Attention by Selection: A Deep Selective Attention Approach to Breast Cancer Classification

Bolei Xu, Jingxin Liu, Xianxu Hou, Bozhi Liu, Jon Garibaldi, *Member, IEEE*, Ian O. Ellis, Andy Green, Linlin Shen, and Guoping Qiu

***Abstract*— Deep learning approaches are widely applied to histopathological image analysis due to the impressive levels of performance achieved. However, when dealing with high-resolution histopathological images, utilizing the original image as input to the deep learning model is computationally expensive, while resizing the original image to achieve low resolution incurs information loss. Some hard-attention based approaches have emerged to select possible lesion regions from images to avoid processing the original image. However, these hard-attention based approaches usually take a long time to converge with weak guidance, and valueless patches may be trained by the classifier. To overcome this problem, we propose a deep selective attention approach that aims to select valuable regions in the original images for classification. In our approach, a decision network is developed to decide where to crop and whether the cropped patch is necessary for classification. These selected patches are then trained by the classification network, which then provides feedback to the decision network to update its selection policy. With such a co-evolution training strategy, we show that our approach can achieve a fast convergence rate and high classification accuracy. Our approach is evaluated on a public breast cancer histopathological image database, where it demonstrates superior performance compared to state-of-the-art deep learning approaches, achieving approximately 98% classification accuracy while only taking 50% of the training time of the previous hard-attention approach.**

***Index Terms*— Histopathological image, reinforcement learning, breast cancer classification, deep learning.**

I. Introduction

BREAST cancer is a major concern among women because of its higher mortality than other cancers [1]. Thus, early detection and accurate assessment are necessary to increase survival rates. In the process of a clinical breast examination, it is usually exhausting and time-consuming for pathologists to provide a diagnostic report. Therefore, there is large demand to develop computer-aided diagnosis (CADx) to relieve the workload of pathologists.

In recent years, deep learning approaches [2]–[4] have been widely applied to histopathological image analysis due to their significant performance on various medical imaging tasks. However, one issue with deep learning approaches is that the size of the original image is usually large. Directly inputting original images to the deep neural network is computationally expensive, and requires days to train on GPUs. Previous approaches address this problem by resizing images to achieve low resolution [5]–[7] or by randomly cropping patches [8] from images. However, both approaches will lead to information loss, and given that the detailed features of an image part with an abnormality could be missing, these approaches might result in misdiagnosis. Another approach is to use a sliding window [9] to crop image patches. However, a large number of patches that are not related to the lesion parts will be selected, given that the abnormality usually resides in a small portion in some cases.

Moreover, one property of the human visual system is that it does not have to process the whole image at once. Therefore, in the task of clinical diagnosis, pathologists first selectively pay attention to the abnormal region, and then investigate the

Manuscript received September 29, 2019; revised December 12, 2019; accepted December 21, 2019. Date of publication December 24, 2019; date of current version June 1, 2020. This research was supported by Natural Science Foundation of China under grants no. 61902253, 91959108, and 61672357, the Science and Technology project of Guangdong Province under grant no. 2018A050501014. *(Corresponding author: Guoping Qiu.)*

Bolei Xu, Xianxu Hou, and Bozhi Liu are with the College of Information Engineering, Shenzhen University, Shenzhen 518060, China, also with the Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen 518060, China, and also with the Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518060, China.

Jingxin Liu is with the College of Information Engineering, Shenzhen University, Shenzhen 518060, China, also with the Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen 518060, China, also with the Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518060, China, and also with the Histo Pathology Diagnostic Center, Shanghai 200444, China.

Jon Garibaldi is with the School of Computer Science, University of Nottingham, Nottingham NG7 2RD, U.K.

Ian O. Ellis and Andy Green are with the Faculty of Medicine and Health Sciences, University of Nottingham, Nottingham NG7 2RD, U.K.

Linlin Shen is with the National Engineering Laboratory for Big Data System Computing Technology, School of Computer Science and Software Engineering, Computer Vision Institute, Shenzhen University, Shenzhen 518060, China, and also with the Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China.

Guoping Qiu is with the College of Information Engineering, Shenzhen University, Shenzhen 518060, China, also with the Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen 518060, China, and also with the School of Computer Science, University of Nottingham, Nottingham NG7 2RD, U.K. (e-mail: guoping.qiu@nottingham.ac.uk).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2962013

region for details. Inspired by this human vision property, a number of works [10]–[12] apply attention-based deep learning approaches to highlight possible lesion parts in the image. There are two main kinds of attention mechanisms: hard attention and soft attention. Hard attention aims to identify a series of regions of interest in the image, while soft attention usually tries to learn weights of importance for each pixel. Since the size of histopathological images is usually large, hard attention has become more popular in some previous works [4], [13], in which the problem is formulated as a partially observed Markov decision process (POMDP) to stochastically sample patches from images through coordinates without directly working on the original image image. However, one problem with these POMDP-based methods is that the sampling process is not efficient, since valueless and redundant patches are also trained by the classifier. It thus requires a long time to achieve convergence. Moreover, in those approaches, a Long Short-Term Memory (LSTM) network not only has to sample image patches but also needs to accomplish classification or regression tasks. Therefore, the model is difficult and unstable to train.

To overcome the aforementioned problems, we propose a deep selective attention approach for histopathological image classification, which is a based on our preliminary conference paper [13]. The proposed approach contains a decision network (DeNet) and a soft-attention classification network (SaNet). The DeNet is developed to select the most useful patches from images for classification. The decision is made based on the learning progress of SaNet and statistics of the incoming data. The main difference between our approach and previous hard-attention work is that not every cropped patch is used for classification. Instead, we seek image patches that can enhance the discriminative ability of the SaNet. In some cases, a cropped patch could be abandoned even if it is related to the lesion part in the image and could be well classified by the SaNet, since this patch might have a very minor effect on improving the discriminative ability of the SaNet. In another case, those patches that are misclassified at the current stage will be selected by the DeNet to correct their predictions. Thus, with the implementation of our approach, the DeNet selects the most useful patches to train the classifier instead of using all the cropped patches as in the previous work. Such a learning strategy enables our approach to achieve a much faster training convergence rate. On the other hand, we construct two networks to conduct selection and classification tasks separately, and we also propose a co-evolution training strategy to ensure the two networks co-operate with each other in the training process; therefore, the whole framework is more stable and easier to train than previous POMDP-based approaches. We evaluate our approach on a public breast cancer dataset (BreaKHis [14]), where our approach outperforms state-of-the-art approaches with significant improvement on the classification accuracy. Moreover, we show that our approach takes much less training time than our previous POMDP-based approach [13].

The main contribution of this paper is threefold and summarized as follows. (1) A novel selective attention mechanism is proposed to find key regions from original histopathological images in BreakHis dataset. This enables SaNet to work with the most useful training samples, which can enhance the discriminative ability of the SaNet and achieve a fast convergence rate. (2) A co-evolution training strategy is also proposed to train the DeNet and SaNet simultaneously, which makes the whole framework more stable and easier to train. (3) This approach demonstrates superior performance to previous state-of-the-art methods on a public breast cancer dataset, which is important for computer-aided diagnosis of breast cancer.

II. Related Work

Two aspects of related work are now reviewed, namely attention-based deep learning approaches, and histopathological image classification.

A. Visual Attention

The concept of an attention mechanism has recently been widely used in the construction of deep neural networks, since it is able to extract meaningful features and ignore unnecessary information. Such attention mechanisms have been successfully applied to learn image features in various image classification tasks [15], [16]. They can also play an important role in other related research fields including image captioning and visual question answering [17]–[19]. Wang *et al.* [15] adopt a deep residual attention network to stack a number of attention blocks in a residual network. An attention mask is finally learned in each block to filter useful information. Hu *et al.* [16] develop a channel-wise attention mechanism to generate attentive features through learning attention weights on each channel. Chen *et al.* [20] propose to learn both spatial and channel-wise attentions in a deep convolutional neural network through a soft-max layer. Schlemper *et al.* [21] and their further work [22] also try to apply an attention mechanism to address classification and segmentation problems of medical imaging. Zhang *et al.* [11] propose an attention residual learning convolutional neural network for skin lesion classification. It consists of multiple attention residual blocks and exploits a self-attention mechanism in deep neural network. Although these attention mechanisms often significantly improve the performance of the deep neural network, they all have to work on the whole image, which requires the original image to be resized into a lower resolution or the use of a sliding window to extract patches from the image. Directly applying these strategies to datasets such as BreakHis would inevitably lead to information loss and / or high computational cost. In comparison, our approach does not directly access to the original image from BreakHis dataset, our approach could instead automatically select key regions through coordinates to save computational cost and keep the details in the images.

