ± 0.041	0.886 ± 0.033
GTE	0.748 ± 0.119	0.911 ± 0.063	$\mathbf{0.869} \pm 0.058$	$\underline{0.911} \pm 0.063$	$\mathbf{0.913} \pm 0.041$
20% patients with noisy labels					
Baseline	0.641 ± 0.259	0.849 ± 0.102	0.796 ± 0.076	0.849 ± 0.102	0.854 ± 0.058
Bootstrap	0.677 ± 0.254	0.837 ± 0.071	0.796 ± 0.050	0.837 ± 0.071	0.859 ± 0.062
Forward	0.656 ± 0.130	0.865 ± 0.060	0.811 ± 0.060	0.863 ± 0.069	0.871 ± 0.043
Backward	0.659 ± 0.133	0.857 ± 0.067	0.806 ± 0.053	0.856 ± 0.069	0.867 ± 0.040
TE	0.644 ± 0.209	0.870 ± 0.064	$\underline{0.812} \pm 0.056$	$\mathbf{0.874} \pm 0.064$	$\underline{0.883} \pm 0.044$
SNTG	0.700 ± 0.147	0.830 ± 0.065	0.797 ± 0.054	0.826 ± 0.063	0.854 ± 0.036
GTE	0.703 ± 0.140	0.855 ± 0.096	$\mathbf{0.815} \pm 0.058$	$\underline{0.872} \pm 0.086$	$\mathbf{0.889} \pm 0.048$



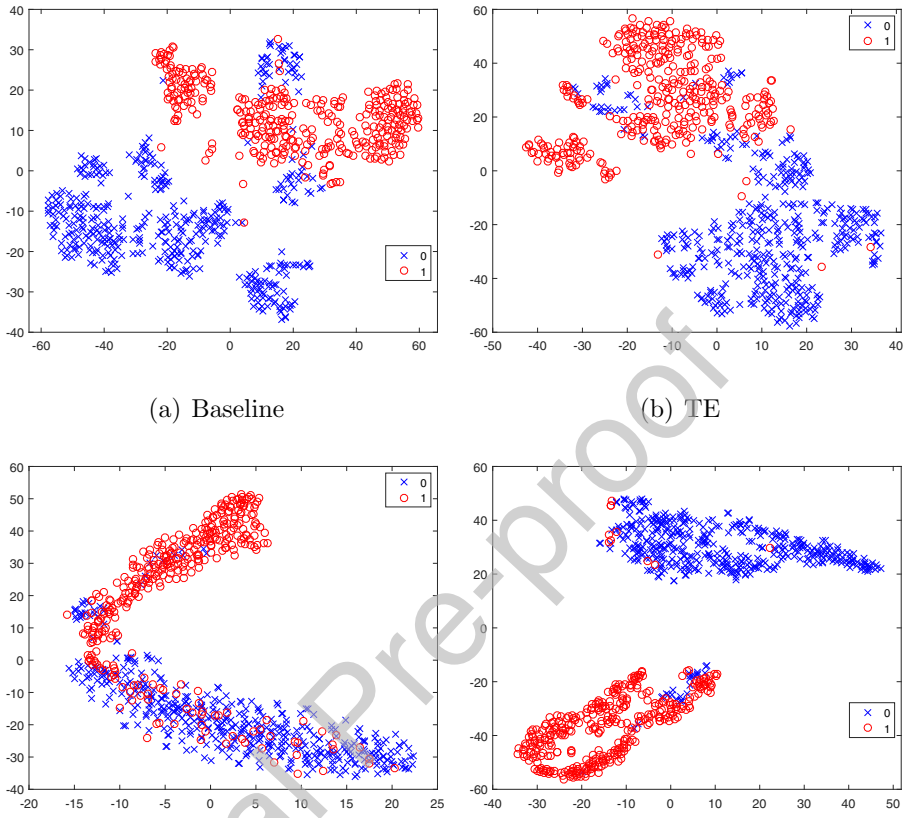
(a) Lung cancer (b) Breast cancer

Fig 5: Results including average accuracy and F-score for semi-supervision and noisy labels on the lung and breast cancer datasets. Each experiment is repeated 10 times and an average result is reported.

