Diagnose Like A Pathologist: Weakly-Supervised Pathologist-Tree Network for Slide-Level Immunohistochemical Scoring

Zhen Chen[1*], **Jun Zhang**[2*], **Shuanlong Che**[3], **Junzhou Huang**[2], **Xiao Han**[2†], **Yixuan Yuan**[1†]
[1] City University of Hong Kong, Hong Kong SAR, China
[2] Tencent AI Lab, Shenzhen, China
[3] KingMed Diagnostics, Guangzhou, China
zchen.ee@my.cityu.edu.hk, junejzhang@tencent.com,shuanlong2008@sina.com, joehhuang@tencent.com, haroldhan@tencent.com, yxyuan.ee@cityu.edu.hk

Abstract

The immunohistochemistry (IHC) test of biopsy tissue is crucial to develop targeted treatment and evaluate prognosis for cancer patients. The IHC staining slide is usually digitized into the whole-slide image (WSI) with gigapixels for quantitative image analysis. To perform a whole image prediction (e.g., IHC scoring, survival prediction, and cancer grading) from this kind of high-dimensional image, algorithms are often developed based on multi-instance learning (MIL) framework. However, the multi-scale information of WSI and the associations among instances are not well explored in existing MIL based studies. Inspired by the fact that pathologists jointly analyze visual fields at multiple powers of objective for diagnostic predictions, we propose a Pathologist-Tree Network (PTree-Net) to sparsely model the WSI efficiently in multi-scale manner. Specifically, we propose a Focal-Aware Module (FAM) that can approximately estimate diagnosis-related regions with an extractor trained using the thumbnail of WSI. With the initial diagnosis-related regions, we hierarchically model the multi-scale patches in a tree structure, where both the global and local information can be captured. To explore this tree structure in an end-to-end network, we propose a patch Relevance-enhanced Graph Convolutional Network (RGCN) to explicitly model the correlations of adjacent parent-child nodes, accompanied by patch relevance to exploit the implicit contextual information among distant nodes. In addition, tree-based self-supervision is devised to improve representation learning and suppress irrelevant instances adaptively. Extensive experiments are performed on a large-scale IHC HER2 dataset. The ablation study confirms the effectiveness of our design, and our approach outperforms state-of-the-art by a large margin.

Introduction

Breast cancer (BC) is the most prevalent type of cancer among females worldwide, especially affecting women aged 20-59 years (Loibl and Gianni 2017). By revealing the proteins amplification in the cells and on cell surfaces, the immunohistochemistry (IHC) test of biopsy tissue is recommended to measure the severity of BC, which is decisive for the further treatment and prognosis of patients. Particularly, human epidermal growth factor receptor 2 (HER2) is a typical IHC diagnostic marker, which is widely used as a therapeutic target. However, the diagnosis of BC requires pathologists to perform visual inspection under the microscope, which is tedious and time-consuming. Moreover, the diagnostic accuracy and both inter-/intra-observer reliability are affected by the pathologists' experience.

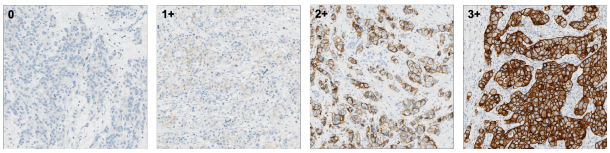

Figure 1: Typical patterns cropped from IHC stained WSIs with 0 to 3+ HER2 scores. The intensity and completeness of cell membrane staining are key evidence for HER2 scoring.

Recently, several deep learning based approaches (Zhu et al. 2017; Tokunaga et al. 2019; Xie et al. 2020) have been developed to assess the whole-slide images (WSIs). WSI is created from glass slides using specialized scanning machines (Farahani, Parwani, and Pantanowitz 2015), which can preserve the IHC slides for reproducible diagnosis. A typical WSI usually contains gigapixels, which makes it difficult to be processed by convolutional neural networks (CNNs) directly. To overcome this challenge, existing works adopted multi-instance learning (MIL), where each WSI is divided into a bag of image patches to estimate the target of the whole slide. In fact, the tissue regions (e.g., invasive tumor regions in HER2 scoring) that contribute to the diagnosis are usually sparse, compared with the large-size WSI. Some works (Wang et al. 2019; Zhao et al. 2020) required regions-of-interest (ROIs) annotations for WSI dataset, which is impractical in application scenarios. On the other hand, some other works randomly extracted patches (Courtiol et al. 2018; Hashimoto et al. 2020) or integrated all patches of tissue (Campanella

[*]Equal contribution; †Corresponding authors.
This work was done when Z. Chen interned at Tencent AI Lab.

et al. 2019) to conduct the diagnosis. However, the discriminative patches may be overwhelmed by massive diagnosis-irrelevant patches. Moreover, it is still challenging to capture both global and local structural information with the selected image patches, although some instance fusion methods (e.g., graph-learning, MinMax pooling, and bag-of-words) have been employed to integrate the local patches to represent the whole slide globally.

In this study, we would like to develop a method that automatically identifies the task-related regions and informatively represents the whole slide for IHC scoring. Motivated by the fact that pathologists jointly analyze visual fields across different powers of objective to perform diagnosis, we propose a focal-aware Pathologist-Tree Network (PTree-Net) to simulate the scenario of reading slide by pathologists. First, a preliminary CNN is trained with thumbnails of WSI pyramid using the slide-level labels (i.e., HER2 scores) to achieve an overall perception of WSI. Then, the Focal-Aware Module (FAM) is devised to revisit the feature maps of each thumbnail from the network, and generate a heatmap that highlights the task-related regions in the WSI. The heatmap indicates where should be seen, which is similar to the scenario that pathologists scan the slide using a low-power objective to discover the informative regions. With the identified attentive regions, we crop image patches layer-by-layer from the WSI pyramid in a coarse-to-fine manner. The cropped patches from high resolution layers provide detailed local information, which is similar to the scenario that pathologists check the cell-level staining condition using a higher-power objective. In order to effectively train a deep learning model using the extracted multi-scale patches, we hierarchically model them in a tree structure by considering the inclusive relationship among the thumbnail of WSI and the multi-scale patches. To explicitly model the tree-structured relationship, we propose a patch Relevance-enhanced Graph Convolutional Network (RGCN) to process coherence among patches. Also, we employ tree-based self-supervision to improve the representation learning and suppress the potential interference of irrelevant patches. In addition, we build the connection between FAM and RGCN to consistently update the selection of task-related regions. By doing so, the WSI can be explored by our PTree-Net following the habits of pathologists.