B. Histopathological Image Classification

Feature engineering is the main issue in achieving accurate multiclass breast cancer classification. Zhang *et al.* [23] employ kernel-based principal component analysis for benign

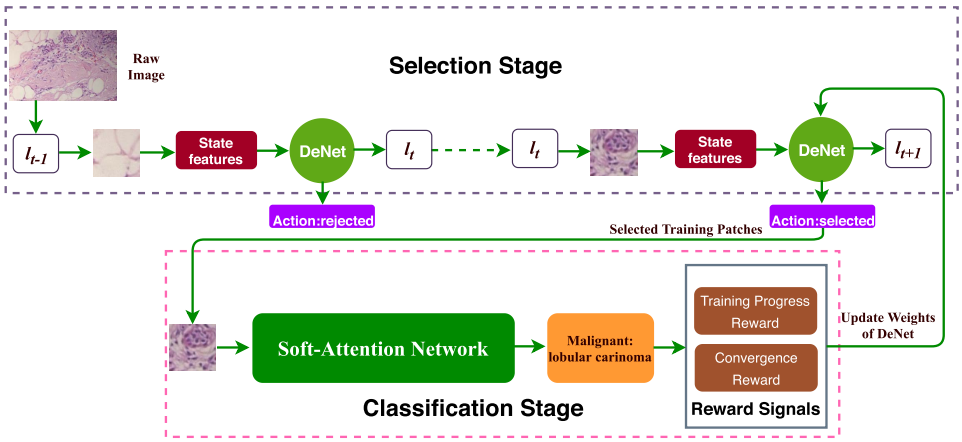


Fig. 1. The overall framework of our deep selective attention network. In each iteration, a recurrent decision network (DeNet) is designed to select possible lesion parts for each image in the mini-batch through the coordinates of the patch center. The selected patches are used to train the SaNet that is constructed based on a soft-attention mechanism. At the end of each iteration, the DeNet receives the reward signals from SaNet to update its selection policy.

and malignant classification of breast cancer histopathological images. Wang *et al.* [24] further utilize four shape and 138 textual features to achieve binary classification. Bahlmann *et al.* [25] transform the RGB patch into two channels, where one channel intensifies hematoxylin stain and another channel reveals eosin stain. These traditional approaches require manually designed features to represent the image content, which may be unable to accurately capture the key attributes of lesion areas.

Deep learning approaches have recently been applied to classification of histopathological images because of their significant performance and end-to-end training strategy. Liu *et al.* [26] propose a deep autoencoding classification network to simultaneously reconstruct and classify input images to learn robust image features. Spanhol *et al.* [9] apply a pre-trained AlexNet to exact image features. Han *et al.* [6] leverage hierarchical feature representation for breast cancer multiclassification. Their approach adopts an end-to-end training scheme to automatically learn hierarchical features from low level to high level, and considers the intraclass and interclass relations in the feature-level space.Song *et al.* [27] combine the deep features with the Fisher vectors. In their further work [28], they use the Fisher vector to encode the CNN-based local features and transform the Fisher vector to high-level discriminative feature space. Gupta and Bhavsar [29] propose to utilize joint color and texture features to classify breast histopathological images. They also explore the representation ability of features from different layers to improve the discrimination of feature representation [30]. Their latest work integrates ResNet features for breast cancer classification [31]. Our previous work [13] first regards the classification process as a POMDP, and then adopts a hybrid-attention mechanism to determine lesion parts in the original image. It enables the network to work with the selected patches instead of the whole input original image, which is able to save computational cost and focus on the lesion parts of the image. However, in this preliminary work, each image patch has to be classified during the training process, which requires a long training time to achieve convergence. To overcome this problem, we develop a new deep learning approach based on this previous work to selectively train the image patches, thus removing the redundant training sample to reduce the training time and increase the classification accuracy of the model.

III. Proposed Method

The proposed deep selective attention network model is composed of a recurrent decision network and a soft attention classification network. In each training iteration, we formulate the histopathological image classification task as a POMDP problem, which means the network does not have full access to the original image and it has to make decisions based on the current observed region. For each image in the mini-batch, it is processed by two stages including "Selection" and "Classification" as shown in Figure 1. In the "Selection" stage, we design a decision network (DeNet) to identify possible lesion regions in the original image based on a hard attention mechanism. In the "Classification" stage, a SaNet utilizes a soft attention mechanism to capture the detailed features of the selected patches and assigns labels to each input patch. In the training process, two networks co-operate with each other to achieve co-evolution. We now delve into the details of this model.

A. "Selection" Stage

In the 'Selection' stage, we design a DeNet to iteratively crop $k = 5$ regions-of-interest (ROIs) from the original image according to the coordinates of the patch center in the image, while it does not directly access the original image. The DeNet is constructed based on a recurrent LSTM network, as shown in Figure 2. LSTM is well-suited for classifying, processing and predicting time series data. In our paper, we formulate the ROI selection as a POMDP problem, where the state features (Table I) are mostly time series data (e.g. location information). The LSTM model can better predict the desired ROI by memorizing the state features in the past time-steps. In comparison, the Convolutional Neural Network is not suitable for processing time series data. When comparing to

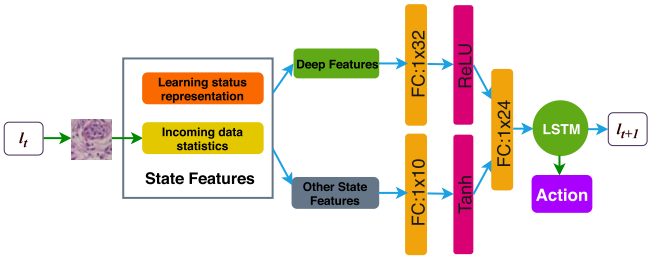


Fig. 2. The detail network structure of DeNet. In time step *t*, DeNet iteratively crops patches from each image in the mini-batch through coordinates. It then makes a decision on whether to select this patch for training and where to crop in the next time step.

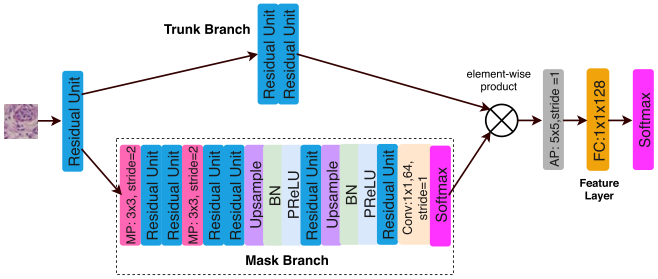


Fig. 3. The structure of SaNet. Here Conv($1 \times 1, 64$) denotes a convolutional layer with kernel size of 1×1 and 64 convolutional filters. 'BN' denotes batch normalization. MP means max-pooling layer. 'AP' denotes the global average pooling layer. 'PReLU' refers to the activation function PReLU is applied. 'Upsample' denotes upsampling by bilinear interpolation. The structure of residual unit is shown in Supplementary Material.

TABLE I
SUMMARY OF THE EACH COMPONENT OF THE STATE FEATURES

	Feature name	Length
Learning status representation	Average historical training loss	1
	The best classification result	1
	Passed iteration number	1
Incoming data statistics	Deep features	128
	Predicted label	1
	Location information l_{t-1}	2
	Ground-truth label	1

Recurrent Neural Network (RNN), LSTM can better handle the gradient vanishing problem in the training process, which means that LSTM is easier to train than the RNN [32]. In each time step, LSTM has two main jobs: (i) to decide where to crop a patch in the original image through a hard-attention mechanism; and (ii) to decide whether the cropped patch is useful to enhance the discriminative ability of SaNet.

The hard-attention mechanism in DeNet is designed to determine the possible lesion parts in each image in the mini-batch. At time step t, a *hard-attention sensor* receives a partial image patch x_t based on the location information l_{t-1} (the center coordinates of the patch in the original image) predicted by DeNet in the last time step. The cropped patch has a much smaller image size than the original image x, which is a coarse region that might be related to an abnormal part. Instead of constructing a new convolutional neural network to extract features of the cropped region, we directly use the *feature layer* of SaNet (Figure 3) to represent image features. There are two advantages to applying such a feature extraction strategy: (i) it is able to save computational cost and facilitate computational speed; and (ii) the features learned by the SaNet can be constructed as part of the learning status of SaNet, and the DeNet can make decisions based on this important feature (we will introduce the details of this point in the following section).

The DeNet parameterized by θ_d models the action policy $\pi_{\theta_d}(\mathcal{S}_t)$ to make decisions based on the state features in each time step. The state features $\mathcal{S} = (\mathcal{F}_E, \mathcal{F}_D)$ are a combination of two features: the learning status representation of SaNet $\mathcal{F}_D$ and the incoming data statistics $\mathcal{F}_E$. The *learning status representation* $\mathcal{F}_E$ of SaNet is constructed as (i) the average number of the historical training loss; (ii) the best classification result so far on the validation dataset; and (iii) the passed iteration number. The *incoming data statistics* $\mathcal{F}_D$ of the current cropped image patch consists of (i) the deep features from the *feature layer* of SaNet; (ii) the predicted label by SaNet; (iii) its ground-truth label; and (iv) the location information l_{t-1}. The details of each component are shown in Table I.

