Graph Segmentation-Based Pseudo-Labeling for Semi-Supervised Pathology Image Classification

HONG-KYU SHIN\textsuperscript{1}, KWANG-HYUN UHMN\textsuperscript{2}, KYUYEON CHOI, ZHIXIN XU, SEUNG-WON JUNG\textsuperscript{2}, (Senior Member, IEEE), AND SUNG-JEA KO\textsuperscript{1}, (Fellow, IEEE)

Department of Electrical Engineering, Korea University, Seoul 02841, South Korea

Corresponding author: Seung-Won Jung (swjang83@korea.ac.kr)

This work was supported in part by the Korea Medical Device Development Fund Grant by the Korean Government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health and Welfare, the Ministry of Food and Drug Safety) under Grant 1711139124 and Grant KMDF_PR_20200901_0096; in part by the Grant of the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute (KHIDI), by the Ministry of Health and Welfare, Republic of Korea, under Grant H11C0940; and in part by the Korea University Grant.

ABSTRACT
Pathology image classification is an important step in cancer diagnosis and precision treatment. Training a pathology image classification model in a fully supervised manner requires exhaustive pixel-level manual annotations from pathologists, which may not be practical in real applications. Semi-supervised learning (SSL) has been widely used to exploit large amounts of unlabeled data to facilitate model training with a small set of labeled data. However, due to the limited annotations, it still suffers from the issue of inaccurate pseudo-labels of unlabeled data. In this paper, we propose a novel framework for semi-supervised pathology image classification, which incorporates graph-based segmentation to refine initial pseudo-labels of tissue regions by considering local and global contextual relationships of patches in whole-slide images (WSIs). Moreover, we define a new energy function for graph construction that allows the graph to take into account the uncertainty of network predictions on unlabeled data. Extensive experiments on two different pathology image datasets demonstrate the effectiveness of our method compared with state-of-the-art SSL baselines. In particular, when using 5\% labeled data, our approach outperforms a strong baseline by 2.81\% AUC.

INDEX TERMS
Graph-based segmentation, pathology, pseudo-labeling, semi-supervised learning.

I. INTRODUCTION

Computational pathology has emerged as a potential standard for diagnosis with the development of scanning techniques. Consequently, there has been a rapid increase in the amount of high-resolution whole slide images (WSIs), which contribute to deep learning being the mainstream methodological choice for analyzing and interpreting histology images [1], [2], [3]. For automated diagnosis such as detecting or classifying the lesions, deep learning methods require region-level or even pixel-level annotations on WSIs since they generally take patches as inputs. Therefore, if there are sufficiently large and finely-grained annotated data on WSIs, deep learning can assure excellent and stable performance as they have already shown in natural image domains [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. However, obtaining such large and exhaustively annotated data on WSIs is an extremely laborious and time-consuming process because (1) WSIs are tremendously large high-resolution images (even with the size larger than 100,000 × 100,000 pixels) and (2) WSIs are highly heterogeneous and complex, in consequence of acquisition procedures. To acquire region-level annotations on WSIs (localizing tumor regions or isolated tumor cells), specialized pathologists need to examine tissue regions at multiple magnification levels to consider both context and details of tissue. This process hinders the potential of fully supervised learning, which requires as much annotated data as possible.

Semi-supervised learning (SSL) has been actively researched in the domain of natural images [14] and recently...
applied to the computational pathology domain to overcome the scarcity of labeled data [15], [16], [17], [18], [19], [20], [21], [22]. SSL requires a small amount of labeled data but a large amount of unlabeled data for training. Among the various SSL methods, including mean teacher [23], virtual adversarial training (VAT) [24], and pseudo-label [25], pseudo-label has been the most widely adopted in computational pathology due to its simplicity and generalizability [17], [21], [22]. However, these pseudo-label-based methods strongly depend on the reliability of pseudo-labels. If the labeled data is highly insufficient, e.g., only one annotated WSI is available, pseudo-labels are inevitably erroneous and lead to imprecise network training.

Self-supervised learning (SelfSL) that trains a model using the labels generated from the data itself can also be adopted to address the major challenge of deficient annotations. Recently, it has been shown that pathology-specific SelfSL can improve the performance of pathology image classification compared to conventional pathology-agnostic SelfSL [20]. Moreover, SelfSL can be combined with SSL when exploiting unlabeled data for training [26]. However, SelfSL only enables the model to perform the main task better by learning useful features from pretext tasks. Thus, the inherent problem of incorrect pseudo-labels cannot be simply solved with the help of SelfSL.

In this paper, we propose a novel SSL framework to solve the inherent problem of inaccurate pseudo-labels by refining them using graph-based segmentation, especially when the annotation is highly limited. Based on the ability of graph on representing and analyzing tissue structure in digital histopathology, we propose the method of refining pseudo-labels considering local and global contextual relationships between patches in WSI.

Our proposed framework first trains the deep neural network using a small amount of labeled training data and obtains inaccurate initial pseudo-labels on unlabeled training data. Considering these pseudo-labels as segmentation seeds for the segmentation of tumor regions from WSI, the graph-cut algorithm is applied to refine them and obtain more precise pseudo-labels. In particular, we newly design an energy function for better segmentation results by utilizing the network prediction. The refined labels are then used as new pseudo-labels of unlabeled data and the network is retrained on both labeled and pseudo-labeled data. Our main contributions are summarized as follows:

- We propose a novel approach to alleviate the inherent limitation of SSL, i.e., strong dependency on the labeled data, by incorporating the classical graph-based segmentation method. To the best of our knowledge, this is the first work that uses graph-based segmentation to refine pseudo-labels.
- We design a regional term correction scheme to make confident pseudo-labels more significant in graph-based segmentation.
- We achieve outperforming performance compared to conventional SSL-based pathology image classification methods on the public pathology datasets, especially when the annotation is highly limited (1-2%).