(Maaten and Hinton, 2008), to project the feature vectors of lung cancer test images obtained by different methods (Baseline, TE, SNTG and GTE) onto a two-dimensional plane. As shown in Fig. 6, GTE can make features between classes more separable, because GTE aims to map images of each class into a cluster, leading to more compact features in each class.

4.4. Effects of Parameter Selection

First, we show the influence of the feature and graph in GTE on semi-supervision and noisy labels. Specifically, we set $\lambda_1 = 0$ to remove the effect of features, and utilize $\mathbf{S} = \mathbf{I}_{n^b}$ to eliminate the influence of graph. We conduct semi-supervised experiments on two databases, and we randomly select 20% patients from the training set of each database to construct a labeled set, and the remaining images are used as unlabeled ones. For noisy label experiments, we also utilize the two training sets and randomly flip 20%



(a) Baseline (b) TE

(c) STNG (d) GTE

Fig 6: Feature projection of lung cancer test images onto a two-dimensional plane by using Baseline, TE, SNTG and GTE. 0 and 1 represent AC and SC, respectively.

patients of each class into the other class. We repeat this process ten times and report the average image accuracy, recall and F-score in Tables 4-5, which illustrate that GTE has superior performance over GTE($\lambda_1 = 0$) and GTE ($\mathbf{S} = \mathbf{I}_{n^b}$). Additionally, GTE($\lambda_1 = 0$) achieves better performance than GTE($\mathbf{S} = \mathbf{I}_{n^b}$) on semi-supervised experiments, but GTE($\mathbf{S} = \mathbf{I}_{n^b}$) performs better than GTE($\lambda_1 = 0$) on experiments for noisy labels. They suggest that graph (mapping samples of each class into a cluster) is more important than

Table 4: Results on the lung caner image dataset. We bold the best results and highlight the second best results via underlines.

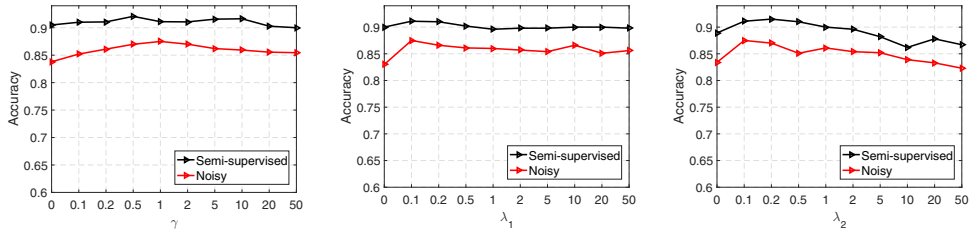
Method	Accuracy			Recall	F-Score
	Class 1	Class 2	Avg		
Semi-supervised					
GTE	0.921 ± 0.046	0.886 ± 0.101	$\mathbf{0.905} \pm 0.041$	$\mathbf{0.886} \pm 0.101$	$\mathbf{0.894} \pm 0.071$
GTE ($\lambda_1 = 0$)	0.935 ± 0.017	0.831 ± 0.150	$\underline{0.890} \pm 0.069$	$\underline{0.831} \pm 0.150$	$\underline{0.862} \pm 0.110$
GTE ($\mathbf{S} = \mathbf{I}_{n^b}$)	0.875 ± 0.086	0.556 ± 0.220	0.735 ± 0.091	0.556 ± 0.220	0.628 ± 0.177
Noisy labels					
GTE	0.895 ± 0.059	0.850 ± 0.097	$\mathbf{0.875} \pm 0.060$	$\mathbf{0.850} \pm 0.097$	$\mathbf{0.855} \pm 0.075$
GTE ($\lambda_1 = 0$)	0.884 ± 0.059	0.756 ± 0.203	0.828 ± 0.106	0.756 ± 0.023	0.783 ± 0.164
GTE ($\mathbf{S} = \mathbf{I}_{n^b}$)	0.936 ± 0.041	0.783 ± 0.147	$\underline{0.869} \pm 0.048$	$\underline{0.790} \pm 0.142$	$\underline{0.822} \pm 0.123$

Table 5: Results on the breast caner image dataset. We bold the best results and highlight the second best results via underlines.