It can be seen from Table I that the length of deep features occupies a much larger proportion of the total state features than other features. Thus, directly fusing all features together will lead to an unbalanced state feature representation. To overcome this problem, we redistribute state features $\mathcal{S}$ as deep features C and the remaining state features Z: $\mathcal{S} = C \cup Z$. A fully connected layer is adopted to encode deep features into a low-dimension feature:

$$C_i' = \phi(\mathrm{W}_c(C_i) + \mathrm{b}_c) \tag{1}$$

where $\mathrm{W}_c \in \mathbb{R}^{l \times l'}$, $\mathrm{b}_c \in \mathbb{R}^{l'}$, l is the dimension of deep features, l' is the dimension of the encoded image representation and $\phi(\cdot)$ is the $ReLU$ activation function. Similarly, another fully connected layer is applied to encode the remaining features Z to generate the encoded feature Z':

$$Z_i' = \phi(\mathrm{W}_z(Z_i) + \mathrm{b}_z) \tag{2}$$

where $\mathrm{W}_z \in \mathbb{R}^{u \times u'}$, $\mathrm{b}_z \in \mathbb{R}^{u'}$, u refers to the dimension of Z and u' is the dimension of the encoded features Z'. We then construct the input state features to LSTM by concatenating C' and Z':

$$h_i = \phi(\mathrm{W}_s(Z_i' || C_i') + \mathrm{b}_s) \tag{3}$$

where $\mathrm{W}_s \in \mathbb{R}^{(l'+u') \times q}$, $\mathrm{b}_s \in \mathbb{R}^q$, q denotes the dimension of the input state features and $||$ denotes a concatenation operation. The decision actions (whether to use the current patch for training and where to crop in the next time step) are finally estimated based on the hidden feature layer of the LSTM with a sigmoid activation function. The selection process will be stopped when $k = 5$ regions are selected for each image in the mini-batch.

B. "Classification" Stage

The "Classification" stage involves a *soft-attention* mechanism $f_s(x_t; \theta_f)$ that is parameterized by θ_f and encodes the

observed image region x_t to a soft-attention map where the valuable information is highlighted. Since the cropped patch has a much smaller size than the original image, the computation of soft attention on the cropped patch requires much fewer resources than working on the original image. This is achieved by a soft attention network (SaNet) as shown in Figure 3. The SaNet contains a mask branch and a trunk branch which is modified based on the work [15]. The trunk branch consists of two residual units (the detailed structure of residual units is shown in Supplementary Material) to extract feature maps from input patches. The soft mask branch aims to learn a mask $M(x_t)$ in the range of [0, 1] by a symmetrical top-down architecture and a softmax layer to normalize the output. Specifically, we implemented max pooling layers twice in the mask branch to increase the receptive field after the residual units. This results in a lower resolution employed to collect the information of global features of the input patch. We then expanded it by performing linear interpolation twice to upsample the feature map after some residual units. It thus modified the size the feature map to be the same as the input patch. After a 1×1 convolution layer, a sigmoid layer is utilized to normalize the output range to [0,1].

The trunk branch outputs the feature map $T(x_t)$ and the mask branch outputs the attention mask $M(x_t)$. The attentive feature map is computed by:

$$\mathcal{A}(x_t) = (1 + M(x_t)) * T(x_t). \tag{4}$$

The whole Eq.(4) is similar to residual learning: in the worst case, the soft-attention mask $M(x_t)$ could be viewed as identical mapping when it approaching 0 and the $A(x_t)$ will be approximately equal to the original features $T(x_t)$, which means that the performance will be no worse than not applying soft-attention mask. The final soft-attention-based feature maps $f_s(x_t; \theta_f)$ are then learned by a global average pooling over the attention map $\mathcal{A}(x_t)$. A fully connected *feature layer* with the $ReLU$ activation function follows, and learns the feature vector of the input patch that also served as part of the state features for DeNet. Finally, we use a softmax layer to classify input patch into 8 types of histopathological tumors.

C. Reward Signal

After the data selection by DeNet, the selected data will be used to train the SaNet. There will be a new observation of state $\mathcal{S}_{t+1}$. A reward signal R_t is designed to reflect the performance of the selection mechanism. In this paper,the reward signal is designed as:

$$R = \sum_{t=1}^{T} (r_p^t + r_c^t), \tag{5}$$

where r_p denotes the training progress of SaNet and r_c represent the convergence performance of SaNet. Both r_p^t and r_c^t are set as the terminal reward: they are only computed at the final time step T in each iteration. In specific, the training progress reward r_p^t is calculated as the accuracy $\tau \in [0, 1]$ on the validation set; and the convergence rate reward r_c^t is estimated as the index i_Γ of the mini-batch in which the validation loss is below a threshold Γ:

$$r_c^T = -\log(\frac{i_\Gamma}{T'}). \tag{6}$$

where T' is a predefined maximum iteration number. It can be seen that the reward signal is designed to encourage high classification accuracy and a fast convergence rate.

D. Network Optimization

In this section, we describe how to optimize the DeNet and the SaNet. In each iteration, the SaNet minimizes cross-entropy loss:

$$\mathcal{L}_c(\hat{y_i}, y_i) = -\frac{1}{N} \sum_{i=1}^{n} [y_i \log(\hat{y_i}) + (1 - y_i) \log(1 - \hat{y_i}))], \tag{7}$$

where $\hat{y_i}$ is the estimated class label by SaNet and y_i is the ground-truth label.

Since the hard attention in DeNet is non-differentiable, we adopt policy gradient [33] to train the DeNet, in order to learn the optimal selection policy $\pi_\theta(a_t|s_{1:t})$. In this paper, we aim to maximize the reward as:

$$J(\theta) = \mathbb{E}_{p(s_{1:T};\theta)}[\sum_{t=1}^{T} (r_p^T + r_c^T)] = \mathbb{E}_{p(s_{1:T};\theta)}[R]. \tag{8}$$

To maximize J, the gradient of J can be approximated by:

$$\begin{aligned} \nabla_\theta J &= \sum_{t=1}^{T} [\nabla_\theta \log \pi_\theta(a_t|s_{1:t}) R] \\ &\approx \frac{1}{K} \sum_{j=1}^{K} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t^j|s_{1:t}^j) R^j \end{aligned} \tag{9}$$

where $j = 1 \ldots K$ is the running episodes. Equation 9 encourages the network to adjust parameters for the chosen probability of actions that would lead to a high cumulative reward, and decrease the probability of actions that would decrease the reward. Although the above gradient estimator provides us with an unbiased estimation (due to the fact that has been shown in [33], [34]), it can have a large variance that makes the training unstable. One training strategy to overcome this problem is through the subtraction of a baseline [34]:

$$\frac{1}{K} \sum_{j=1}^{K} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t^j|s_{1:t}^j)(R^j - b_t) \tag{10}$$

where b_t is the average reward value in historical epochs. The estimation of Equation 10 can have the same expectation as Equation 9 but may have lower variance.

E. Testing Phase

In the testing stage, five ROIs from each test image will be selected by DeNet. Each of the five patches will be assigned a class label by the SaNet. In some cases, there might be different labels assigned to the five patches, in which case we adopt a majority vote strategy to decide the label of the test image. For instance, if three patches are predicted as ductal

TABLE II
THE DATA STATISTICS OF BREAKHIS DATASET

Class	Subclass	#Patients (training,validation,test)	Magnification factors				Total
			40×	100×	200×	400×	
Benign	A	4 (2, 1, 1)	114	113	111	106	444
	F	10 (5, 2, 3)	253	260	264	237	1014
	TA	7 (4, 1, 2)	109	121	108	115	453
	PT	3 (1, 1, 1)	149	150	140	130	569
Malignant	DC	38 (21, 6, 11)	864	903	896	788	3451
	LC	5 (2, 1, 2)	156	170	163	137	626
	MC	9 (4, 2, 3)	205	222	196	169	792
	PC	6 (3, 1, 2)	145	142	135	138	560
Total		82 (42, 15, 25)	1995	2081	2013	1820	7909

carcinoma, while the remaining two patches are assigned labels of lobular carcinoma, then the final label assigned to the test image will be ductal carcinoma. Occasionally, five patches cannot make final decision with a majority vote (e.g. two of them are predicted as ductal carcinoma, two as lobular carcinoma, and one as mucinous carcinoma), in which case we then utilize the DeNet to select more patches until one predicted class obtains a majority.

IV. EXPERIMENT

A. Dataset

We evaluated our approach on a public dataset BreaKHis [14]. The dataset contains 7,909 images and eight subclasses of breast cancers, which have been collected from 82 patients and include 58 for malignant and 24 for benign classes. Both benign and malignant breast tumors are labeled by pathologists employing microscopes. Therefore, these tumor tissue images are captured at four kinds of optical magnifications of 40×, 100×, 200×, and 400×. The dataset contains four histopathological types of benign breast tumors: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA); and four malignant tumors: ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC). The class distribution of BreakHis dataset is shown in Table II. The digitalized images in the BreakHis dataset are the individual image patches obtained by an Olympus BX-50 system microscope with resolution of 700×460 from the breast tissue slides. For more details about the dataset, please refer to [14].