The remainder of the paper is organized as follows. Section II reviews recent work on SSL and pathology image classification. The details of our method are described in Section III. The experimental results, ablation studies, and performance comparisons are provided in Section IV. Finally, Section V presents our conclusions.

II. RELATED WORK

A. SEMI-SUPERVISED LEARNING

SSL is a copiously studied methodology to learn effectively with small labeled data and leverage a large amount of unlabeled data [14]. For the natural image classification task, various SSL methods have been developed based on pseudo-label [25], [27], [28], [29], [30], [31], consistency regularization [23], [32], virtual adversarial training (VAT) [24], and self-supervision [26]. These methods have also been combined [33], [34], [35] to improve classification performance further.

Pseudo-labeling [25] is one of the most popular SSL methods. The goal of pseudo-labeling is to create pseudo-labels for unlabeled data, where the pseudo-labels are obtained from the predictions of the model trained on labeled data. The model is then retrained using both labeled and pseudo-labeled data. Towards this direction, Lee et al. [25] directly used network predictions to obtain hard pseudo-labels. Shi et al. [27] additionally considered confidence scores based on the density of local neighbors in the feature space, and Iscen et al. [28] proposed a label propagation method to assign pseudo-labels to unlabeled data. Moreover, another study demonstrated theoretical support of using network predictions as pseudo-labels [29].

However, pseudo-labeling inherently has a strong dependency on the reliability of pseudo-labels; thus, there have been various approaches to tackle this problem. Arazo et al. [30] showed that naive pseudo-labeling is prone to overfit to incorrect pseudo-labels due to the confirmation bias and proposed to generate soft pseudo-labels. Rizve et al. [31] argued that conventional pseudo-labeling-based methods underperform due to incorrect pseudo-labels and proposed an uncertainty-aware pseudo-label selection method. Wang and Wu [29] proposed a re-prediction strategy to mitigate the problem of pseudo-labels being uncertain in updating them by network predictions. For the human pose and shape estimation, SPIN [36] addressed this issue by combining deep learning-based and optimization-based methods, where the optimization-based method is used to refine pseudo-labels.

B. PATHOLOGY IMAGE CLASSIFICATION

Due to the availability of abundant digital pathological images, the demand for the development of effective deep learning-based automated diagnosis methods has increased in clinical practice [1], [37]. Most of the existing deep learning-based classification methods use small patches
extracted from WSIs for training because of the extremely high dimensions of WSIs [4], [5], [6], [7], [8], [10], [11], [12], [13]. Earlier works on the patch-based classification used simple architectures [4], [5], [6], [7], e.g., a three-layer CNN was used to identify invasive ductal carcinoma in breast cancer [4]. In more recent years, deep architectures such as VGGNet [38], InceptionNet [39], ResNet [40], and MobileNet [41] have been used [8], [10], [11], [13]. In particular, pathologist-level classification performance was demonstrated on adenocarcinoma histology patterns using the ResNet architecture [13], and even much higher Gleason score prediction accuracy compared to the average accuracy of general pathologists was shown on prostatectomy WSIs using the InceptionNet-based architecture [10]. Although deep learning-based methods have shown pathologist-level [13] or even better [10] performances, the above patch-based methods are trained in a fully supervised manner with a massive amount of labeled data. In computational pathology, however, obtaining a large amount of labeled data for patch-based classification is challenging since patch-level annotation is laborious and time-consuming. To mitigate this challenge, there have been various attempts to apply SSL in the pathology image domain [15], [17], [18], [19], [20], [21], [22]. Peikari et al. [15] introduced a clustering-based SSL method to identify the structure of the data space and leverage the unlabeled data. Su et al. [18] proposed the mean teacher [23]-based approach for nuclear classification, which enforces local and global consistency of the predictions under different perturbations. Gao et al. [19] used MixMatch [33] to train multiple binary CNNs for renal cell carcinoma subtype classification when only few point annotations are available. Recently, Koohbanani et al. proposed an SSL-SelfSL combined approach to classify pathology images, demonstrating the superiority of pathology-specific self-supervision over pathology-agnostic self-supervision.

Pseudo-labeling-based approaches have also proven effective in the computational pathology domain [17], [21], [22]. To this end, Jaiswal et al. [17] investigated the effectiveness of pseudo-labels in breast cancer detection of lymph node metastases. In addition, Shaw et al. [21] utilized pseudo-labeling in the form of a teacher-student chain to fine-tune the model for colorectal cancer classification. Silva-Rodriguez et al. [22] also proposed a teacher-student framework for Gleason score prediction. In particular, the teacher model is trained via the multiple instance learning framework, and the student model is trained on pseudo-labels generated by the teacher model. Although the above methods [17], [21], [22] investigated the effectiveness of pseudo-labels in pathology image classification, the inherent problem of pseudo-label based-methods, i.e., the problem of strong dependency on the reliability of pseudo-labels, still exists. Therefore, it is desirable to find a way to obtain more reliable pseudo-labels and design a framework to use them as supervision for effective network training.

III. METHODS
In this section, we first formally define our problem and the notations used in this paper. We then provide an overview and details of the proposed SSL framework. Finally, we provide more details about pseudo-label refinement, a key component of our proposed method.