The main contributions of our work are summarized as follows.

1. We present a Pathologist-Tree Network (PTree-Net) [1] to efficiently capture and exploit multi-scale features of WSI pyramid, which simulates the scenario of reading slide by pathologists. To the best of our knowledge, our work represents the first attempt to leverage the WSI pyramid in a tree structure.
2. We propose a Focal-Aware Module (FAM) to discover the task-related regions. With multi-scale patches cropped at attentive regions, a patch Relevance-enhanced Graph Convolutional Network (RGCN) is devised to model the explicit correlations of adjacent parent-child nodes and exploit the implicit relation among distant nodes.
3. To overcome the information bottleneck of MIL framework, we propose the tree-based self-supervision, including semantic consistency and sparse constraint, to improve the representation learning and suppress the contributions of potential irrelevant patches. We also unify the FAM and RGCN to jointly guide the region selection.
4. The extensive experiments on the dataset with $1,105$ WSI of IHC HER2 slides demonstrate the effectiveness of our PTree-Net, which outperforms other multi-instance learning methods.

[1] The codes are available at https://github.com/franciszchen/PTree-Net.

Related Works

HER2 Scoring

The IHC test indicates the cancer severity, and HER2 is a targeted IHC diagnostic marker for breast cancer, which is significant to conduct treatment and prognosis. The HER2 slides can be classified into four scores (i.e., 0, 1+, 2+, and 3+) according to the American Society of Clinical Oncology and the College of American Pathologists (ASCO/CAP) guidelines (Wolff et al. 2018). Specifically, the tissues scoring 0 or 1+ contain no or faint membrane staining in less than 10% invasive tumor cells respectively, which are classified as negative. A score of 3+ represents more than 10% invasive tumor cells are observed with intense and complete circumferential staining membrane. As the borderline of two cases, those tissues with more than 10% weak membrane staining or no more than 10% strong complete membrane staining in tumor areas, are regarded as equivocal with a score of 2+. Examples of these four HER2 scores are demonstrated in Fig. 1.

Recently, there are a few studies that attempted to achieve automatic HER2 scoring. For example, Saha *et al.* (Saha and Chakraborty 2018) conducted cell segmentation and HER2 scoring using Trapezoidal LSTM units on 2048×2048 patches, rather than the entire WSI. Qaiser *et al.* (Qaiser and Rajpoot 2019) also achieved the patch-level HER2 scoring with the help of reinforcement learning. To achieve the diagnosis of gigapixels WSI, Vandenberghe *et al.* (Vandenberghe et al. 2017) and Khameneh *et al.* (Khameneh, Razavi, and Kamasak 2019) predicted the HER2 status using the results of cell classification and cell membranes segmentation, respectively. It is worth noting that both of these two works required additional pixel-wise annotations and employed human-designed rules to integrate the output of networks into the HER2 score prediction, while our PTree-Net handles the entire WSI in an end-to-end manner using only the slide-level labels.

Multi-Instance Learning

Multi-instance learning (MIL) is a typical form of weakly-supervised learning, where the label is assigned to a bag of instances while each instance within the bag has no specific label (Zhou and Xu 2007; Wang et al. 2018). Ilse *et al.* (Ilse, Tomczak, and Welling 2018) introduced two kinds of attention mechanisms into MIL framework for the first time to

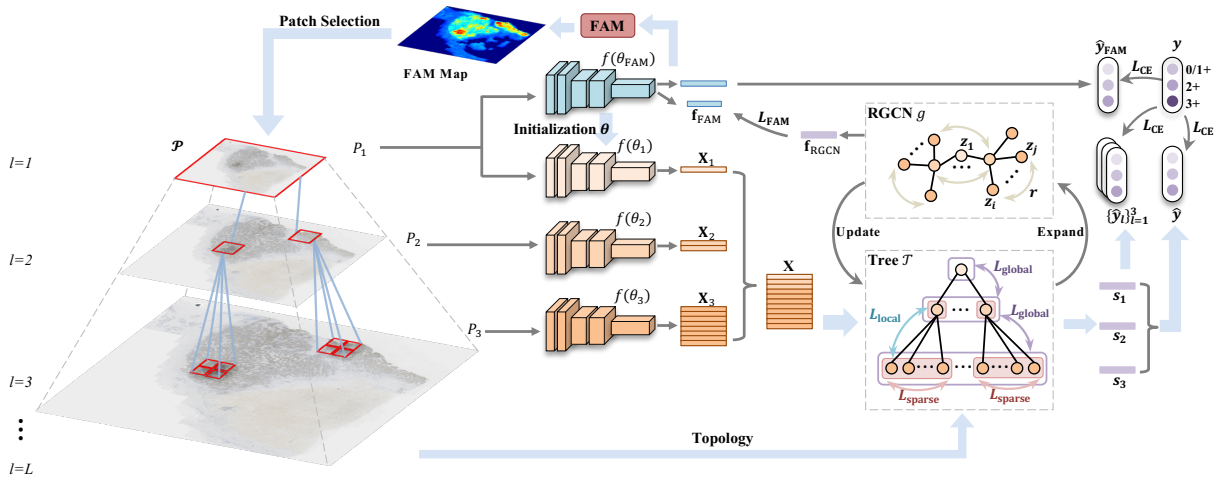

Figure 2: The framework of PTree-Net. P_l represents the set of patches extracted from the l-level WSI pyramid $\mathcal{P}$. Under the guidance of FAM, attentive patches are extracted from $\mathcal{P}$ hierarchically, followed by level-wise feature extraction networks f, and a tree structure $\mathcal{T}$ is constructed based on the topology among patches. The RGCN g iteratively processes and updates the node features of the tree $\mathcal{T}$. Finally, level-wise semantic features $\{\boldsymbol{s}_l\}_{l=1}^{3}$ are integrated from $\mathcal{T}$ and aggregated into the diagnosis $\hat{y}$. Note that PTree-Net in the figure uses L=3 levels of WSI pyramid.