B. Implementation

In the experiment, we first randomly divide the BreaKHis dataset on patient-wise into a training (70%) dataset and a testing (30%) dataset following the experimental protocol of [14] and [6]. To estimate the convergence reward signal r_c^t in Equation 6, we further use the 25% of training dataset (i.e., 15 patients) for validation, and we use the remaining 75% of the training dataset (i.e., 42 patients) for training the networks, while the testing patient number (i.e., 25 patients) remains the same as the experimental protocol in [14] for a fair comparison. The details of data splitting of each fold is shown in Table II. In all the experiment, the training dataset is used to train the deep learning models, the validation dataset is applied to fine-tune the hyperparameters and the test dataset is used to evaluate the learned approaches. Thus, for all the tables in the experiment, we reported the classification accuracy on the test dataset. The results are obtained with the average of five trials and both classification accuracy and standard deviation are reported following the approach of previous work [14]. This experimental protocol is applied independently to every magnification as done in [14].

Before training, we augment images in BreakHis dataset by applying rotation, and horizontal and vertical flips, which results in 3 times the original training data. The image size in the dataset is 740×460. For the DeNet, the following settings are applied:

1) The weights are uniformly initialized between $(-0.01, 0.01)$.
2) The bias values are initialized as 0 in the FC layers.
3) We use l_2 normalization to normalize input state features.
4) The batch size is set to 4, the learning rate is set to 0.001 and the *Adam* optimizer is applied.
5) The threshold Γ in Equation 6 is set to 0.25, which is discussed in Section IV-J, and we set the predefined iteration number $T' = 200$.

For the SaNet, we choose the Adam optimizer with a learning rate of 0.01 that exponentially decays over epochs, and the batch size is set to 20. The experiment is conducted on a workstation with four NVIDIA 1080 Ti GPUs, and the code is implemented based on PyTorch.

TABLE III
PERFORMANCE COMPARISON OF MAGNIFICATION-SPECIFIC SYSTEM (IN %) AT THE PATIENT LEVEL. THE NUMBER INDICATES THE CLASSIFICATION ACCURACY WITH THE STANDARD DEVIATION. "N/A" DENOTES THE AUTHORS DID NOT REPORT THE CORRESPONDING DATA. "RAW" MEANS NO DATA AUGMENTATION IN TRAINING. "AUG" MEANS DATA AUGMENTATION IN TRAINING. THE RESULTS ARE EVALUATED ON THE TEST DATASET

Methods	Magnification factors			
	40×	100×	200×	400×
DRAN [15]	92.6 ± 1.8	89.0 ± 1.6	90.8 ± 2.2	91.2 ± 2.4
SEnet [16]	88.8 ± 2.1	90.2 ± 1.8	89.5 ± 2.4	92.8 ± 2.0
VGG-16 [35]	76.8 ± 2.2	82.1 ± 2.8	82.0 ± 2.2	87.9 ± 2.5
VGG-19 [35]	86.9 ± 5.2	85.4 ± 3.5	85.2 ± 4.4	85.7 ± 8.8
ResNet-50 [36]	90.5 ± 1.6	86.8 ± 2.0	93.9 ± 2.5	93.2 ± 2.1
Spanhol [14]	83.8 ± 4.1	82.1 ± 4.9	85.1 ± 3.1	82.3 ± 3.8
Spanhol [9]	90.0 ± 6.7	88.4 ± 4.8	84.6 ± 4.2	86.1 ± 6.2
Gupta *et al.* [29]	86.7 ± 2.3	88.6 ± 2.7	90.3 ± 3.7	88.3 ± 3.0
Sequential [30]	94.7 ± 0.8	95.9 ± 4.2	96.7 ± 1.1	89.1 ± 0.1
FV+CNN [27]	90.0 ± 3.2	88.9 ± 5.0	86.9 ± 5.2	86.3 ± 7.0
Song *et al.* [37]	88.5 ± 2.7	90.8 ± 4.4	89.2 ± 3.2	89.2 ± 7.9
MIL+CNN [38]	$81.3\pm$ n/a	$80.4\pm$ n/a	$77.6\pm$ n/a	$79.1\pm$ n/a
MIL [39]	$89.5\pm$ n/a	$89.0\pm$ n/a	$88.8\pm$ n/a	$87.7\pm$ n/a
PIM [31]	97.0 ± 1.2	96.1 ± 1.0	94.7 ± 1.2	90.9 ± 2.1
DenseNet [40]	$94.2\pm$ n/a	$97.9\pm$ n/a	$96.3\pm$ n/a	$95.2\pm$ n/a
CSDCNN [6]	94.1 ± 2.1	93.2 ± 1.4	94.7 ± 3.6	93.5 ± 2.7
ISBI'19 [13]	97.5 ± 1.6	96.2 ± 1.3	97.4 ± 2.5	95.4 ± 1.5
Ours (RAW)	97.8 ± 0.3	97.5 ± 0.4	97.9 ± 0.3	96.7 ± 0.4
Ours (AUG)	$\mathbf{98.2 \pm 0.2}$	$\mathbf{97.9 \pm 0.2}$	$\mathbf{98.5 \pm 0.2}$	$\mathbf{97.8 \pm 0.3}$

C. Evaluation Metrics

The performance of our approach is first evaluated by the patient recognition rate (PRR). PRR aims to calculate a ratio of correctly classified tissues to the total number of tissues.

TABLE IV
PERFORMANCE COMPARISON OF MAGNIFICATION-SPECIFIC SYSTEM (IN %) AT THE IMAGE LEVEL. THE NUMBER INDICATES THE CLASSIFICATION ACCURACY WITH THE STANDARD DEVIATION. "RAW" MEANS NO DATA AUGMENTATION IN TRAINING. "AUG" MEANS DATA AUGMENTATION IN TRAINING. THE RESULTS ARE EVALUATED ON THE TEST DATASET

Methods	Magnification factors			
	40×	100×	200×	400×
DRAN [15]	91.2 ± 2.2	88.5 ± 1.8	93.1 ± 2.8	94.5 ± 2.6
SEnet [16]	86.4 ± 2.0	91.5 ± 1.8	89.5 ± 2.4	92.8 ± 2.0
VGG-16 [35]	72.6 ± 2.8	80.0 ± 2.6	82.8 ± 2.4	88.2 ± 2.6
VGG-19 [35]	80.9 ± 1.6	81.1 ± 3.0	82.2 ± 1.9	80.2 ± 3.8
ResNet-50 [36]	90.6 ± 1.9	89.0 ± 1.8	92.8 ± 2.4	94.0 ± 2.0
CNN [9]	85.6 ± 4.8	83.5 ± 3.9	83.6 ± 1.9	80.8 ± 3.0
FV+CNN [27]	86.8 ± 2.5	85.6 ± 3.8	83.8 ± 2.5	81.6 ± 4.4
Song *et al.* [37]	87.5 ± 1.6	88.6 ± 3.6	85.5 ± 2.0	85.0 ± 4.6
CSDCNN [6]	92.8 ± 2.1	93.9 ± 1.9	93.7 ± 2.2	92.9 ± 1.8
SE-ResNet [41]	94.4 ± 0.3	94.5 ± 0.2	93.7 ± 0.1	92.9 ± 0.4
DenseNet [40]	$93.6\pm$ n/a	$97.4\pm$ n/a	$95.9\pm$ n/a	$94.7\pm$ n/a
ISBI'19 [13]	95.2 ± 1.4	94.6 ± 1.5	95.0 ± 1.6	93.8 ± 1.4
Ours (RAW)	96.5 ± 0.4	97.2 ± 0.3	98.0 ± 0.3	96.8 ± 0.5
Ours (AUG)	$\mathbf{98.0 \pm 0.2}$	$\mathbf{98.3 \pm 0.2}$	$\mathbf{98.4 \pm 0.2}$	$\mathbf{97.2 \pm 0.3}$

TABLE V
THE PERFORMANCE COMPARISON OF DIFFERENT ATTENTION STRATEGIES AT BOTH THE PATIENT LEVEL AND IMAGE LEVEL. THE RESULTS ARE THE CLASSIFICATION ACCURACY (%) WITH THE STANDARD DEVIATION. "−" DENOTES REMOVE AND "+" DENOTES THE NEW REPLACEMENT. THE RESULTS ARE EVALUATED ON THE TEST DATASET