A. PROBLEM FORMULATION
Let $D_l = \{(x_i^l, y_i^l)\}_{i=1}^{N_l}$ and $D_u = \{x_i^u\}_{i=1}^{N_u}$ denote the labeled and unlabeled data, respectively. $x_i^l$ represents the
\( \text{\textsuperscript{p}th} \) labeled WSI consisting of image patches, i.e., \( X_i^l = \{x_{ik}\}_{k=1}^{n_i} \). The \( \text{\textsuperscript{p}th} \) unlabeled WSI \( X_j^u \) is defined similarly. \( N_l \), \( N_u \), and \( n_i \) are the numbers of labeled and unlabeled WSIs and the number of patches in \( X_i^l \), respectively. \( Y_i = \{y_{ik}\}_{k=1}^{N_l} \) is the set of classification labels corresponding to \( X_i^l \), where \( y_{ik} \in \{0, 1\} \) for our binary (normal/tumor) classification task. In general, the number of labeled WSIs is much smaller than the number of unlabeled WSIs, i.e., \( N_l \ll N_u \). The goal of the proposed method is thus to train an accurate patch-level classifier \( f \) by using both the limited labeled data \( D_l^l \) and large unlabeled data \( D_u^u \).

**B. OVERVIEW**

Fig. 1 is an illustration of the overview of the proposed framework. First, we extract all patches from tumor and normal regions in the labeled data to construct training data \( D_l^l = \{(x_i, y_i)\}_{i=1}^{N_l} \), where \( N \) is the total number of patches in \( D_l^l \), i.e., \( N = \sum_{i=1}^{N_l} n_i \). Then, our patch-level classifier \( f \) (CNN in Fig. 1) is trained on \( D_l^l \) as follows:

\[
\theta_l = \arg \min_{\theta} \sum_{(x, y) \in D_l^l} \mathcal{L}(f(x; \theta), y),
\]

where \( \theta_l \) is a set of network parameters after training and \( \mathcal{L} \) denotes the cross-entropy loss function. Next, for each WSI in the unlabeled data, we extract non-overlapping patches in the sliding-window fashion and feed them to the CNN pre-trained on \( D_l^l \). We then obtain pseudo-labels of the extracted patches according to the prediction results. Specifically, for each patch \( x_{ik} \) in \( X_j^u \), its network prediction \( p^{ik}_{jk} \) is obtained as \( f(x_{ik}; \theta_l) \). The pseudo-label \( \hat{y}_{jk} \) is then obtained by thresholding \( p^{ik}_{jk} \). For each \( X_j^u \), we obtain the set of patch-level predictions \( P^u_j = \{p^{ik}_{jk}\}_{k=1}^{n_j} \) and the set of patch-level pseudo-labels \( \hat{Y}_j^u = \{\hat{y}_{jk}\}_{k=1}^{n_j} \), where \( n_j \) is the number of patches in \( X_j^u \). From the patch-level pseudo-labels of all WSIs in the unlabeled data \( D_u^u \), we obtain the pseudo-labeled data \( D_p^u = \{X_j^u, \hat{Y}_j^u\}_{j=1}^{N_u} \). Then, \( D_p^u \) is added to the initial training data \( D_l^l \), yielding combined data \( D_{lp}^l = D_l^l \cup D_p^u \). Using \( D_{lp} \), \( f \) can be retrained as follows:

\[
\theta_{lp} = \arg \min_{\theta} \sum_{(x, y) \in D_{lp}^l} \mathcal{L}(f(x; \theta), y),
\]

where \( \theta_{lp} \) is a set of network parameters after training with \( D_{lp} \). However, the above pseudo-labeling can lead to imprecise network training since it greatly depends on the network predictions, and the network predictions can be erroneous, especially when labeled data is limited. Therefore, we propose to use graph-based segmentation for pseudo-label refinement. Specifically, for each \( X_j^u \), we first construct a graph structure based on the position and features of each patch. We then perform graph-based segmentation [42] using a set of patch-level pseudo-labels as initial seeds. We can thus obtain a set of refined-labels \( \hat{Y}_j^u \) for each \( X_j^u \). A detailed description of pseudo-label refinement is provided in the next sub-section. After completing the pseudo-label refinement for all WSIs in the unlabeled data \( D_u^u \), we obtain refined pseudo-labeled data \( D_{lp}^{r} = \{X_j^u, \hat{Y}_j^u\}_{j=1}^{N_u} \). \( D_{lp}^{r} \) is added to the initial training data \( D_l^{r} \), and the CNN is retrained on \( D_{lp}^{r} = D_l^{r} \cup D_{lp}^{r} \) as follows:

\[
\theta_{lp}^{r} = \arg \min_{\theta} \sum_{(x, y) \in D_{lp}^{r}} \mathcal{L}(f(x; \theta), y),
\]

where \( \theta_{lp}^{r} \) is a set of network parameters after training with \( D_{lp}^{r} \).

**C. PSEUDO-LABEL REFINEMENT**

A key component of the proposed framework is to refine the pseudo-label of unlabeled data and provide better supervision to the CNN for patch-level classification, as illustrated in Fig. 3. To consider the local and global contextual relationships of patches in unlabeled WSI \( X_j^u \), we first construct a slide-wise graph structure using patches from an unlabeled WSI \( X_j^u \in D_u^u \). Then, we define an energy function \( E \) to obtain a refined pseudo-label \( \hat{Y}_j^u \) from the initial pseudo-label \( \hat{Y}_j^u \). Here, the initial pseudo-label \( \hat{Y}_j^u \) is used as initial seeds for estimating the parameters of the energy function \( E \), and the prediction result \( P^u \) is used to modify the energy function \( E \) to leverage the reliability of the initial pseudo-label. The refined pseudo-label \( \hat{Y}_j^u \) is obtained by minimizing the energy function \( E \) and used to construct refined pseudo-labeled data \( D_{lp}^{r} \), which is added to the initial training data \( D_l^{r} \) to retrain the CNN using (3).