Method	Accuracy			Recall	F-Score
	Class 1	Class 2	Avg		
Semi-supervised					
GTE	0.738 ± 0.111	0.949 ± 0.016	$\mathbf{0.895} \pm 0.023$	$\underline{0.951} \pm 0.016$	$\mathbf{0.922} \pm 0.014$
GTE ($\lambda_1 = 0$)	0.556 ± 0.223	0.967 ± 0.029	$\underline{0.861} \pm 0.040$	$\mathbf{0.967} \pm 0.029$	$\underline{0.913} \pm 0.021$
GTE ($\mathbf{S} = \mathbf{I}_{n^b}$)	0.485 ± 0.320	0.924 ± 0.045	0.811 ± 0.092	0.924 ± 0.045	0.881 ± 0.055
Noisy labels					
GTE	0.703 ± 0.140	0.855 ± 0.096	$\mathbf{0.815} \pm 0.058$	$\mathbf{0.872} \pm 0.086$	$\mathbf{0.889} \pm 0.048$
GTE ($\lambda_1 = 0$)	0.575 ± 0.255	0.813 ± 0.080	0.752 ± 0.054	0.813 ± 0.080	0.829 ± 0.038
GTE ($\mathbf{S} = \mathbf{I}_{n^b}$)	0.694 ± 0.147	0.851 ± 0.060	$\underline{0.810} \pm 0.054$	$\underline{0.863} \pm 0.060$	$\underline{0.877} \pm 0.038$

features (forming consensus feature predictions) on semi-supervision, while feature consistency is more significant on the robustness to noisy labels.

Then we present the effect of three essential parameters γ, λ_1 and λ_2 in GTE for semi-supervision and noisy labels with lung cancer images. For semi-supervised experiments, we randomly select 50% patients from the lung cancer training set to construct a labeled set, and utilize the remaining images as unlabeled ones; for noisy label experiments, we also utilize the



(a) $\lambda_1 = 0.1, \lambda_2 = 0.1$ (b) $\gamma = 1, \lambda_2 = 0.1$ (c) $\gamma = 1, \lambda_1 = 0.1$

Fig 7: The effect of different values of γ, λ_1 and λ_2 on the proposed GTE for semi-supervision and noisy labels with lung cancer images. Each experiment is repeated 10 times and an average result is reported.

lung cancer training set and randomly flip 20% patients of each class into the other class. We run the experiments ten times and report the average image accuracy of different values of γ, λ_1 and λ_2 during the range of $[0, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50]$ in Fig 7, upon which we can see that $\gamma = 1, \lambda_1 = \lambda_2 = 0.1$ can achieve the best or suboptimal accuracy for GTE on both semi-supervision and noisy labels. Similar findings can be observed on other percentages and breast cancer images.

5. Conclusion and Future Work

In this paper, we propose a novel robust deep architecture to leverage the semantic information of both labeled and unlabeled data. The proposed architecture utilizes EMA to create ensemble targets for feature and label predictions, and then adopts a graph to map labeled images of each class into a cluster to boost the strength of ensemble predictions. Meanwhile, it applies the consistency cost on feature representations and label predictions of all training images, in order to form consensus predictions under different

configurations. Experiments on lung and breast cancer datasets demonstrate the effectiveness and efficiency of the proposed method. Additionally, experiments illustrate that mapping labeled images of each class into a cluster is more beneficial to semi-supervised classification and forming consensus feature predictions is more helpful to the model robustness.

Although the proposed robust semi-supervised deep method has achieved promising performance on the two datasets, there is still much work to explore: (i) deep neural networks on semi-supervised learning and with noisy labels for multi-class histopathology image classification; (ii) training deep neural networks with extremely noisy labels; (iii) robust semi-supervised deep hashing for histopathology image retrieval (because image retrieval can not only provide the class information of images, but also search the most similar images); (iv) robust semi-supervised deep neural networks for whole-slide image classification.

References

Atwood, J., Towsley, D., 2016. Diffusion-convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems. pp. 1993–2001.

Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P. M., Rueckert, D., 2017. Semi-supervised learning for network-based cardiac mr image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 253–260.

Chang, H., Han, J., Zhong, C., Snijders, A. M., Mao, J.-H., 2017. Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedi-

cal applications. IEEE transactions on pattern analysis and machine intelligence 40 (5), 1182–1194.

Chen, P., Gao, L., Shi, X., Allen, K., Yang, L., 2019. Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. Computerized Medical Imaging and Graphics.