Accuracy at	Methods	Magnification factors			
		40×	100×	200×	400×
Patient Level	-H.A	97.6 ± 1.2	96.1 ± 1.1	96.8 ± 1.4	95.4 ± 1.2
	-S.A, +ResNet-18	95.4 ± 1.2	94.2 ± 1.5	96.0 ± 1.6	94.5 ± 2.2
	-S.A, +ResNet-50	94.8 ± 1.5	95.2 ± 1.4	95.8 ± 1.5	93.2 ± 1.6
	-DeNet	91.8 ± 1.8	90.6 ± 1.9	89.8 ± 1.4	88.6 ± 1.8
	$-r_p^t$	85.5 ± 1.8	86.8 ± 2.0	83.9 ± 2.5	82.2 ± 2.1
	$-r_c^t$	98.0 ± 0.4	95.2 ± 0.4	94.8 ± 0.6	96.6 ± 0.6
	-Equation 3	93.9 ± 0.6	94.2 ± 0.5	94.5 ± 0.4	93.6 ± 0.5
	Random Selection	52.8 ± 11.3	59.6 ± 13.6	55.4 ± 10.2	58.5 ± 9.8
	Random Initialization	96.8 ± 0.3	95.2 ± 0.4	96.0 ± 0.4	94.3 ± 0.4
	Full Model	98.2 ± 0.2	97.9 ± 0.2	98.5 ± 0.2	97.8 ± 0.3
Image Level	-H.A	96.8 ± 1.2	95.4 ± 1.1	97.2 ± 1.0	95.8 ± 1.0
	-S.A, +ResNet-18	94.4 ± 0.4	95.8 ± 0.4	96.2 ± 0.2	95.2 ± 0.5
	-S.A, +ResNet-50	95.5 ± 0.5	94.0 ± 0.6	94.8 ± 0.5	96.6 ± 0.5
	-DeNet	91.5 ± 1.7	92.6 ± 2.0	90.8 ± 1.4	89.2 ± 1.8
	$-r_p^t$	82.6 ± 1.9	86.0 ± 2.2	84.8 ± 2.4	88.0 ± 2.0
	$-r_c^t$	97.8 ± 0.4	94.6 ± 0.8	95.5 ± 0.5	96.8 ± 0.6
	-Equation 3	93.2 ± 0.6	94.0 ± 0.7	93.6 ± 0.5	94.2 ± 0.5
	Random Selection	48.9 ± 11.7	50.3 ± 10.6	46.7 ± 13.2	45.2 ± 13.8
	Random Initialization	95.2 ± 0.5	96.1 ± 0.4	96.5 ± 0.5	94.8 ± 0.4
	Full Model	98.0 ± 0.2	98.3 ± 0.2	98.4 ± 0.2	97.2 ± 0.3

It is formulated as:

$$PRR = \frac{\sum_{i=1}^{N} ACC_i}{N}, \quad ACC = \frac{N_{rec}}{N_p} \tag{11}$$

where N is the total number of patients in the testing data. N_{rec} is the correctly classified tissues of patient p and N_p is the total tissue number from patient p.

We then evaluate the recognition rate at the image level (IRR), which aims to solely estimate the image classification rate without taking into account the patient information. If the network correctly classified N_{rec} images among N_{all} total images, then the recognition rate at the image level is formulated as:

$$IRR = \frac{N_{rec}}{N_{all}} \tag{12}$$

D. Comparing to Baseline Methods

As our proposed approach utilizes attention learning and residual learning, we first compared our approach with the deep residual attention learning approach (DRAN) [15] and one state-of-the-art attention network SEnet [16]. We also compared our approach with other well-known deep learning frameworks including VGG-16 [35], VGG-19 [35] and ResNet-50 [36] (all of them are first pretrained on the ImageNet dataset and then the whole network is fine-tuned on the BreaKHis dataset).

The results are shown in Table III (patient level) and Table IV (image level). It can be seen from both tables that our approaches achieves the best performance when compared with all the baseline models. The average accuracy of the patient level of our approach (AUG) is 98.1% over different magnifications, and the average accuracy of the image level is 97.9%. It can be noticed that by employing the data augmentation strategy, there is a slight performance improvement: the average accuracy increases from 97.5% (RAW) to 98.1% (AUG) at the patient level, while it increases from 96.6% (RAW) to 97.9% (AUG) at the image level. The improvement is mainly due to the larger training data size with data augmentation. It enables the network to avoid overfitting and capture more information from the augmented images. It can also be observed that the standard deviation decreases when data augmentation is applied. This means the network is more stable with a larger training dataset that learns a more discriminative feature representation.

It is clear that our approach significantly outperforms both the attention-based approaches (DRAN and SEnet) and non-attention deep learning frameworks (VGG-16, VGG-19 and ResNet-50). The reason for the superiority of our approach can be attributed to two factors: (i) we employ a hard-attention mechanism to avoid image resizing as performed in these deep learning methods, and thus prevent information loss; and (ii) since the size of the dataset is relatively small (7,909 in total), it is not necessary to employ a very deep neural network. Using a very deep network would inevitably lead to the overfitting problem, which reduces the network performance on the testing dataset. It can also be validated by the comparison between VGG-16 and VGG-19, where VGG-19 does not demonstrate superior performance to VGG-16 with a deeper network structure.

We also present the confusion matrices of different magnification factors in Figure 4. It can be seen that the confusion between PC and MC mainly contributes to the performance drop due to their high similarity.

E. Comparing to State-of-the-Art Methods

We also compared our proposed deep learning framework with the state-of-the-art approaches that reported their results on the BreaKHis Dataset ([6], [9], [13], [14], [27], [29], [30], [37]–[41]). The results are shown in Table III (patient level) and Table IV (image level), demonstrating our approach

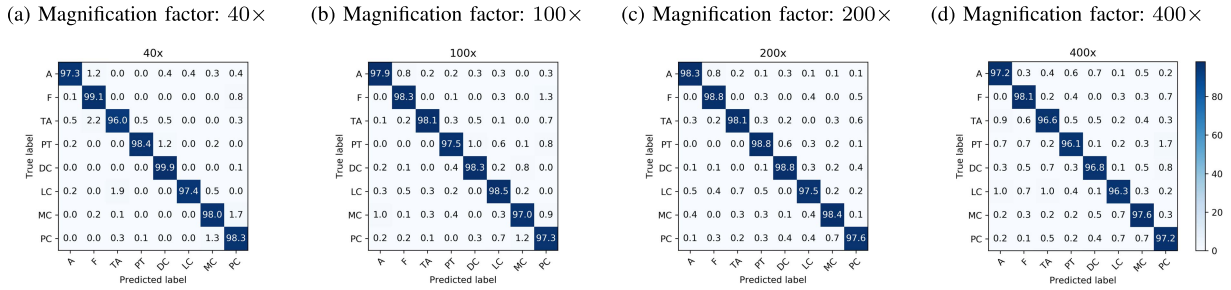


Fig. 4. Confusion matrices of different magnification factors. The number is computed as the average of five trials. The entry in the *i*-th row and *j*-th column refers to the percentage (%) of the testing images of class *i* that are classified as class *j*.

outperforms all previous approaches. It should be noted that our approach achieves much higher accuracy than most CNN approaches [27], [38], [39]. We believe this is achieved by the well-designed attention mechanisms that select useful regions for the SaNet. Specifically, the hard-attention mechanism in DeNet identifies the regions that are most related to the part with an abnormality, and the soft-attention mechanism in SaNet highlights those abnormal features. Our approach thus prevents resizing the images in BreakHis dataset which might lead to information loss, and enables the network to process the image through small-size image patches in order to save computational cost. When compared with our previous work [13], we show that the modified model in this paper can better predict class labels with higher accuracy and is more stable during training, achieving a lower standard deviation. It is mainly due to the selection mechanism involved in our new approach. The selection mechanism is able to provide the most appropriate training samples for the SaNet. It can thus prevent the SaNet from training by employing the noisy training samples, which enhances the discriminative ability of the SaNet. The low standard deviation (around 0.2) also demonstrates our approach is stable and not sensitive to the input data. The model can learn optimal parameters based on the validation dataset.

We also observed that a large number of approaches ([9], [13], [27], [30], [31], [37], [39], [41]) exhibit worse performance on 400× magnification. The main reason is that an image patch at 400× magnification is more likely to contain incomplete tissue structures with smaller receptive field to capture information from original image, which might lead to misclassification in some cases.

F. Significance Study

We compared the performance of our approach (AUG) with some previous methods which have publicly available code. These methods include DRAN [15], SEnet [16], ResNet-50 [36], and the previous state-of-the-art method ISBI'19 [13]. Friedman's test was applied to detect differences between the performance of the various methods, followed by a post-hoc application of a two-sample paired sign test on each pair of groups to identify where any differences lie. Statistical comparisons were carried out using the "friedman.test" package in R (version 3.6.1). Tests were carried out for both patient level data and image level data, for each of the magnifications 40×, 100×, 200× and 400×. The details of these statistical tests are shown in Supplementary Material. Friedman's test confirmed that there are differences among all the compared approaches, and the post-hoc test results confirmed that our approach provides statistically significant improvement over the compared methods.