adaptively adjust the importance of each instance. The MIL framework is suitable for the diagnosis of WSI slides, especially when no pixel-level/cell-level annotation is available. Specifically, Campanella *et al.* (Campanella et al. 2019) adopted a Recurrent Neural Network (RNN) to integrate the extracted features of each patches into the prediction of WSI. Wang *et al.* (Wang et al. 2019) combined global and local information to recalibrate the importance of each instance, but the localization network required the pixel-level annotations by pathologists. Hashimoto *et al.* (Hashimoto et al. 2020) randomly cropped multi-scale patches of WSI pyramid and applied instance-wise attention to aggregate the features of scale-specific extractor networks. In the experiment, we compared our work with these state-of-the-art WSI works based on MIL framework.

PTree-Net

Generally, pathologists scan the slide using a low-power objective to discover the informative regions and further check the cell-level staining condition using a higher-power objective. To simulate such reading scenario, we propose a PTree-Net to automatically discover the task-related regions and sparsely represent the WSI with tree-structured image patches, which can avoid the problem of exhaustively inspecting the entire WSI. Specifically, given a histopathology WSI pyramid $\mathcal{P}$, our PTree-Net explores regions-of-interest from top to bottom of WSI pyramid and predicts the WSI label y by jointly considering discriminative patches $\{p_k\}_{k=1}^{K} \subseteq \mathcal{P}$. The overview of our PTree-Net is illustrated in Fig. 2.

As niduses usually distribute sparsely within tissues, we devise a focal-aware module (FAM) to process the thumbnail of slide at a low-power objective and figure out the most attentive regions related with diagnostic task, by utilizing the spatial information retained in the feature maps of CNNs. With the attentive regions indicated by FAM, informative patches at different magnifications are extracted from the WSI pyramid $\mathcal{P}$ to provide more details for further investigation. It is worth noting that there is an explicit relationship among the thumbnail and these patches, where the hierarchical structure suggests a parent-child relationship between low-power objective one and high-power objective one. The global context and local appearance can be jointly captured.

Note that without ambiguity, the thumbnail is also described as a patch in some cases. Hence, we build a tree structure $\mathcal{T}$ to model this inter-patch relationship, where the thumbnail of WSI is regarded as the root node of the tree, and the attentive patches of higher-power objective serve as the child nodes, and so on. In this way, the child node represents the zoom-in visual field of the parent node. In this section, we are going to introduce how PTree-Net investigates the tree-structure in detail, including FAM, RGCN, tree-based self-supervision, and loss functions.

Focal-Aware Module

Benefiting from the spatial information retrained in the feature maps of pretrained CNNs, previous work (Zhou et al. 2016) indicated the importance of regions contributing to a specific category prediction by class activation maps (CAM). Specifically, the localization map $\mathbf{m}_c$ for category c $(1 \leq c \leq C)$ is calculated as $\mathbf{m}_c(x, y) = \sum_k \mathbf{w}(c, k)\mathbf{F}_k(x, y)$, where $\mathbf{F}$ is feature maps and $\mathbf{w}$ is parameters of the following fully-connected layer, and k represents the channel dimension. In the HER2 scoring task, instead of category-wise maps, a heatmap is expected to indicate where is the most informative for diagnosis, which guides the further patch cropping at higher resolutions.

With WSI thumbnails as input, a CNN f is trained to predict the HER2 scores. Since HER2 scoring is determined by

the cell membrane staining of invasive tumor regions in the entire slide, the thumbnail downsampled by an appropriate scale contains color and structural information to produce a roughly acceptable prediction[2], which satisfies the limitation of hardware resources. For each position (x, y), we apply a softmax function on the category-wise importance, as $\mathbf{m}_c^{'} = \frac{exp(\mathbf{m}_c)}{\sum_c exp(\mathbf{m}_c)}$, where $\mathbf{m}_c^{'}$ represents the confidence to classify this position into category c. Considering that the discriminative regions contribute to definitive diagnosis while the irrelevant areas tend to produce ambiguous results, we adopt the standard deviation of category-wise confidence to measure the spatial importance towards diagnosis.

$$\mathrm{FAM}(x, y) = \sqrt{\frac{\sum_{c=1}^{C} \left(\mathbf{m}_c' - \frac{1}{C}\right)^2}{C}}, \tag{1}$$

where $\frac{1}{C}$ represents the average of normalized importance scores. The position with larger FAM score contains more task-related information and requires further investigation of higher-power objective details.