1) **GRAPH CONSTRUCTION**

A graph \( G = (\mathcal{V}, \mathcal{E}) \) consists of a set of nodes \( \mathcal{V} \) and a set of edges \( \mathcal{E} \). In this work, we use a patch-graph such that the patches from an unlabeled WSI are defined as the nodes of our graph. For simplicity, we shall omit the superscript \( u \) and the subscript \( j \). A WSI and its patches are thus denoted as \( X \) and \( \{x_{ik}\}_{k=1}^{n_i} \), respectively. Also, each pair of connected nodes is defined as a single edge \( e = \{x_p, x_q\} \in \mathcal{E} \), where \( p \) and \( q \) are node indexes, and the 8-neighbors are used to define connectivity. Here, the interconnected edges between graph nodes are called \( n\)-links (\( N \)), which represent the informative relationship between neighboring nodes. In addition to the graph nodes, there are two special terminal nodes, called \( S \) and \( T \). In our graph, \( S \) and \( T \) correspond to the tumor and normal regions, respectively.
normal nodes, respectively. Other types of edges connecting graph nodes to these terminals are called t-links (T), which are denoted by [v,S] and [v,T] for all v ∈ V. Non-negative weights w_e are assigned to every n-link and t-link. Fig. 2 shows an example of a constructed graph.

2) ENERGY FUNCTION
We define \( \tilde{Y} = [y_1, y_2, \ldots, y_n] \) as a binary vector whose element \( y_k \) specifies whether the patch \( x_k \) corresponds to tumor or normal (1 or 0), where \( n \) is the number of patches in a WSI. In other words, \( \tilde{Y} \) represents a set of segmentation labels of all patches. To obtain an optimal segmentation, an energy function that encodes the regional and boundary properties is defined as follows:

\[
E(\tilde{Y}) = \lambda R(\tilde{Y}) + B(\tilde{Y}),
\]

where \( R(\tilde{Y}) \) and \( B(\tilde{Y}) \) are the regional and boundary terms, respectively, and \( \lambda \) is a coefficient that specifies relative importance of \( R(\tilde{Y}) \) and \( B(\tilde{Y}) \). The regional term \( R(\tilde{Y}) = \sum_k R_k(y_k) \) computes the penalty of assigning \( x_k \) to tumor (\( y_k = 1 \)) or normal (\( y_k = 0 \)), which is assigned on each t-links to represent the global contextual relationships of patches in WSI. Motivated by MAP-MRF formulations [43], this penalty term is defined using a negative log-likelihood as follows:

\[
R_k(\gamma) = -\ln P(h_k | \gamma),
\]

where \( P(h_k | \gamma) \) denotes a Gaussian probability distribution function:

\[
P(h_k | \gamma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_\gamma|}} \exp \left( -\frac{1}{2} (h_k - \mu_\gamma)^T \Sigma_\gamma^{-1} (h_k - \mu_\gamma) \right),
\]

where \( h_k \) is the \( d \)-dimensional feature vector extracted from the patch \( x_k \), and \( \mu_\gamma \in \mathbb{R}^d \) and \( \Sigma_\gamma \in \mathbb{R}^{d \times d} \) are the mean vector and covariance matrix of the feature vectors, respectively. In this study, we used the average RGB vector of the patch as the feature vector, i.e., \( h_k = [r_j, g_j, b_j] \). \( \gamma \) denotes a set of class labels, and in our case, \( \gamma = \{0, 1\} \). To estimate \( \mu_\gamma \) and \( \Sigma_\gamma \), we use two Gaussian mixture models (GMMs) and apply the expectation maximization algorithm [44].

The boundary term in (4) is defined as

\[
B(\tilde{Y}) = \sum_{\{x_p, x_q\} \in \mathcal{N}} B_{x_p, x_q} \cdot \delta_{y_p \neq y_q},
\]

and \( \delta_{y_p \neq y_q} \) is given as

\[
\delta_{y_p \neq y_q} = \begin{cases} 1, & \text{if } y_p \neq y_q, \\ 0, & \text{if } y_p = y_q. \end{cases}
\]

The boundary term \( B(\tilde{Y}) \) computes the penalty of assigning different labels to two adjacent nodes, which is assigned on each n-links to represent the local contextual relationships of patches in WSI, according to the boundary penalty function \( B_{x_p, x_q} \), which measures the similarity of them. If \( x_p \) and \( x_q \) are similar in the feature space, \( B_{x_p, x_q} \) assigns a high penalty and vice versa. Specifically, the boundary penalty function is defined as

\[
B_{x_p, x_q} = \frac{\exp(-\beta \| f_p - f_q \|^2)}{d(x_p, x_q)},
\]

where \( d(x_p, x_q) = 1 \) if \( x_p \) belongs to the 4-neighbor of \( x_q \), and \( d(x_p, x_q) = \sqrt{2} \) otherwise.