G. Ablation Study

We evaluate each component in the deep selective attention framework. We design four baselines:

1) *Removing hard attention (−H.A):* we do not utilize coordinates to select image patches. Instead, we resize the image to a size of 112×112 and then use DeNet to select appropriate training images for SaNet.
2) *Removing Soft Attention (−S.A, +ResNet):* the whole SaNet is removed, and the patches selected by the DeNet are classified by ResNet-18 or ResNet-50. In other words, SaNet is replaced by ResNet-18 or ResNet-50.
3) *Removing DeNet (-DeNet):* the DeNet is removed from the framework. The image is resized to 112×112 and classified by SaNet.
4) *Removing $\boldsymbol{r}_p^t$ or $\boldsymbol{r}_c^t$ from reward function ($-\boldsymbol{r}_p^t$, $-\boldsymbol{r}_c^t$):* either $\boldsymbol{r}_p^t$ or $\boldsymbol{r}_c^t$ is removed from the reward function (Equation 5).
5) *Not Applying Equation 3 (-Equation 3):* Equation 3 is not applied to achieve balanced feature representation. In this case, the learning status representation are directly fused with the incoming data statistics to represent the input state features in DeNet (the length of the state features is 135). The state features are then fed into a fully connected layer($1 \times 1 \times 24$) and followed by the ReLU activation function to obtain the embedded state features, which are then used as the input to the LSTM.
6) *Random Selection and Random Initialization:* Random Selection means we randomly crop 5 patches from each image for training SaNet where DeNet is not applicable. Random Initialization means the weights of DeNet is randomly initialized with a normal distribution (the default mean value of normal distribution in PyTorch is 0 and standard deviation is 1.0).

Other settings remain the same as those described in Section IV-B.

TABLE VI

COMPARATIVE CLASSIFICATION RESULTS (IN %) OF DIFFERENT PATCH SIZES WITH DIFFERENT NUMBERS OF ROIS AND DIFFERENT MAGNIFICATION-SPECIFIC SYSTEMS AT THE PATIENT LEVEL. THE RESULTS ARE EVALUATED ON THE TEST DATASET (THE VALIDATION ACCURACY IS SHOWN IN BRACKETS)

Magnification Factors	Patch Size	Number of ROIs			
		3	5	7	9
40×	56 × 56	62.1 ± 4.0 (64.2 ± 3.8)	76.2 ± 1.9 (77.4 ± 1.5)	91.4 ± 1.7 (92.8 ± 1.5)	90.8 ± 0.8 (91.4 ± 1.0)
	112 × 112	74.3 ± 1.2 (75.0 ± 1.0)	**98.2 ± 0.2(98.6 ± 0.2)**	97.8 ± 0.3 (98.1 ± 0.2)	97.6 ± 0.2 (97.7 ± 0.2)
	224 × 224	83.4 ± 0.8 (86.2 ± 0.6)	97.6 ± 0.2 (98.1 ± 0.2)	96.4 ± 0.4 (98.0 ± 0.2)	96.8 ± 0.4 (97.8 ± 0.2)
100×	56 × 56	60.8 ± 4.4 (65.7 ± 3.0)	75.8 ± 1.9 (78.6 ± 1.6)	92.0 ± 1.6 (94.1 ± 1.2)	92.6 ± 0.6 (93.8 ± 0.6)
	112 × 112	76.5 ± 1.0 (78.8 ± 0.6)	**97.9 ± 0.2(98.5 ± 0.2)**	97.0 ± 0.3 (97.6 ± 0.2)	97.6 ± 0.2 (98.0 ± 0.2)
	224 × 224	82.0 ± 0.9 (83.9 ± 0.7)	97.2 ± 0.2 (98.0 ± 0.2)	96.0 ± 0.4 (97.2 ± 0.3)	96.3 ± 0.4 (96.8 ± 0.3)
200×	56 × 56	59.6 ± 3.9 (62.3 ± 2.8)	74.4 ± 1.6 (77.8 ± 1.2)	92.8 ± 1.5 (94.7 ± 1.3)	91.8 ± 0.5 (93.2 ± 0.3)
	112 × 112	78.8 ± 0.8 (80.2 ± 0.4)	**98.5 ± 0.2(99.3 ± 0.2)**	98.2 ± 0.2 (98.6 ± 0.2)	97.8 ± 0.3 (98.5 ± 0.2)
	224 × 224	80.1 ± 1.1 (82.4 ± 0.8)	95.5 ± 0.1 (96.7 ± 0.2)	94.8 ± 0.2 (95.3 ± 0.2)	96.6 ± 0.2 (97.8 ± 0.2)
400×	56 × 56	56.9 ± 4.2 (63.3 ± 3.6)	73.2 ± 1.5 (77.6 ± 1.2)	93.3 ± 1.4 (95.8 ± 1.2)	92.2 ± 0.6 (94.7 ± 0.5)
	112 × 112	79.2 ± 0.8 (83.6 ± 0.8)	**97.8 ± 0.3(98.5 ± 0.2)**	97.0 ± 0.3 (97.6 ± 0.2)	96.8 ± 0.4 (98.6 ± 0.2)
	224 × 224	78.8 ± 1.3 (84.0 ± 0.8)	95.8 ± 0.1 (97.2 ± 0.2)	95.6 ± 0.2 (97.8 ± 0.2)	96.0 ± 0.2 (98.4 ± 0.2)

TABLE VII

COMPARATIVE CLASSIFICATION RESULTS (IN %) OF DIFFERENT PATCH SIZES WITH DIFFERENT NUMBERS OF ROIS AND DIFFERENT MAGNIFICATION-SPECIFIC SYSTEMS AT THE IMAGE LEVEL. THE RESULTS ARE EVALUATED ON THE TEST DATASET (THE VALIDATION ACCURACY IS SHOWN IN BRACKETS)

Magnification Factors	Patch Size	Number of ROIs			
		3	5	7	9
40×	56 × 56	62.8 ± 4.5 (67.2 ± 3.8)	78.6 ± 2.2 (82.0 ± 1.9)	92.1 ± 1.6 (94.3 ± 1.4)	92.2 ± 0.8 (94.6 ± 0.8)
	112 × 112	75.8 ± 0.6 (78.4 ± 0.5)	**98.0 ± 0.2(99.4 ± 0.2)**	96.8 ± 0.3 (97.5 ± 0.2)	97.7 ± 0.3 (98.6 ± 0.2)
	224 × 224	72.2 ± 1.1 (74.0 ± 0.8)	97.2 ± 0.2 (98.0 ± 0.2)	97.4 ± 0.3 (98.3 ± 0.2)	96.6 ± 0.3 (97.8 ± 0.2)
100×	56 × 56	55.2 ± 4.3 (59.3 ± 3.2)	76.5 ± 2.3 (78.1 ± 1.6)	94.1 ± 1.4 (96.7 ± 1.2)	92.6 ± 0.6 (94.8 ± 0.6)
	112 × 112	72.0 ± 1.0 (73.5 ± 0.8)	**98.3 ± 0.2(99.5 ± 0.2)**	97.4 ± 0.2 (98.8 ± 0.2)	97.0 ± 0.2 (98.6 ± 0.2)
	224 × 224	74.0 ± 0.9 (75.8 ± 0.8)	98.0 ± 0.2 (98.6 ± 0.2)	95.8 ± 0.4 (97.9 ± 0.2)	96.6 ± 0.3 (97.2 ± 0.3)
200×	56 × 56	52.2 ± 4.8 (55.9 ± 3.3)	71.0 ± 2.9 (73.2 ± 2.0)	93.8 ± 1.4 (95.6 ± 1.0)	92.0 ± 0.4 (95.4 ± 1.0)
	112 × 112	71.2 ± 0.8 (72.6 ± 0.4)	**98.4 ± 0.2(99.0 ± 0.1)**	97.6 ± 0.2 (98.5 ± 0.1)	98.2 ± 0.3 (98.8 ± 0.2)
	224 × 224	77.1 ± 1.2 (79.3 ± 0.6)	96.8 ± 0.2 (97.6 ± 0.2)	95.2 ± 0.2 (97.7 ± 0.2)	96.7 ± 0.2 (98.3 ± 0.1)
400×	56 × 56	50.5 ± 4.4 (54.6 ± 3.1)	68.8 ± 2.7 (70.5 ± 1.5)	94.5 ± 1.5 (96.0 ± 1.1)	93.8 ± 0.6 (95.7 ± 0.3)
	112 × 112	70.6 ± 0.9 (72.9 ± 0.6)	**97.6 ± 0.3(98.8 ± 0.1)**	96.3 ± 0.3 (97.9 ± 0.2)	96.5 ± 0.3 (97.5 ± 0.2)
	224 × 224	79.3 ± 1.1 (82.4 ± 0.9)	97.0 ± 0.2 (98.1 ± 0.1)	96.8 ± 0.2 (97.4 ± 0.2)	96.4 ± 0.2 (97.5 ± 0.3)

The results are shown in Table V. It can be seen that the model is able to achieve the best performance when both hard and soft attention mechanisms are applied. When the hard attention is removed, we have to resize the image to a lower resolution, which inevitably leads to information loss. The details of the lesion part could be abandoned in the resizing process. Thus, the performance of classification slightly drops at both the image-level and patient-level classification. When the soft attention is replaced by the ResNet, we find that the performance also decreases dramatically. The decreased performance is due to all the image regions that are equally processed by the ResNet to extract image features. This means some redundant features are also processed by the network, which might contain noisy features that lead to misclassification. Thus, it is essential to apply the soft attention mechanism to highlight useful features and encourage the network to neglect those with unnecessary information.