Patch Relevance Enhanced GCN

To leverage the tree-topology in networks, we employ the Graph Convolution Network (GCN) (Kipf and Welling 2017) to exploit the features of adjacent nodes. Denote the d-dimensional feature vector of each node sent to GCN as $\boldsymbol{x} \in \mathbb{R}^d$, and features of all nodes are packed into $\mathbf{X} \in \mathbb{R}^{K \times d}$. Specifically, the graph is constructed with the nodes of tree, $\mathcal{V}(\mathcal{T})$. Since the edges of tree, $\mathcal{E}(\mathcal{T})$, represent the parent-child relationship among connected parent patch and child patch, the adjacency matrix $\mathbf{A} \in \{0, 1\}^{K \times K}$ is generated as follows:

$$\mathbf{A}(m, n) = \begin{cases} 1 & \text{if } (\boldsymbol{x}_m, \boldsymbol{x}_n) \in \mathcal{E}(\mathcal{T}) \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Consequently, the graph with adjacency matrix $\mathbf{A}$ reveals the tree topology of WSI pyramid, and GCN can explicitly integrate multi-scale features using the expanded form of tree structure. However, the global dependencies among distant nodes are ignored. To overcome this bottleneck, we devise the patch relevance to enrich GCN with implicit dependencies among distant nodes. Specifically, the affinity φ_{ij} is first calculated between every two nodes $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ at the semantic space:

$$\varphi_{ij} = (\mathbf{W}_1 \boldsymbol{x}_i)^T \cdot \mathbf{W}_2 \boldsymbol{x}_j, \tag{3}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable parameters, and $\cdot$ denotes the matrix multiplication. To refine nodes with the knowledge of entire tree, the patch relevance $\boldsymbol{r}_i$ for i-th node is generated by accumulating features of all nodes weighted by the normalized affinity:

$$\boldsymbol{r}_i = \sigma\Big(\sum_{j=1}^{K} \frac{\exp(\varphi_{ij})}{\sum_{j=1}^{K} \exp(\varphi_{ij})} \mathbf{W}_3 \boldsymbol{x}_j\Big), \tag{4}$$

where $\boldsymbol{r}_i \in \mathbb{R}^d$ and $\mathbf{W}_3$ is learnable parameters. To impose the inter-node relationship into GCN, we concatenate the node features (i.e., $\boldsymbol{x}_i$ and relevance $\boldsymbol{r}_i$) and adopt a fully-connected layer $\mathbf{W}_{\mathrm{FC}}$ to reduce the feature dimension, as $\boldsymbol{x}_i^{'} = \mathbf{W}_{\mathrm{FC}}[\boldsymbol{x}_i, \boldsymbol{r}_i]$. By introducing the patch relevance to GCN, our RGCN is enhanced with the global contextual information as:

$$\mathbf{Z} = \sigma\left(\hat{\mathbf{D}}^{-\frac{1}{2}} \hat{\mathbf{A}} \hat{\mathbf{D}}^{\frac{1}{2}} \mathbf{X}^{'} \mathbf{W}\right), \tag{5}$$

where $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix $\mathbf{A}$ with self-loops, $\hat{\mathbf{D}} = \sum_j \hat{\mathbf{A}}_{i,j}$ is diagonal degree matrix, and $\mathbf{W} \in \mathbb{R}^{d \times d^{'}}$ is a learnable matrix for any graph convolutional layer. σ is a non-linear function (i.e., ReLU operation in our method). According to the tree structure $\mathcal{T}$ in Fig. 2, $\mathbf{Z}$ is further integrated into level-wise semantic features, which are used to generate the final prediction.

Tree-based Self-Supervision

For a conventional WSI classification task, the number of samples are limited, but the image size are huge. Therefore, it is difficult to learn a robust model with the image-level label only. To address such a problem, we expect to build a self-supervision mechanism utilizing the local properties of the tree structure to alleviate the insufficiency of supervision information. For a parent node $\boldsymbol{z}^p$ and its child nodes $\{\boldsymbol{z}_q^c\}_{q=1}^{Q}$, as they present the same image region at different scales, it is intuitive to have the assumption that they hold consistent semantic information towards the diagnosis. Specifically, we calculate attention-based importance weights $\boldsymbol{w}^p \in \mathbb{R}^Q$ for child nodes of the parent node $\boldsymbol{z}^p$, as $\boldsymbol{w}^p = \frac{h((\mathbf{Z}^c)^T)}{\|h((\mathbf{Z}^c)^T)\|_{\mathrm{F}}}$, where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm and h is a 1×1 convolutional layer with output channel of Q. The Q is fixed in our implementation and the parameters of h layer are shared among subtrees, which brings a negligible overhead to PTree-Net. To suppress the interference of diagnosis-irrelevant patches, we impose the sparse constraint on the importance weights $\boldsymbol{w}^p$ of the entire tree $\mathcal{T}$ using the L1-norm (Liu et al. 2017), as follows:

$$L_{\mathrm{sparse}} = \sum_{\forall \mathcal{T}} \|\boldsymbol{w}^p\|_1. \tag{6}$$

Then, we apply the local semantic consistency between the weighted features of child nodes and their parent node:

$$L_{\mathrm{local}} = \sum_{\forall \mathcal{T}} \left\| \boldsymbol{z}^p - \sum_{q=1}^{Q} w_q^p \boldsymbol{z}_q^c \right\|^2, \tag{7}$$

where w_q^p is a scalar importance for q-th child node $\boldsymbol{z}_q^c$ of parent node $\boldsymbol{z}^p$. By doing this, the multi-scale input from the same region tends to maintain consistent semantic information. In addition, we further investigate the global semantic consistency between different levels of tree:

$$L_{\mathrm{global}} = \sum_{l=1}^{L} \|\boldsymbol{s}_l - \boldsymbol{s}_{l+1}\|^2, \tag{8}$$

where $\boldsymbol{s}_l$ is the semantic feature vector of the l-th level, which can be calculated by weighted accumulating all nodes of the l-th level to reduce the instance dimension, as $\boldsymbol{s}_l = \sum_{\forall P_{l-1}} \sum_{q=1}^{Q} w_q^p \boldsymbol{z}_q^c$. Particularly, the thumbnail node serves as the semantic feature vector of the top-level.

[2]The TN + CNN with kappa of 0.8 in Table 1 and Table 2.

Algorithm 1: The pipeline of training PTree-Net.