3) REGIONAL TERM CORRECTION
When computing the Gaussian probability in (6), the standard graph-cut-based image segmentation [42] requires initial seed points, which are assumed to be given by user interactions. Specifically, manually labeled nodes should be given as initial seed points to estimate the parameters of each GMM. In contrast, we use the initial pseudo-labels \( \tilde{Y} \) of unlabeled data for initial seed points, which are obtained by applying the CNN parameterized by \( \theta_I \) in (1). However, using the initial
pseudo-labels \( \hat{Y} \) as the seed points to compute the regional term in (5) can lead to undesirable results since they are hard-labels obtained by thresholding the network predictions. The pseudo-labels resulting from different probabilities are expected to influence the energy function differently. In other words, the less the network prediction becomes confident, the more likely \( \hat{Y} \) becomes erroneous. To this end, we define a new regional term as follows:

\[
R'_k(\hat{Y}) = - \ln P(h_k|\hat{Y}) \cdot z_Y(p_k), \quad (10)
\]

and \( z_Y(p_k) \) is defined as

\[
z_Y(p_k) = \begin{cases} p_k, & \text{if } Y = 0, \\ 1 - p_k, & \text{if } Y = 1, \end{cases} \quad (11)
\]

where \( p_k \) and \( 1 - p_k \) represent the probability of \( x_k \) to be tumor and normal, respectively. Note that \( z_Y \) is inversely proportional to the prediction probability since the regional term measures the penalty of label assignments. Using the new regional term, the energy function in (4) is modified as:

\[
E(\hat{Y}) = \lambda R(\hat{Y}) + B(\hat{Y}), \quad (12)
\]

where \( R' = \sum_k R'_k(\hat{Y}) \).

4) OPTIMIZATION

The set of edges consisting of \( n \)-links and \( t \)-links is denoted as \( E = \mathcal{N} \cup \mathcal{T} \). Here, the boundary term and the regional term in (12) are assigned to \( \mathcal{N} \) and \( \mathcal{T} \), respectively. Given the constructed graph and its associated edges and their weights, the graph cut with the minimum energy, called min-cut, is obtained by applying the max-flow algorithm [42]. The resultant segmentation label \( \hat{Y} \) is obtained from the min-cut and used as refined pseudo-labels. After obtaining refined pseudo-labels from all unlabeled WSIs, we construct refined pseudo-labeled data \( D_{lp}^{tr} = \{X_j^u, \hat{Y}_j^u\}_j^{N_u} \) and combine \( D_{lp}^{tr} \) to \( D_{lp}^{tr} \) to retrain the CNN on new training data \( D_{lp}^{tr} = D_{lp}^{tr} \cup D_{lp}^{tr} \) via (3).

IV. EXPERIMENTS AND RESULTS

A. DATASETS

1) Camelyon16

The Camelyon16 challenge dataset [45] consists of 399 H&E stained WSIs obtained from patients who underwent surgery for breast cancer at two hospitals: Radboud University Medical Center (RUMC) and University Medical Center Utrecht (UMCU). RUMC WSIs were produced using a digital slide scanner (Pannoramic 250 Flash II, 3DHISTECH) with a 20× objective lens (specimen-level pixel size: 0.243 \( \mu m \times 0.243 \mu m \)), and UMCU WSIs were produced using a digital slide scanner (NanoZoomer-XR Digital slide scanner C12000-01, Hamamatsu Photonics) with a 40× objective lens (specimen-level pixel size: 0.226 \( \mu m \times 0.226 \mu m \)). For each WSI, tumor regions were exhaustively annotated by pathologists. The numbers of the training and test WSIs are 270 and 129, respectively. We randomly sampled 28 WSIs from the training WSIs for validation.

2) TCGA

We used another dataset from The public Cancer Genome Atlas (TCGA) repository, which provides various cancer-related data, including WSIs from multiple institutions [46], [47], [48]. We first collected a total of 2313 tissue slide images of patients with kidney cancer from the TCGA kidney renal clear cell carcinoma (TCGA-KIRC, 1637), papillary cell carcinoma (TCGA-KIRP, 471), and chromophobe (TCGA-KICH, 205) databases. We split the collected WSIs into 714 normal slides and 1599 tumor slides regardless of their sub-types. Among the 1599 tumor WSIs slides, we selected 199 WSIs with “100% percent tumor” annotation for our experiments since the WSIs from TCGA do not have region-level annotations. Of these 913 WSIs, we used 550 WSIs for training, 85 WSIs for validation, and 277 WSIs for test.

B. DATA PREPROCESSING

For both Camelyon16 and TCGA datasets, we extracted non-overlapping patches with the size of 128 × 128 at 10× magnification from each WSI. If WSIs with 10× magnification are not available, we extracted patches with the size of 256 × 256 at 20× magnification or 512 × 512 at 40× magnification and resized them to 128 × 128 for our experiments. The patches with more than 50% of the background pixels (i.e., pixels with intensity values higher than 200 in the HSV space) were removed. The overall data statistics are shown in Table 1.

C. IMPLEMENTATION DETAILS

We used the ResNet50 architecture [40] for all our experiments. We pretrained the network for a maximum of 100 epochs with the initial training data \( D_{lp}^{tr} \) from the labeled data. We terminated the training earlier when the validation loss did not decrease for five consecutive epochs and used the model with the best validation loss for inference. For this pretraining, we used the Adam optimizer, batch size of 64, and learning rate of \( 10^{-5} \).

When retraining the network with both labeled data and refined pseudo-labeled data \( D_{lp}^{tr} \), we trained the network for a maximum of 30 epochs, starting from the model with the best validation loss. For this retraining, we used 64 batches, consisting of 8 labeled patches and 56 pseudo-labeled patches. We also terminated the training earlier using the same criterion.