Cheplygina, V., de Bruijne, M., Pluim, J. P., 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255.

Galea, M. H., Blamey, R. W., Elston, C. E., Ellis, I. O., 1992. The nottingham prognostic index in primary breast cancer. Breast cancer research and treatment 22 (3), 207–219.

Ghosh, A., Kumar, H., Sastry, P., 2017. Robust loss functions under label noise for deep neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 1919–1925.

Gurcan, M. N., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., Yener, B., 2009. Histopathological image analysis: A review. IEEE Reviews in Biomedical Engineering 2, 147.

Haeusser, P., Mordvintsev, A., Cremers, D., 2017. Learning by association-a versatile semi-supervised training method for neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. Vol. 3. p. 6.

Kamnitsas, K., Castro, D. C., Folgoc, L. L., Walker, I., Tanno, R., Rueckert, D., Glocker, B., Criminisi, A., Nori, A., 2018. Semi-supervised learning via compact latent space clustering. arXiv preprint arXiv:1806.02679.

Kapil, A., Meier, A., Zuraw, A., Steele, K., Rebelatto, M., Schmidt, G., Brieu, N., 2018. Deep semi supervised generative learning for automated pd-l1 tumor cell scoring on nsclc tissue needle biopsies. arXiv preprint arXiv:1806.11036.

Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kipf, T. N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Kraus, O. Z., Ba, J. L., Frey, B. J., 2016. Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics 32 (12), i52–i59.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems. pp. 1097–1105.

Kumar, A., Sattigeri, P., Fletcher, T., 2017. Semi-supervised learning with gans: manifold invariance with improved inference. In: Proceedings of Advances in Neural Information Processing Systems. pp. 5534–5544.

Laine, S., Aila, T., 2016. Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242.

Lee, D.-H., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICML Workshop on Challenges in Representation Learning. Vol. 3. p. 2.

Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., Li, L.-J., 2017. Learning from noisy labels with distillation. In: Proceedings of International Conference on Computer Vision. pp. 1928–1936.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., Sánchez, C. I., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60–88.

Luo, Y., Guan, T., Yu, J., Liu, P., Yang, Y., 2018. Every node counts: Self-ensembling graph convolutional networks for semi-supervised learning. arXiv preprint arXiv:1809.09925.

Luo, Y., Zhu, J., Li, M., Ren, Y., Zhang, B., 2017. Smooth neighbors on teacher graphs for semi-supervised learning. arXiv preprint arXiv:1711.00258.

Maaten, L. v. d., Hinton, G., 2008. Visualizing data using t-sne. Journal of machine learning research 9 (Nov), 2579–2605.

Miyato, T., Maeda, S.-i., Ishii, S., Koyama, M., 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Mnih, V., Hinton, G. E., 2012. Learning to label aerial images from noisy data. In: Proceedings of International Conference on Machine Learning). pp. 567–574.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., Tewari, A., 2013. Learning with noisy labels. In: Proceedings of Advances in Neural Information Processing Systems. pp. 1196–1204.

Odena, A., 2016. Semi-supervised learning with generative adversarial networks. arXiv preprint arXiv:1606.01583.

Patrini, G., Rozza, A., Menon, A. K., Nock, R., Qu, L., 2017. Making deep neural networks robust to label noise: A loss correction approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2233–2241.

Peikari, M., Salama, S., Nofech-Mozes, S., Martel, A. L., 2018. A cluster-then-label semi-supervised learning approach for pathology image classification. Scientific Reports 8 (1), 7193.

Quellec, G., Cazuguel, G., Cochener, B., Lamard, M., 2017. Multiple-instance learning for medical image and video analysis. IEEE reviews in biomedical engineering 10, 213–234.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T., 2015. Semi-supervised learning with ladder networks. In: Proceedings of Advances in Neural Information Processing Systems. pp. 3546–3554.

Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A., 2014. Training deep neural networks on noisy labels with bootstrapping. arXiv preprint arXiv:1412.6596.

Ren, M., Zeng, W., Yang, B., Urtasun, R., 2018. Learning to reweight examples for robust deep learning. arXiv preprint arXiv:1803.09050.

Ruder, S., Plank, B., 2018. Strong baselines for neural semi-supervised learning under domain shift. arXiv preprint arXiv:1804.09530.