Input : The WSI dataset $\{\mathcal{P}_i, y_i\}_{i=1}^N$;
The feature extraction network f;
The RGCN g;
Loss factors λ_{local}, λ_{global} and λ_{sparse};
Output: The trained PTree-Net;

Initialize $f(\theta_{\text{FAM}})$ for FAM;
while $f(\theta_{\text{FAM}})$ reaches convergence **do**
  for each $\mathcal{P}$ and y in $\{\mathcal{P}_i, y_i\}_{i=1}^N$ **do**
    Extract thumbnail p_1 from WSI $\mathcal{P}$;
    Inference on $f(p_1, \theta_{\text{FAM}})$;
    Minimize $L_{\text{CE}}(f(p_1, \theta_{\text{FAM}}), y)$.
  end
end
Initialization: $\theta_1, \theta_2, \cdots, \theta_L := \theta_{\text{FAM}}$;
while g reaches convergence **do**
  for each $\mathcal{P}$ and y in $\{\mathcal{P}_i, y_i\}_{i=1}^N$ **do**
    Extract thumbnail p_1 from WSI $\mathcal{P}$;
    Obtain FAM map by inference of $f(p_1, \theta_{\text{FAM}})$;
    Extract patches from different levels of $\mathcal{P}$ according to FAM map, as $\{p_k\}_{k=1}^K \subseteq \mathcal{P}$;
    Generate adjacency matrix $\mathbf{A}$ during patch extraction;
    Obtain features of all patches $\mathbf{X}$ using corresponding $\{f(\theta_l)\}_{l=1}^L$;
    Conduct diagnosis of $\mathcal{P}$ with inference of $g(\mathbf{X}, \mathbf{A})$;
    Minimize $L(g(\mathbf{X}, \mathbf{A}), y)$ in Eq. 10;
    Update FAM by minimizing Eq. 11.
  end
end
The PTree-Net with f and g is well-trained.

Aggregation and Loss Functions

After obtaining the semantic feature vectors of all levels $\{\boldsymbol{s}_l\}_{l=1}^L$, we concatenate them in channel dimension and generate the diagnostic prediction $\hat{y}$ using a fully-connected layer. Also, each semantic feature vector $\boldsymbol{s}_l$ is equipped with an auxiliary classifier and obtains $\hat{y}_l$ to ameliorate the training. To achieve the task of HER2 scoring, we employ cross entropy loss L_{CE} on these predictions with label y:

$$L_{\text{CE}} = L_{\text{CE}}(\hat{y}, y) + \sum_{l=1}^{L} L_{\text{CE}}(\hat{y}_l, y). \tag{9}$$

Therefore, the loss function of RGCN is defined as follows:

$$L = L_{\text{CE}} + \lambda_{\text{local}} L_{\text{local}} + \lambda_{\text{global}} L_{\text{global}} + \lambda_{\text{sparse}} L_{\text{sparse}}, \tag{10}$$

where λ_{local}, λ_{global}, and λ_{sparse} are trade-off factors to adjust the importance of loss components. Considering RGCN with multi-scale inputs can produce more accurate prediction, we further update FAM with the semantic features of RGCN as well as the scoring supervision, as follows:

$$L_{\text{FAM}} = L_{\text{CE}}(\hat{y}_{\text{FAM}}, y) + \lambda_{\text{FAM}} \left\| \mathbf{f}_{\text{FAM}} - \mathbf{f}_{\text{RGCN}} \right\|_{\text{F}}, \tag{11}$$

where $\mathbf{f}_{\text{FAM}}$ and $\mathbf{f}_{\text{RGCN}}$ are the feature maps generated by independent branches derived from FAM and RGCN.

Through optimizing RGCN with Eq. 10 and FAM with Eq. 11 respectively, our PTree-Net can obtain the remarkable performance only using the weakly slide-level labels. The training pipeline is summarized in Algorithm 1.

Experiment

Dataset

Our HER2 scoring dataset consists of $1,105$ HER2-stained WSI slides, including 410 slides of 0/1+, 522 slides of 2+, and 173 ones of 3+. The sections were collected from KingMed Diagnostics and scanned with multiple image scanners. All tissues and data were retrieved under the permission of the institutional research ethics board of the institution. Region cropping is implemented to remove massive background while preserving tissues. All WSIs are standardized into a 3-level pyramid of uniform magnifications, where the magnifications of the bottom-level (l=3), the middle-level (l=2) and thumbnail (l=1) are $10\times$ ($1.0\mu\text{m/pixel}$), $5\times$ ($2.0\mu\text{m/pixel}$) and $1.25\times$ ($8.0\mu\text{m/pixel}$), respectively. The bottom-level resolution of WSI slide is up to $40,500 \times 21,000$ and the average resolution is $14,130 \times 11,030$ pixels. Correspondingly, the average resolution of thumbnails is $1,767 \times 1,379$. Each WSI is marked with a HER2 scoring label by an experienced pathologist. The comparison experiment was implemented in four-fold cross validation.

Implementation and Evaluation Strategy

Considering the medical significance of HER2 scoring, we conduct 3-category prediction, namely negative (0/1+), equivocal (2+) and positive (3+), which is consistent with previous works (Vandenberghe et al. 2017; Khameneh, Razavi, and Kamasak 2019). We adopt the ResNet-18 (He et al. 2016) for the feature extraction network as f, which converts patches into 512-dimensional features $\mathbf{X}$. The RGCN is conducted with two graph convolutional layers of Eq. 5, which processes node features into 128 dimensions and 32 dimensions, respectively. The semantic feature vector $\boldsymbol{s}$ is 32-dimensional for each level. The features to update FAM, $\mathbf{f}_{\text{RGCN}}$ and $\mathbf{f}_{\text{FAM}}$, are 64 dimensions generated by two fully-connected layers. According to the intensity of FAM map, we crop patches from unselected regions in turn, which are limited to no more than 10% WSI area and no more than 8 patches. The size of cropped patches at bottom-level and middle-level is fixed as 512×512, and each parent node contains $Q = 4$ child nodes. To avoid over-fitting, the child nodes of the same parent node are shuffled, which results in the permutation-invariant attribute for PTree-Net. In addition, we apply a foreground mask based on the threshold at RGB space to bound the heatmap generated by FAM.