---

TABLE 1. Numbers of WSIs and patches in the dataset.

| Dataset          | Train   | Validation | Test    |
|------------------|---------|------------|---------|
| Camelyon16       | 242     | 28         | 129     |
| patches          | 6,181,236 | 771,745   | 3,339,672 |
| TCGA             | 350     | 85         | 277     |
| WSIs             | 2,445,539 | 386,918   | 1,235,440 |

---

H.-K. Shin et al.: Graph Segmentation-Based Pseudo-Labeling for Semi-Supervised Pathology Image Classification
TABLE 2. Experimental results on Camelyon16 with different percentages of labeled WSIs. Each score represents the mean AUC ± standard deviation obtained by five different random samplings of the training WSIs. The supervised upper bound performance using the whole training WSIs (242 WSIs) is 94.62%.

| Method             | 1%          | 2%          | 5%          | 10%         | 20%         |
|--------------------|-------------|-------------|-------------|-------------|-------------|
| Supervised baseline| 69.19±8.75  | 77.96±3.80  | 85.49±0.97  | 87.42±3.02  | 89.33±1.98  |
| Mean Teacher [23]  | 69.86±16.21 | 79.83±4.52  | 86.31±1.20  | 88.31±2.17  | 89.88±1.92  |
| VAT [24]           | 69.56±12.54 | 80.22±4.17  | 86.30±0.56  | 87.82±3.03  | 89.80±1.99  |
| MixMatch [33]      | 70.84±13.02 | 80.86±3.70  | 86.42±1.13  | 88.47±1.74  | 89.91±1.94  |
| Self-Path [20]     | 70.71±14.25 | 82.01±4.15  | 86.95±2.09  | 89.13±2.01  | 90.72±1.59  |
| Pseudo-label [25]  | 68.12±15.64 | 81.57±4.08  | 86.90±1.95  | 89.84±4.36  | 91.41±1.75  |
| Soft-label [30]    | 70.05±9.81  | 82.17±4.47  | 87.56±1.55  | 90.03±3.61  | 91.69±2.02  |
| UPS [31]           | 71.17±9.16  | 83.19±4.31  | 88.52±1.05  | 90.57±3.07  | 91.91±2.12  |
| Proposed w/o correction | 70.13±13.31 | 82.44±4.11  | 87.05±1.04  | 89.16±2.26  | 89.97±1.24  |
| Proposed           | 71.84±12.45 | 84.09±5.51  | 89.76±1.28  | 91.57±3.40  | 92.92±1.83  |

D. RESULTS

We compared our proposed framework with a set of widely used semi-supervised baselines, namely Mean Teacher [23], VAT [24], MixMatch [33], and pseudo-label [25]. We also compared ours with the state-of-the-art SSL approaches, Soft-label [30], UPS [31], and Self-Path [20]. Note that Soft-label [30] and UPS [31] are recently proposed methods to solve the problem of strong dependency on the reliability of pseudo labels, by generating soft pseudo-labels and by selecting pseudo-labels based on uncertainty, respectively. Also, Self-Path [20] is specially designed for pathology image classification with pathology-specific self-supervision. In particular, we used JigMag [20], which achieved overall high performances, as the self-supervision for implementing Self-Path.

We evaluated the model performance in terms of areas under the curves (AUCs) for different percentages of labeled training WSIs. Specifically, we randomly sampled 1%, 2%, 5%, 10%, or 20% of the WSIs from the training dataset to construct the labeled set and regarded the remaining WSIs as the unlabeled set. The supervised baseline was trained on the sampled labeled WSIs only. The other baselines were trained on both labeled and unlabeled WSIs.

1) RESULTS ON Camelyon16

As shown in Table 2, our proposed method outperforms all other baselines at all percentage settings. For instance, when the percentages of the labeled WSIs are 5%, 10% and 20%, our proposed method achieved average AUCs of 89.76%, 91.57% and 92.92%, respectively, which are the highest among all the compared methods and in particular 4.30%, 4.15% and 3.59% higher than the supervised baseline. Specifically, when the percentage of the labeled WSI is 2%, our proposed method achieved an average AUC of 84.09%, which is 6.13% higher than the supervised baseline. Furthermore, with 1% labeled WSIs, the performance of Pseudo-label [25] decreased to 68.12%, which is even inferior to the supervised baseline. This result indicates that erroneous pseudo-labels lead to imprecise network training, especially when the annotation is highly limited. Meanwhile, our proposed method achieved an average AUC of 71.84%, demonstrating that the proposed method effectively refine inaccurate pseudo-labels and thus helpful for network training. Also, the performance of the proposed method is 0.67% and 1.79% higher than that of Soft-label [30] and UPS [31], respectively, which shows that the proposed method refines pseudo-labels more precisely using graph representation. Moreover, the proposed method shows comparable performance to the fully supervised model (92.92% vs. 94.62%) using only 20% of training WSIs as labeled WSIs. We also present visualization results in Fig. 4. We can see that the proposed method predicts the tumor tissue region better than other semi-supervised baselines.