DeNet is also a key component to improve the classification accuracy. We can see that there is a large reduction in the classification accuracy when DeNet is not applicable. In such a case, the image has to be resized to fit the input shape of SaNet, which leads to information loss. On the other hand, all the training images are used to train SaNet, which might include redundant and noisy samples. Thus, its classification performance is not comparable to the full model where DeNet is applied.

When evaluating the components of the reward function (Equation 5), we can see that r_p^t or r_c^t is important to enhance the classification accuracy. Specifically, the classification accuracy drops dramatically when r_p^t is removed. This means the classification accuracy on the validation dataset is a key reward signal to reflect the training progress of SaNet, and thus the DeNet can provide the most appropriate patches based on this reward signal. When r_c^t is removed from the reward function, there is a slight reduction on the classification accuracy. It is due to its ability to encourage the DeNet to reject redundant and useless training samples for classification in order to achieve a fast convergence rate. However, when r_p^t is used, the DeNet can continue to select key regions for classification. Thus, removing r_c^t will cause a smaller reduction in the accuracy than removing r_p^t.

We then evaluate the effectiveness of Equation 3 on the feature balance. It can be seen that the classification accuracy drops dramatically when Equation 3 is not applied. In that case, the deep features dominate all the remaining state features (128 versus 7), and the DeNet may heavily rely on the deep features to make decisions that could lead to non-optimal choices. Thus, it is necessary to achieve a feature balance, that is, to embed the deep features and other features, as done in Figure 2.

It can also be seen that the random selection strategy dramatically lowers the classification accuracy. The main reason is that random selection has no way of knowing which regions

contain useful information and it could instead cropping unnecessary or noisy patches for training SaNet. We also found that random initialization also slightly reduces the prediction accuracy and increases the standard deviation. The reason is that random initialization usually sets the bias term in DeNet to non-zero values, resulting in too many patches being filtered in the early stages.

H. How Does the Number and Size of ROI Affect Classification Accuracy?

We also investigate how the number and size of ROI in the selection stage affects the classification accuracy. We evaluate this by utilizing 3, 5, 7, and 9 ROIs selected by DeNet, with patch sizes set to 56×56, 112×112 and 224×224 respectively. We report different numbers of ROI and patch sizes of four different magnification factors. According to the description of BreakHis dataset, the effective pixel sizes of four magnification factors are: 0.49μm ($40\times$)), 0.20 μm ($100\times$), 0.10 μm ($200\times$), and 0.05 μm ($400\times$). The classification results of both test dataset accuracy and validation dataset are shown in Table VI (patient-level) and Table VII (image-level) respectively. We selected the hyperparameters based on the performance of validation dataset. As the validation dataset is involved in the training process to select hyperparameters, its performance is usually better than the performance of test dataset (test dataset is unseen in the training phase).

It can be seen that the best performance is achieved by selecting 5 ROIs with a patch size of 112×112. When more regions are selected, the performances are close to 5 ROIs, since the most important features are included in 5 ROIs and any additional patches would be redundant for the SaNet. When fewer regions are selected, we can observe that the classification accuracy drops dramatically. It is caused by the missing information with fewer selected patches. Additionally, it can be seen the standard deviation increases with fewer ROIs, implying that the training is unstable and not well trained with few ROIs.

Another finding which can be observed is that a large number of ROIs are required to achieve high classification accuracy when setting the patch size to 56×56. This is a reasonable result, since more information can be obtained with the increasing number of ROIs when the receptive field is small. The best performance is reached with 7 ROIs for a patch size of 56×56. However, its best performance is still lower than that obtained with patch sizes of 112×112 and 224×224. This result is caused by the small receptive field such that the soft attention mechanism is unable to capture its detailed patch features for classification. Furthermore, it can be noted that an increase in the magnification factor can decrease the classification accuracy of a small patch size (e.g., 56×56) in the case of fewer ROIs (3 and 5). The reason is that the small patch size can capture only minor information from the high-magnification factor images.

However, these results do not mean it is always better to make the patch size larger. Comparing the performance of patch sizes of 112×112 and 224×224, 224×224 demonstrates better performance only when the number of ROIs is small (e.g., 3 ROIs), since the larger patch can receive more information in that case. However, when the number of ROIs increases, there is enough information for classification, and thus, the patch size of 224×224 achieves no advantage over the patch size of 112×112.

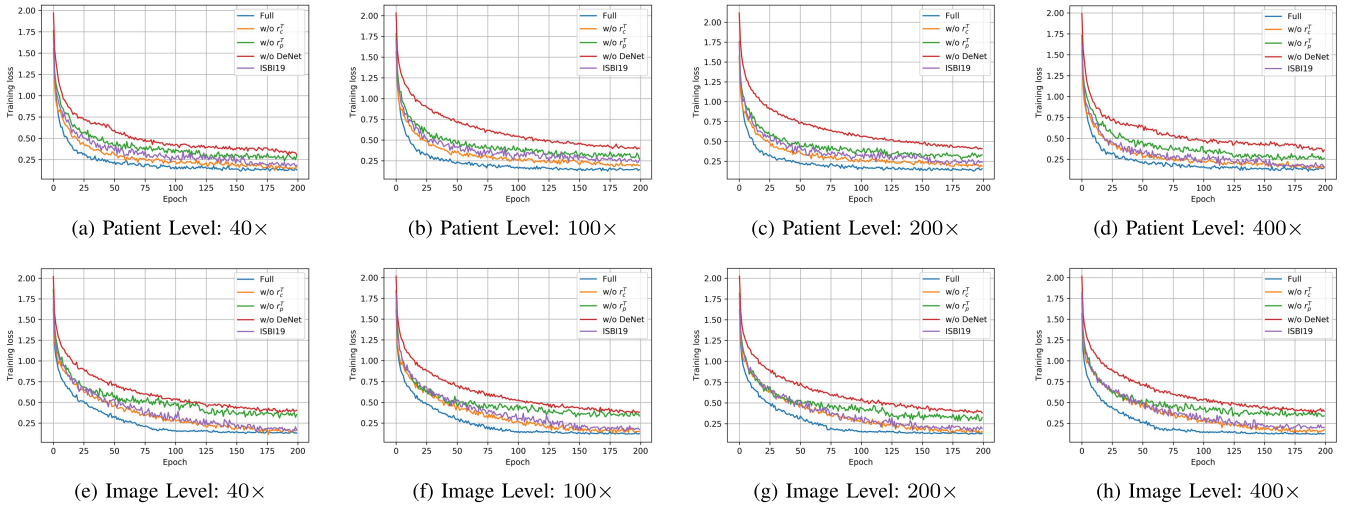
It can also be noticed that the larger number of ROIs can provide lower standard deviation of classification accuracy. This is because more training samples can provide more details of lesion part for SaNet. Thus, the SaNet can provide more stable results with the increasing number of ROIs (e.g. the standard deviation can be lower than 1 when there are 9 ROIs for all three patch sizes). However, more ROIs can also be redundant when there is already enough information for SaNet. We can see that a patch size of 112×112 results in slightly better performance than a patch size of 224×224 when the number of ROIs is larger than 3. The main reason is that redundant and possibly noisy features can be included in the larger receptive field patch when there is already enough feature information for classification. Thus, we found that by setting the patch size to 112×112 and the patch number to 5, the model is able to reach the best classification performance. We have also visualized the selected patches in the supplementary material.

I. Convergence Analysis

We then compare the convergence performances between our approach and four different baselines.The four baseline models are (1) removal of the reward signal r_p^T from Equation 5 (w/o r_p^T); (2) removal of the reward signal r_c^T from Equation 5; (3) removal of the whole DeNet and use of only the SaNet for training (w/o DeNet); and (4) our previous POMDP approach [13] (ISBI'19). We calculate the cross-entropy classification loss by the end of each epoch on the training dataset. The results of different optical magnifications of both patient and image levels are shown in Figure 5.

We observe that the application of DeNet can contribute significantly to achieving faster convergence and lower loss. When DeNet is applied, it takes approximately 75 to 100 epochs to achieve convergence with a lower loss value. When DeNet is not applied, it requires more than 175 epochs to converge and causes a relatively higher training loss. This shows that DeNet can effectively select the most appropriate training data for classification, and therefore the redundant data are removed from training to achieve a fast convergence rate.