All models in the experiment are implemented in PyTorch (Paszke et al. 2019) and trained until convergence on a single NVIDIA P40 GPU. The networks are optimized by Adam (Kingma and Ba 2015) with the batch size of 16. The learning rate is initialized as 1×10^{-4} and divided by 10 after every 10 epochs. We empirically set λ_{local}, λ_{global} and λ_{sparse} as 10, and λ_{FAM} as 5.

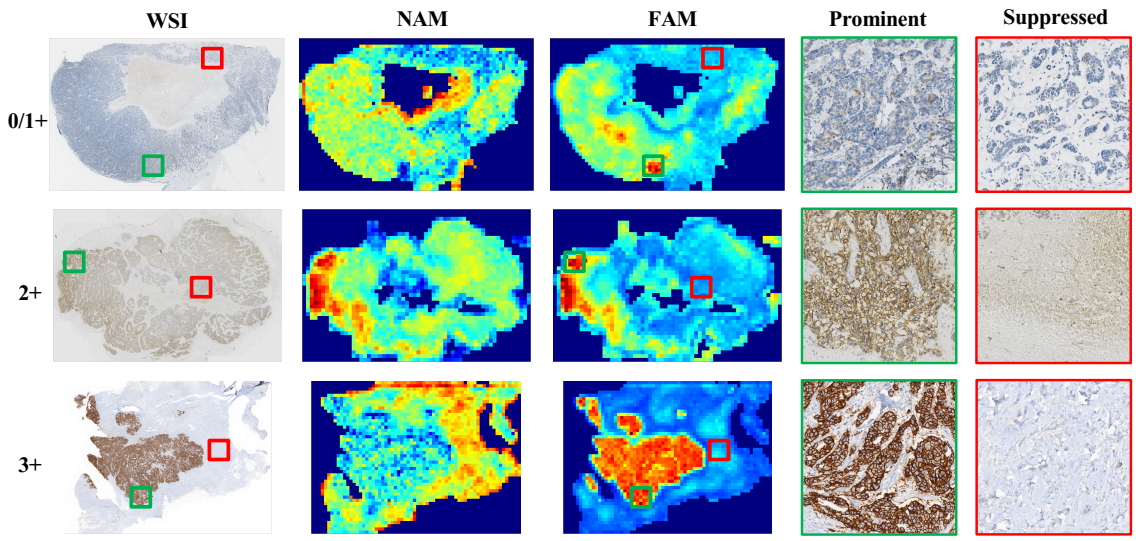

Figure 3: The visualization of FAM and detailed patches at high-power objective. Note that the heatmap of FAM and NAM (Feng et al. 2017) are generated using the same network with the same parameters. Note that the red box represents the suppressed region of FAM map, which is not utilized by the PTree-Net.

The algorithms in our experiment are evaluated by four kinds of statistical metrics, including the accuracy, the F1 score, Cohen's kappa, and Matthews correlation coefficient (MCC). Specifically, the accuracy indicates the overall agreement between predictions and labels across categories. The F1 score is first calculated in each category and then conducted in a macro average. Considering the progressive relationship of HER2 scores, we calculate Cohen's kappa with quadratic weighting, which has been widely used in similar tasks (Bulten et al. 2019). The MCC represents the consistency between predictions and labels.

Qualitative Analysis of Region Selection

To verify the effectiveness of our FAM, we demonstrate the heatmap of FAM and Nodule Activation Map (NAM) (Feng et al. 2017) at cases of various scores. In Fig. 3 (a) and (b), compared with NAM, the heatmap produced by FAM highlights sparse attentive regions and suppresses massive irrelevant areas (e.g., background and stroma cells), which is in line with the characteristics of WSI. We suppose that the softmax of category-wise importance maps can restrain noise, and the second-order information across categories is utilized in Eq. 1. In the last two columns of Fig. 3, the most attentive regions and random crops of the suppressed area are extracted from a higher-power objective. For example, the details of Fig. 3 (c) confirm the presence of intense and complete staining membrane, which provide significant information for the decision of 3+. On the other hand, the patch of suppressed area contains no epithelial cells. The other examples in Fig. 3 also confirm to this observation.

Quantitative Comparison

We evaluate the proposed PTree-Net by comparing it with other state-of-the-art methods (Courtiol et al. 2018; Ilse, Tomczak, and Welling 2018; Campanella et al. 2019; Wang et al. 2019; Hashimoto et al. 2020). As the RMDL method (Wang et al. 2019) requires pixels-level annotations to train a patch-selection network, we reproduce the RMDL in two cases, including the random cropping and FAM-based patch selection.

As shown in Table 1, our PTree-Net achieves the best performance on four evaluation metrics, including the accuracy of 89.70%, F1 of 90.77%, kappa of 89.28% and MCC of 83.34%. Compared with the baseline ResNet-18 using thumbnails of WSI, the multi-scale information and tailor-made architecture bring a 7.44% F1 and 11.83% MCC improvement to PTree-Net. The more than 2% kappa gap between the MinMax method (Courtiol et al. 2018) and attention-based MIL (Ilse, Tomczak, and Welling 2018) confirms the the contribution of instance features aggregation to WSI diagnosis, as MinMax layer discards abundant information while attention mechanism adaptively refines instances. The two rows of RMDL prove our FAM can provide the diagnosis-related regions for down-stream networks, with an increase of 3.12% in MCC over random cropping. Based on the attention mechanism in (Ilse, Tomczak, and Welling 2018), multi-scale patches further bring a 1.62% improvement of F1 to MS-DA-MIL (Hashimoto et al. 2020). Following these comparison, our PTree-Net explores multi-scale information more reasonably and integrates the features of each instance more effectively, which achieves the outperforming performance in HER2 scoring.