Sapkota, M., Shi, X., Xing, F., Yang, L., 2018. Deep convolutional hashing for low-dimensional binary embedding of histopathological images. IEEE journal of biomedical and health informatics 23 (2), 805–816.

584 Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis.
585 Annual Review of Biomedical Engineering 19, 221–248.

586 Shi, X., Sapkota, M., Xing, F., Liu, F., Cui, L., Yang, L., 2018. Pairwise based deep
587 ranking hashing for histopathology image classification and retrieval. Pattern
588 Recognition 81, 14–22.

589 Shi, X., Xing, F., Xu, K., Xie, Y., Su, H., Yang, L., 2017. Supervised graph hashing
590 for histopathology image retrieval and classification. Medical image analysis 42,
591 117–128.

592 Sparks, R., Madabhushi, A., 2016. Out-of-sample extrapolation utilizing semi-
593 supervised manifold learning (ose-ssl): Content based image retrieval for
594 histopathology images. Scientific Reports 6, 27306.

595 Su, H., Shi, X., Cai, J., Yang, L., 2019. Local and global consistency regular-
596 ized mean teacher for semi-supervised nuclei classification. In: International
597 Conference on Medical Image Computing and Computer-Assisted Intervention.
598 Springer, pp. 559–567.

599 Sukhbaatar, S., Fergus, R., 2014. Learning from noisy labels with deep neural
600 networks. arXiv preprint arXiv:1406.2080 2 (3), 4.

601 Tang, X., Guo, F., Shen, J., Du, T., 2018. Facial landmark detection by semi-
602 supervised deep learning. Neurocomputing 297, 22–32.

603 Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-
604 averaged consistency targets improve semi-supervised deep learning results. In:
605 Proceedings of Advances in Neural Information Processing Systems. pp. 1195–
606 1204.

607 Vahdat, A., 2017. Toward robustness against label noise in training deep discrim-
608 inative neural networks. In: Proceedings of Advances in Neural Information
609 Processing Systems. pp. 5596–5605.

610 Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., Belongie, S. J., 2017.
611 Learning from noisy large-scale datasets with minimal supervision. In: Pro-
612 ceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
613 pp. 6575–6583.

614 Weston, J., Ratle, F., Mobahi, H., Collobert, R., 2012. Deep learning via semi-
615 supervised embedding. In: Neural Networks: Tricks of the Trade. Springer, pp.
616 639–655.

617 Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X., 2015. Learning from massive
618 noisy labeled data for image classification. In: Proceedings of the IEEE Confer-
619 ence on Computer Vision and Pattern Recognition. pp. 2691–2699.

620 Xing, F., Xie, Y., Su, H., Liu, F., Yang, L., 2017. Deep learning in microscopy
621 image analysis: A survey. IEEE transactions on neural networks and learning
622 systems 29 (10), 4550–4568.

623 Xu, K., Chen, H., Liu, S., Chen, P.-Y., Weng, T.-W., Hong, M., Lin, X., 2019.
624 Topology attack and defense for graph neural networks: An optimization per-
625 spective. arXiv preprint arXiv:1906.04214.

626 Xu, K., Liu, S., Zhao, P., Chen, P.-Y., Zhang, H., Fan, Q., Erdogmus, D., Wang,
627 Y., Lin, X., 2018. Structured adversarial attack: Towards general implementa-
628 tion and better interpretability. arXiv preprint arXiv:1808.01664.

629 Xu, Y., Zhu, J.-Y., Eric, I., Chang, C., Lai, M., Tu, Z., 2014. Weakly super-

vised histopathology cancer image segmentation and classification. Medical image analysis 18 (3), 591–604.

Yarowsky, D., 1995. Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 189–196.

Zhang, J., Peng, Y., 2017. Ssdh: semi-supervised deep hashing for large scale image retrieval. IEEE Transactions on Circuits and Systems for Video Technology.

Zhang, X., Liu, W., Dundar, M., Badve, S., Zhang, S., 2015. Towards large-scale histopathological image analysis: Hashing-based image retrieval. IEEE Transactions on Medical Imaging 34 (2), 496–506.

Zhou, Y., Goldman, S., 2004. Democratic co-learning. In: Proceedings of International Conference on Tools with Artificial Intelligence. pp. 594–602.