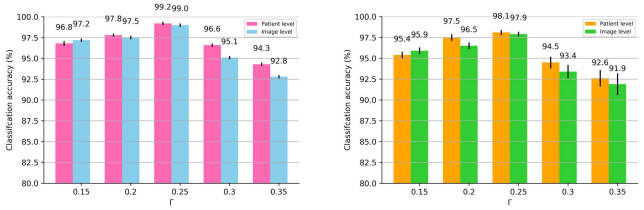
The reward signal r_p^T, which reflects the classification ability of SaNet, is also a key implementation to achieve early convergence. We can see that the classification loss will converge to a relatively higher value when r_p^T is absent from the reward function. This means that the DeNet cannot provide appropriate training samples for SaNet, and the SaNet will converge to a local minimum in the gradient descent optimization process. Similarly, the reward signal r_c^T also contributes to network convergence. It can be observed from the experimental results that the application of r_c^T demonstrates a better ability to achieve a fast convergence rate than when not applying it. The reason is that r_c^T is a signal to indicate how fast the network can achieve a low loss value, and it thus encourages DeNet



(a) Patient Level: 40× (b) Patient Level: 100× (c) Patient Level: 200× (d) Patient Level: 400×

(e) Image Level: 40× (f) Image Level: 100× (g) Image Level: 200× (h) Image Level: 400×

Fig. 5. We evaluate the effectiveness of different baseline models with respect to their convergence rates. The evaluation is conducted with four magnification factors individually, and both image-level and patient-level classification losses are presented.



(a) Validation set (b) Test set

Fig. 6. Comparative results of different values of threshold Γ on the validation adn test datasets. The results are the average classification accuracy over four different magnifications.

to select optimal training samples to accelerate the training process. It can also be seen that the design of the reward signal is a key element to achieve stable training. When either r_p^T or r_c^T is absent in the reward function, the training loss fluctuation increases compared to the full model and even the sole SaNet (w/o DeNet) scenarios.

When comparing with our previous work [13], we observe that the proposed approach in this paper is more stable to train and able to quickly reach a lower loss value. This is mainly due to the patch selection mechanism developed in this paper, which prevents every cropped patch from being classified in the training stage. In addition, in our previous work [13], the patch cropping and classification tasks are accomplished in the same network, which makes it difficult for training. In this paper, we divide the two tasks between DeNet and SaNet, and we also develop a training strategy to make the two networks co-operate with each other in the training phase. It thus makes the whole framework easier and more stable to train.

J. Threshold Analysis

Next, we evaluate the influence of threshold Γ in Equation 6 on the classification performance. The classification results of validation dataset and test dataset are shown in Figure 6. The value of Γ is selected based on the performance of validation dataset. It can be seen that the best performance is achieved when setting $\Gamma = 0.25$. When setting Γ to a higher value, the classification accuracy drops dramatically. The reason is that it is easy to achieve such a high validation loss at the early stage of training, which reduces the incentive of SaNet to reach a lower validation loss. When setting Γ to a smaller value, there is a minor reduction on the classification accuracy. The reason is that it is difficult for the SaNet to achieve such a low validation loss, and thus the reward gain between each iteration is relatively smaller. This makes it ambiguous for the DeNet to update its selection policy with such tiny changes in the reward feedback.

K. Computational Time Analysis

Finally, in our experiment, the model takes approximately 4 hours to train on a workstation with four NVIDIA GTX 1080 Ti GPUs. In comparison, the model in our previous work [13] takes approximately 8 hours to achieve convergence. This means the training time reduces by 50% with the training patch selection mechanism. Most of the redundant and unnecessary patches are not used to train the SaNet.

In the testing phase, our model is also fast to predict class labels for each image. Although we have two networks in our approach, the network structure of DeNet is very simple, with several fully connected layers and an LSTM. It only takes less than 6 ms to infer a class label for each image in the testing phase. Such fast online testing speed suggests that it could be applied in a routine clinical workflow.

L. Limitations

We are also aware that the current study has some limitations. Firstly, the images in BreakHis dataset are the cropped regions from raw histopathological data. Although our approach does not have to resize image in BreakHis dataset as done in previous work, it has not been evaluated on the whole slide dataset. Secondly, the size of the BreakHis dataset is relatively small with data from only 82 patients in total, combined with the small size of the testing dataset, this may mean that the results are biased. Thirdly, only one dataset is evaluated in this paper. How well our method can generalize to other datasets needs further study. In future work, we will

try to evaluate our method on more and larger whole slide datasets to test the generalization capability of our approach.

V. Conclusion

In this paper, we introduce a novel deep hybrid attention network, applied to breast cancer histopathological image classification. The hard attention mechanism in the network can automatically determine the useful regions from the images in BreakHis dataset, and thus does not have to resize the image for the network to avoid information loss. The selection mechanism in our framework is able to reduce training time by 50% when compared with the previous POMDP-based approach. We evaluate our approach on a public dataset, for which it achieves approximately 98% accuracy at four different magnifications.

References

[1] *Cancer Facts & Figures*, Amer. Cancer Soc., New York, NY, USA, 2008.

[2] H. D. Couture *et al.*, "Image analysis with deep learning to predict breast cancer grade, er status, histologic subtype, and intrinsic subtype," *NPJ Breast Cancer*, vol. 4, no. 1, p. 30, 2018.

[3] D. Bardou, K. Zhang, and S. M. Ahmad, "Classification of breast cancer based on histology images using convolutional neural networks," *IEEE Access*, vol. 6, pp. 24680–24693, 2018.

[4] T. Qaiser and N. M. Rajpoot, "Learning where to see: A novel attention model for automated immunohistochemical scoring," *IEEE Trans. Med. Imag.*, vol. 38, no. 11, pp. 2620–2631, Nov. 2019.

[5] F. A. Spanhol, L. S. Oliveira, P. R. Cavalin, C. Petitjean, and L. Heutte, "Deep features for breast cancer histopathological image classification," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 1868–1873.

[6] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast cancer multi-classification from histopathological images with structured deep learning model," *Sci. Rep.*, vol. 7, no. 1, p. 4172, 2017.

[7] M. Jannesari *et al.*, "Breast cancer histopathological image classification: A deep learning approach," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 2405–2412.

[8] A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin, "Deep convolutional neural networks for breast cancer histology image analysis," in *Proc. Int. Conf. Image Anal. Recognit.* Cham, Switzerland: Springer, 2018, pp. 737–744.

[9] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 2560–2567.

[10] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen, and Z. Liu, "Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1959–1970, Aug. 2019.

[11] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2092–2103, Sep. 2019.

[12] M. Tang, Z. Zhang, D. Cobzas, M. Jagersand, and J. L. Jaremko, "Segmentation-by-detection: A cascade network for volumetric medical image segmentation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1356–1359.

[13] B. Xu *et al.*, "Look, investigate, and classify: A deep hybrid attention method for breast cancer classification," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 914–918.

[14] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455–1462, Jul. 2016.

[15] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[17] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5532–5540.

[18] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 289–297.

[19] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[20] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.

[21] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019.

[22] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv: 1804.03999*. [Online]. Available: https://arxiv.org/abs/1804.03999

[23] Y. Zhang, B. Zhang, F. Coenen, J. Xiao, and W. Lu, "One-class Kernel subspace ensemble for medical image classification," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, p. 17, 2014.

[24] P. Wang, X. Hu, Y. Li, Q. Liu, and X. Zhu, "Automatic cell nuclei segmentation and classification of breast cancer histopathology images," *Signal Process.*, vol. 122, pp. 1–13, May 2016.

[25] C. Bahlmann, A. Patel, J. Johnson, J. Ni, A. Chekkoury, and P. Khurd, "Automated detection of diagnostically relevant regions in H&E stained digital pathology slides," *Proc. SPIE*, vol. 8315, Feb. 2012, Art. no. 831504.

[26] J. Liu, B. Xu, L. Shen, J. Garibaldi, and G. Qiu, "HEp-2 cell classification based on a deep autoencoding-classification convolutional neural network," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 1019–1023.

[27] Y. Song, J. J. Zou, H. Chang, and W. Cai, "Adapting Fisher vectors for histopathology image classification," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 600–603.

[28] Y. Song, H. Chang, H. Huang, and W. Cai, "Supervised intra-embedding of Fisher vectors for histopathology image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 99–106.

[29] V. Gupta and A. Bhavsar, "Breast cancer histopathological image classification: Is magnification important?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 769–776.

[30] V. Gupta and A. Bhavsar, "Sequential modeling of deep features for breast cancer histopathological image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2335-1–2335-7.

[31] V. Gupta and A. Bhavsar, "Partially-independent framework for breast cancer histopathological image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR)*, Jun. 2019, pp. 1–8.

[32] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2342–2350.

[33] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 229–256, 1992.

[34] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 1057–1063.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv: 1409.1556*. [Online]. Available: https://org/arxiv.abs/1409.1556

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Aug. 2016, pp. 770–778.

[37] Y. Song *et al.*, "Feature learning with component selective encoding for histopathology image classification," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 257–260.

[38] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3460–3469.

[39] K. Das, S. Conjeti, A. G. Roy, J. Chatterjee, and D. Sheet, "Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 578–581.

[40] M. Nawaz, A. A. Sewissy, and T. H. A. Soliman, "Multi-class breast cancer classification using deep learning convolutional neural network," *Int. J. Adv. Comput. Sci. Appl*, vol. 9, no. 6, pp. 316–332, 2018.

[41] Y. Jiang, L. Chen, H. Zhang, and X. Xiao, "Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module," *PLoS ONE*, vol. 14, no. 3, 2019, Art. no. 0214587.