Ablation Study

We analyze the contributions of PTree-Net components in Table 2. The experiment is conducted on one of the four-fold divisions. Our PTree-Net achieves the accuracy of 89.74%, F1 of 90.85% and kappa of 89.57%. The comparison between *w/o* FAM and PTree-Net indicates that FAM brings

Methods	Accuracy (%)	F1 (%)	Kappa (%)	MCC (%)
Thumbnail (TN) + CNN	82.51 ± 0.55	83.33 ± 1.35	80.29 ± 2.05	71.51 ± 0.88
MinMax (Courtiol et al. 2018)	84.71 ± 0.96	85.14 ± 1.41	84.11 ± 1.26	75.79 ± 1.90
Attention MIL (Ilse, Tomczak, and Welling 2018)	86.45 ± 0.67	86.85 ± 0.58	86.86 ± 0.87	78.78 ± 1.11
Gated Attention MIL (Ilse, Tomczak, and Welling 2018)	86.81 ± 0.52	87.86 ± 0.78	86.75 ± 1.44	79.49 ± 1.48
MIL-RNN (Campanella et al. 2019)	86.18 ± 0.95	87.05 ± 1.26	85.64 ± 1.09	77.79 ± 1.53
Random+RMDL (Wang et al. 2019)	85.44 ± 1.28	85.49 ± 2.30	83.80 ± 1.90	76.02 ± 2.10
FAM+RMDL (Wang et al. 2019)	87.18 ± 0.79	87.82 ± 1.05	86.44 ± 1.29	79.14 ± 1.41
MS-DA-MIL (Hashimoto et al. 2020)	87.09 ± 0.81	88.47 ± 0.55	86.24 ± 1.18	79.07 ± 1.15
PTree-Net	$\mathbf{89.70} \pm 1.22$	$\mathbf{90.77} \pm 1.29$	$\mathbf{89.28} \pm 1.33$	$\mathbf{83.34} \pm 1.92$

Table 1: The comparison between PTree-Net and state-of-the-art methods. Average and standard deviation are calculated based on the results in four-fold cross validation

Methods	Acc (%)	F1 (%)	Kappa (%)	MCC (%)	Param (10^7)
TN + CNN	82.78	82.58	80.24	71.73	1.12
w/o FAM	87.18	87.92	86.67	79.16	3.39
fixed FAM	88.65	89.21	87.85	81.49	**4.52**
w/o RGCN	86.81	87.42	86.92	79.15	4.48
w/o Relev.	87.91	88.56	87.56	80.44	4.49
w/o TS	89.01	89.70	88.91	82.48	4.51
PTree-Net	**89.74**	**90.85**	**89.57**	**83.64**	**4.52**

Table 2: Ablation study of PTree-Net.

Target \ Predict	0/1+	2+	3+
0/1+	**88.01**±7.58	11.99±7.58	0.00±0.00
2+	9.38±2.71	**89.16**±3.43	1.46±1.13
3+	0.00±0.00	4.76±5.09	**95.24**±5.09

Table 3: 4-fold confusion matrix of PTree-Net.

a 2.93% F1 increase over random cropping, and the knowledge of RGCN to update FAM improves kappa with 1.72%. We also replace RGCN with fully-connected layers to process node features of tree, which proves that the usage of tree topology contributes a 3.43% improvement to F1. The patch relevance and the tree-based self-supervision (abbreviated as *TS*) result in a 3.20% and 1.16% improvement of MCC, respectively. Moreover, the proposed PTree-Net components are efficient, with a slight overhead of parameters.

Discussion

We further present the confusion matrix of four-fold cross validation in Table 3. PTree-Net can confidently distinguish WSIs with 3+ HER2 score, with 95.24% WSIs of 3+ are classified correctly. As the equivocal cases, 89.16% WSIs of 2+ are classified correctly, and the misjudgment between 2+ and 0/1+ is the bottleneck of performance, where 11.99% of 0/1+ samples are misjudged into 2+ and 9.38% in turn. Thus, we present two examples of these mistakes in Fig. 4. In Fig. 4 (a), PTree-Net found the region with intense and complete cell membrane staining and predicted the WSI as 2+, although the attentive red regions are very sparse. However, the proportion of this area is not enough to report the slide as HER2 2+, which inspired us to explicitly model

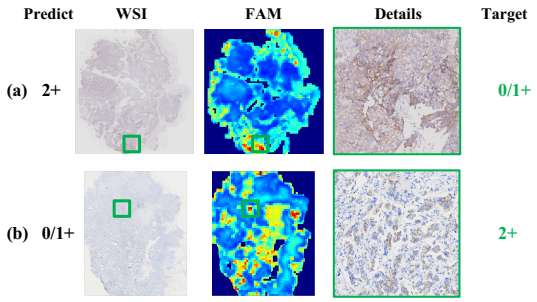

Figure 4: Misjudgment of PTree-Net between 0/1+ and 2+.

region proportion in the future. The Fig. 4 (b) presents the moderate to intense basolateral/lateral staining membranes, which is relatively rare but should be diagnosed as HER2 2+. In our experiment, it is not easy to capture such subtle patterns with not enough training samples.

Considering Vandenberghe *et al.* (Vandenberghe et al. 2017) and Khameneh *et al.* (Khameneh, Razavi, and Kamasak 2019) required additional annotations, we are not able to reproduce these two works. Particularly, Vandenberghe *et al.* (Vandenberghe et al. 2017) achieved kappa of 69% on 71 WSIs, and Khameneh *et al.* (Khameneh, Razavi, and Kamasak 2019) resulted in a 79% kappa on 52 WSIs. In contrast, our PTree-Net obtains 89.28% on four-fold cross validation of 1,105 WSIs. We attribute the performance advantage of PTree-Net to fully exploit the information of the entire WSI in an end-to-end manner, instead of human-designed features or rules.