To investigate the effectiveness of the regional term correction, we compared the proposed method with and without applying it in pseudo-label refinement. Table 2 shows that the proposed method without the regional term correction outperformed several semi-supervised baselines. For example, with 2% labeled WSIs, the proposed method without the regional term correction achieved average AUCs of 82.44%, outperforming all other semi-supervised baselines except for UPS [31], which achieved average AUCs of 83.19%. With the regional term correction, the performance scores were further increased by 1.71%, 1.65%, 2.71%, 2.41% and 2.95% when the percentages of labeled WSIs are 1%, 2%, 5%, 10% and 20%, respectively, which are the highest among all other compared methods. These improvements are also evident in the WSIs overlaid with pseudo-labels, as visualized in Fig. 5. Note that the proposed method without the regional term correction significantly increased both true positives (in red boxes) and false positives (in blue boxes) compared to the initial pseudo-labels. Meanwhile, the proposed method with the regional term correction increased true positives and decreased false positives, demonstrating the effectiveness of the regional term correction.

2) RESULTS ON TCGA

Similar to the Camelyon16 dataset, we observe that our proposed method outperforms the supervised and semi-supervised baselines on the TCGA dataset as shown in Table 3. With 1%, 5%, 10% and 20% labeled WSIs, our proposed method achieved average AUCs of 75.83%,
FIGURE 4. Qualitative comparison of the proposed method and other semi-supervised baselines on the test WSI samples from Camelyon16. Only 1% of training WSIs are used as labeled WSIs. The predicted tumor regions are overlaid in red colors. Black contours indicate the boundaries of the ground truth tumor region. From left to right, the results are obtained by the supervised baseline, Mean Teacher, VAT, MixMatch, Self-Path with JigMag, Pseudo-label [25], Soft-label, UPS, and the proposed method. The red and blue boxes show the zoomed-in areas of the WSIs.

The average AUCs of the proposed method are 92.77%, 94.78% and 96.15%, respectively, which are the highest among all the compared methods and in particular 3.90%, 2.76%, 1.54% and 1.53% higher than the supervised baseline. With 2% labeled WSIs, Self-Path with JigMag achieved the highest average AUC of 84.18%, which is 0.10% better than the proposed method. With 5% labeled WSIs, the performance of the proposed method is 92.77%, which is 1.42% and 0.9% higher than the Soft-label [30] and UPS [31], respectively, demonstrating the superiority of pseudo-label refinement in proposed method. Overall, our proposed method outperforms all other approaches, especially when annotation budgets are limited. Compared to the fully supervised baseline trained on whole training WSIs, the proposed framework achieved a comparable performance of an average AUC of 96.15% using only 20% labeled WSIs.

Table 3 also demonstrates that the proposed method without the regional term correction outperformed several semi-supervised baselines on the TCGA dataset. For example, with 10% labeled WSIs, the proposed method without the regional term correction achieved average AUCs of 94.33%,
TABLE 3. Experimental results on TCGA with different percentages of labeled WSIs. Each score represents the mean AUC ± standard deviation obtained by five different random samplings of the training WSIs. The supervised upper bound performance using the whole training WSIs (550 WSIs) is 98.62%.

| Method                | % of Labeled WSIs |
|-----------------------|-------------------|
| Supervised baseline   | 1%                | 2%                | 5%                | 10%               | 20%               |
| Mean-Teacher [23]     | 71.93±12.67       | 80.43±4.33        | 90.01±1.33        | 93.24±0.91        | 94.62±0.93        |
| VAT [24]              | 73.77±5.49        | 80.74±4.02        | 90.14±0.53        | 93.30±0.78        | 95.10±1.21        |
| MixMatch [33]         | 72.90±11.17       | 81.21±4.47        | 90.20±1.29        | 93.43±0.73        | 94.66±0.67        |
| Self-Path [20]        | 73.49±9.09        | 81.24±4.00        | 90.59±1.46        | 94.19±1.05        | 95.30±1.09        |
| Pseudo-label [25]     | 74.56±8.44        | 84.18±3.64        | 90.90±0.40        | 94.25±1.36        | 95.29±0.63        |
| Soft-label [30]       | 73.48±13.71       | 83.13±4.29        | 91.01±1.60        | 94.30±0.99        | 95.39±1.26        |
| UPS [31]              | 74.09±10.93       | 83.52±5.97        | 91.35±1.38        | 94.32±0.86        | 95.38±1.19        |
| Proposed w/o correction | 75.83±10.93 | 84.08±4.59       | 92.77±1.74        | 94.78±0.95        | 96.15±1.20        |
| Proposed              | 74.85±8.78        | 83.33±4.60        | 91.05±1.23        | 94.33±0.91        | 95.04±1.10        |

FIGURE 5. WSI samples and their overlaid pseudo-labels in red colors: (a) The initial pseudo-labels obtained from the network predictions, (b) refined pseudo-labels obtained using the proposed refinement without regional term correction, (c) refined pseudo-labels obtained using the proposed refinement with regional term correction. The black contours indicate the boundaries of the ground truth tumor region. The blue and red boxes correspond to the normal and tumor region examples. Note that the regional term correction contributes to reducing the false positives (in blue boxes) while maintaining true positives (in red boxes).

outperforming all other semi-supervised baselines except for UPS [31], which achieved average AUCs of 94.50%. With the regional term correction, the performance scores were further increased by 1.28%, 0.75%, 1.72%, 0.45% and 1.11% when the percentages of labeled WSIs are 1%, 2%, 5%, 10% and 20%, respectively, which are the highest among all other compared methods.