Conclusion

In this paper, we propose a PTree-Net to hierarchically exploit the multi-scale features of WSI pyramid. Specifically, FAM is first supervised by slide labels to estimate the diagnosis-related regions. With more detailed patches from attentive regions, we devise RGCN to explore the tree structure of multi-scale patches from explicit and implicit perspectives. Finally, PTree-Net is optimized by tree-based self-supervision to improve the representation learning and suppress the contributions of potential irrelevant patches. Extensive experiments show that our PTree-Net outperforms the state-of-the-art on a large-scale IHC HER2 dataset.

Acknowledgments

This work is supported by Shenzhen-Hong Kong Innovation Circle Category D Project SGDX2019081623300177 (CityU 9240008) and CityU SRG 7005229.

References

Bulten, W.; Pinckaers, H.; van Boven, H.; Vink, R.; de Bel, T.; van Ginneken, B.; van der Laak, J.; de Kaa, C. H.-v.; and Litjens, G. 2019. Automated gleason grading of prostate biopsies using deep learning. *arXiv preprint arXiv:1907.07980* .

Campanella, G.; Hanna, M. G.; Geneslaw, L.; Miraflor, A.; Silva, V. W. K.; Busam, K. J.; Brogi, E.; Reuter, V. E.; Klimstra, D. S.; and Fuchs, T. J. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine* 25(8): 1301–1309.

Courtiol, P.; Tramel, E. W.; Sanselme, M.; and Wainrib, G. 2018. Classification and disease localization in histopathology using only global labels: A weakly-supervised approach. *arXiv preprint arXiv:1802.02212* .

Farahani, N.; Parwani, A. V.; and Pantanowitz, L. 2015. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International* 7(23-33): 4321.

Feng, X.; Yang, J.; Laine, A. F.; and Angelini, E. D. 2017. Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules. In *MICCAI*, 568–576. Springer.

Hashimoto, N.; Fukushima, D.; Koga, R.; Takagi, Y.; Ko, K.; Kohno, K.; Nakaguro, M.; Nakamura, S.; Hontani, H.; and Takeuchi, I. 2020. Multi-scale Domain-adversarial Multiple-instance CNN for Cancer Subtype Classification with Unannotated Histopathological Images. In *CVPR*, 3852–3861.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based Deep Multiple Instance Learning. In *International Conference on Machine Learning (ICML)*, 2127–2136.

Khameneh, F. D.; Razavi, S.; and Kamasak, M. 2019. Automated segmentation of cell membranes to evaluate HER2 status in whole slide images using a modified deep learning network. *Computers in biology and medicine* 110: 164–174.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Liu, Z.; Li, J.; Shen, Z.; Huang, G.; Yan, S.; and Zhang, C. 2017. Learning efficient convolutional networks through network slimming. In *ICCV*, 2736–2744.

Loibl, S.; and Gianni, L. 2017. HER2-positive breast cancer. *The Lancet* 389(10087): 2415–2429.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703* .

Qaiser, T.; and Rajpoot, N. M. 2019. Learning where to see: A novel attention model for automated immunohistochemical scoring. *IEEE Transactions on Medical Imaging* 38(11): 2620–2631.

Saha, M.; and Chakraborty, C. 2018. Her2Net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. *IEEE Transactions on Image Processing* 27(5): 2189–2200.

Tokunaga, H.; Teramoto, Y.; Yoshizawa, A.; and Bise, R. 2019. Adaptive weighting multi-field-of-view CNN for semantic segmentation in pathology. In *CVPR*, 12597–12606.

Vandenberghe, M. E.; Scott, M. L.; Scorer, P. W.; Söderberg, M.; Balcerzak, D.; and Barker, C. 2017. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Scientific reports* 7(1): 1–11.

Wang, S.; Zhu, Y.; Yu, L.; Chen, H.; Lin, H.; Wan, X.; Fan, X.; and Heng, P.-A. 2019. RMDL: Recalibrated multi-instance deep learning for whole slide gastric image classification. *Medical image analysis* 58: 101549.

Wang, X.; Yan, Y.; Tang, P.; Bai, X.; and Liu, W. 2018. Revisiting multiple instance neural networks. *Pattern Recognition* 74: 15–24.

Wolff, A. C.; Hammond, M. E. H.; Allison, K. H.; Harvey, B. E.; Mangu, P. B.; Bartlett, J. M.; Bilous, M.; Ellis, I. O.; Fitzgibbons, P.; Hanna, W.; et al. 2018. Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline focused update. *Archives of pathology & laboratory medicine* 142(11): 1364–1382.

Xie, C.; Muhammad, H.; Vanderbilt, C. M.; Caso, R.; Yarlagadda, D. V. K.; Campanella, G.; and Fuchs, T. J. 2020. Beyond Classification: Whole Slide Tissue Histopathology Analysis By End-To-End Part Learning. In *Medical Imaging with Deep Learning*, 843–856.

Zhao, Y.; Yang, F.; Fang, Y.; Liu, H.; Zhou, N.; Zhang, J.; Sun, J.; Yang, S.; Menze, B.; Fan, X.; et al. 2020. Predicting Lymph Node Metastasis Using Histopathological Images Based on Multiple Instance Learning With Deep Graph Convolution. In *CVPR*, 4837–4846.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*, 2921–2929.

Zhou, Z.-H.; and Xu, J.-M. 2007. On the relation between multi-instance learning and semi-supervised learning. In *International Conference on Machine Learning (ICML)*, 1167–1174.

Zhu, X.; Yao, J.; Zhu, F.; and Huang, J. 2017. Wsisa: Making survival prediction from whole slide histopathological images. In *CVPR*, 7234–7242.