V. CONCLUSION

In this paper, we propose a semi-supervised deep learning framework for pathology image classification, which incorporates a graph-based segmentation into pseudo-labeling process to obtain accurate labels of unlabeled data. The proposed framework refines initial pseudo-labels based on graph-based segmentation which considers local and global contextual relationships between patches in a WSI. Also, for better segmentation, we newly formulate the energy function which leverages the reliability of the initial pseudo-labels. The high-quality pseudo-labels generated by the proposed framework are used as supervision signal to train the model. The experimental results on two independent datasets demonstrate that our proposed framework outperforms state-of-the-art semi-supervised learning baselines.

The proposed method has some limitations. First, during the pseudo-label refinement process, we extract the feature vector of a given patch by simply averaging color intensities. The performance could be improved if the features that can better reflect the characteristics of pathology images are utilized. In the future, we will investigate various pathology-specific features such as the shape and morphology of cell nuclei for pseudo-label refinement. Second, the experiments were conducted for breast and kidney cancer classification tasks only. Further studies on various cancer classification tasks are desired to explore the generalizability of our approach.
[41] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, arXiv:1704.04861.

[42] Y. Boykov and G. Funka-Lea, “Graph cuts and efficient ND image segmentation,” Int. J. Comput. Vis., vol. 70, no. 2, pp. 109–131, Nov. 2006.

[43] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[44] C. Lu, M. Mahmood, N. Jha, and M. Mandal, “Automated segmentation of the melanocytes in skin histopathological images,” IEEE J. Biomed. Health Inform., vol. 17, no. 2, pp. 284–296, Mar. 2013.

[45] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, I. A. Van Der Laak, M. Hermen, Q. F. Manson, and M. Balkenhol, “Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer,” JAMA, vol. 318, no. 22, pp. 2199–2210, Dec. 2017.

[46] D. A. Gutman, J. Cobb, D. Somanna, Y. Park, F. Wang, T. Kurc, J. H. Saltz, D. J. Brat, L. A. Cooper, and J. Kong, “Cancer digital slide archive: An informatics resource to support integrated in silico analysis of TCGA pathology data,” J. Amer. Med. Inform. Assoc., vol. 20, no. 6, pp. 1091–1098, 2013.

[47] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, “The cancer genome atlas (TCGA): An immeasurable source of knowledge,” Contemp. Oncol., vol. 19, no. 1A, pp. 68–77, 2015.

[48] L. A. Cooper, E. G. Demicco, J. H. Saltz, R. T. Powell, A. Rao, and A. J. Lazar, “PanCancer insights from the cancer genome atlas: The pathologist’s perspective,” J. Pathol., vol. 244, no. 5, pp. 512–524, Apr. 2018.

HONG-KYU SHIN received the B.S. degree in electronics engineering from Korea University, in 2017. He entered the Computer Vision and Image Processing Laboratory with the Department of Electronic Engineering, Korea University, in 2017. His current research interests include image processing and computer vision.

KWANG-HYUN UHMN received the B.S. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2015 and 2022, respectively. He is currently a Research Professor with the Research Institute of Information and Communication Technology, Korea University, Seoul, South Korea.

KYUEYON CHOI received the B.S. and M.S. degrees in electrical engineering from Korea University, in 2020 and 2022, respectively. His research interests include image processing and computer vision.

ZHIXIN XU received the B.S. degree in electrical engineering from Korea University, in 2022, where he is currently pursuing the M.S. degree in electrical engineering. His current research interests include image processing and computer vision.

SEUNG-WON JUNG (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2005 and 2011, respectively. From 2011 to 2012, he was a Research Professor with the Research Institute of Information and Communication Technology, Korea University. From 2012 to 2014, he was a Research Scientist with the Samsung Advanced Institute of Technology, Yongin, South Korea. From 2014 to 2020, he was an Assistant Professor with the Department of Multimedia Engineering, Dongguk University, Seoul. In 2020, he joined the Department of Electrical Engineering, Korea University, where he is currently an Associate Professor. He has published over 70 peer-reviewed articles in international journals. His research interests include image processing and computer vision. He received the Haedong Young Scholar Award from the Institute of Electronics and Information Engineers, in 2019.

SUNG-JEA KO (Fellow, IEEE) received the B.S. degree in electronic engineering from Korea University, in 1980, and the M.S. and Ph.D. degrees in electrical and computer engineering from The State University of New York at Buffalo, in 1986 and 1988, respectively. From 1988 to 1992, he was an Assistant Professor with the Department of Electrical and Computer Engineering, University of Michigan Dearborn. In 1992, he joined the Department of Electronic Engineering, Korea University, where he is currently a Professor. He has published over 210 international journal articles. He also holds over 60 registered patents in fields, such as video signal processing and computer vision.

He is a member of the National Academy of Engineering of Korea. He was a recipient of the Best Paper Award from the IEEE Asia-Pacific Conference on Circuits and Systems, in 1996; the Haedong Best Paper Award from the Institute of Electronics and Information Engineers (IEIE), in 1997; the 1999 LG Research Award; the Research Excellence Award from Korea University, in 2004; the Technical Achievement Award from the IEEE Consumer Electronics (CE) Society, in 2012; the 15-Year Service Award from the TPC of ICCE, in 2014; the Chester Sall Award (First Place Transaction Paper Award) from the IEEE CE Society, in 2017; and the Science and Technology Achievement Medal from the Korean Government, in 2020. He served as the General Chairperson for ITC-CSCC 2012 and the General Chairperson for IEICE 2013. He was the President of the IEIE, in 2013; the Vice President of the IEEE CE Society, from 2013 to 2016; and the Distinguished Lecturer of the IEEE, from 2015 to 2017. He is also an Editorial Board Member of the IEEE TRANSACTIONS ON CONSUMER ELECTRONICS.