Breast Cancer Histopathological Images Segmentation and Classification using Vision Transformer

Anonymous CVPR 2021 submission

Paper ID ****

Abstract

Breast cancer is one of the wide-spread diseases in the world. Its detection at an early stage is crucial for increasing the chance of successful treatment of patients. However, pathologists often need to scan a large set of Whole Slide Images (WSIs) to identify the regions and the type of tissues, which is laborious task. There have been several deep learning based approaches to reduce the burden. Previous approaches made use of convolutional neural networks (CNNs) as feature extractor, which often struggle to generalize in different domains. Moreover, these approaches often takes sophisticated post-processing techniques and relies on patch-based technique treating each patch as an image. In this paper, we propose an end-to-end method that exploits the expressive capacity of Vision Transformer (ViT) and its variants. Our method can be put into three main phases. First, we build a classification model for breast cancer diagnosis. Second, we design a ViT based segmentation model for WSIs. Finally, we investigate the combination of these models for domain adaptation in other types of tissue segmentation. The preliminary experiment on BreakHis dataset has demonstrated the proposed method efficiently competes with the existing state-of-the-art results.

1. Introduction

Breast cancer [23] is one of the most wide spread types of cancer whose detection and categorization is not straightforward even for expert pathologists [19]. In fact, pathologists have to scan a large set of Whole Slide Images (WSI), which can be in the order of gigapixels to localize the regions of tumor as well as to identify the type of tumor. To ensure that there is no malpractice, this has to be done in several magnification levels[16]. Fortunately, due to recent advancements in deep learning specifically in pattern recognition[21], there has been a growing interest in applying deep learning to tissue segmentation and breast cancer classification using WSI[17, 24].

However, WSI are very large to segment with current deep learning methods without reducing the images' dimensions. Even though reducing the image size may have little effects on the cancer classification task (benign and malignant), detecting the region of interest through tissue type segmentation requires that images not be resized. To overcome such limitation, patch-based methods have been introduced[14] for both tissues semantic segmentation and cancer classification. The patch-based methods consist of dividing WSI and ground-truth segmentation mask into small chunks of images. Patch-based approaches have been broadly investigated[31] and are actually commonly used in large scale medical image segmentation. The extracted patches can be saved on the disk, but this can become impractical given the number of the images. Another challenge is that some patches do not have distinctive features and may require a good post-processing to achieve a desired outcome. Moreover, patches are treated independently without much regard to their global relationship in the whole image.

In this paper, we propose a new approach that exploits the current state-of-the-art Vision Transformer architecture to take advantage of WSI at once by considering each image as a sequence of patch tokens. Such an approach is fast to train for both cancer types classification and tissues type semantic segmentation with less effort in preprocessing and post-processing. Although similar architecture has been introduced in other medical images segmentation related work [5], to our knowledge, this is the first time this method is applied to WSI for breast cancer images segmentation and tissue classification. In most existing related work where Vision Transformer is used have been to exploit its pretrained feature extraction capability. We consider a similar approach as machine translation when dealing with semantic segmentation, where a sequence of images is given as input to produce a sequence of segmented masks.

The work of this paper can be divided into three parts. First, we propose Vision Transformer model for breast cancer classification. Second, we use similar base architecture

for tissue type segmentation. Finally, we use the semantic segmentation model as feature extractor for cancer classification and investigate further domain adaptation. The contribution and the novelty of this paper are summarized as follows:

- We propose the first Vision Transformer based cancer tissues segmentation and cancer types classification using WSI of breast. To our best knowledge, this approach has not yet been applied to WSI and for large scale image segmentation and classification.
- The proposed method require less postprocessing compared to the existing patch-based approaches.
- The proposed method is fast to train and achieve competitive results compared to existing complex convolution neural network model on classification tasks.

2. Related work

WSI Classification. Using WSI is one of the best ways to integrate deep learning in cancer diagnosis, specifically for breast cancer. In breast cancer diagnosis, the classification can be binary (benign or malignant) or multiclass. Hameed et al [13] used an ensemble of deep learning model such VGG16 and VGG19 to design a classifier model for non-carcinoma and carcinoma breast cancer histopathology images identification. However, this model is hard to train due to the complexity of VGG network. In the same vein of binary classification, Abdullah-Al Nahid et al. [1] employed CNN and LSTM among other features extraction methods to design a model that can accurately classify cancer type, benign or malignant. Similar approach has been proposed in [2] where they exploited the Inception recurrent CNN architecture to improve the performance on BreakHis dataset[4].

On the other hand, some studies included the sub-classes of cancer in a multiclass classification approach [8]. [22] proposed a comparison study of several deep learning approaches for breast cancer classification. Furthermore, [4] conducted a survey where they provided a performance analysis of existing state-of-the-art methods on BreakHis dataset. [30] is one of these best approaches for binary classification. They inserted squeeze and excitation module to a ResNet model and adjusted the optimization parameters to increase the classification performance. The existing methods are mostly based on patch-based approach where the challenges arise. One of the challenges is to distinguish informative patches from non-informative patches. It is also challenging to classify an image based on several sequences of patches. Additionally, dividing large images into patches and save to disk will triple the memory requirement (original, patches, and online execution). Our approach leverages all these challenges of existing method and can achieve competitive results. With the transformer architectures, we are still using patch-based approach but this method can take variable length of patches while extracting relevant image features with attention mechanism.

WSI Segmentation. There is a large body of work on WSI segmentation[26]. Most recent works on WSI segmentation focus on deep learning[6]. However, applying deep learning for WSI segmentation is a challenging task because of the large size of the images where resizing is not generally recommended for accuracy. Therefore, most existing deep learning approaches exploit patch-based based segmentation[18, 11]. In [10], active learning was used to enhance the performance of patch based WSIs segmentation. Similarly in [29], pretrained deep convolutional neural network (CNN) was used to improve the performance of patch-based segmentation of brain WSIs for tumor detection. To investigate the deep learning model generalization, Mahendra Khened et al.[15] proposed a deep learning framework for histological tissue segmentation. In their study, they explored different datasets including a breast cancer WSI dataset and evaluated the model's uncertainty and generalization performance against distributional shift.

Moreover, several studies such as [12] exclusively studies breast cancer WSI segmentation. In addition to WSI segmentation for cancer or disease detection, deep learning methods are being used for cancer classification from WSI[28], specifically for the breast cancer. Most existing WSI segmentation methods are patch-based, therefore they require more additional post processing processing to combine the patches. Furthermore, some patches can be non-informative. Those methods commonly use deep CNN architectures which can slow the training process and difficult to train on computation resource limited devices. Transformer architectures can extract meaningful features from natural images.

Transformer. Transformer architecture was first proposed by Vaswani et al.[25]. The essence of the architecture lies in the multi-headed self-attention (MHA) mechanism to learn the relationships among sequential tokens. MHA can be expressed as the following

$$\begin{aligned} &\text{MHA}(Q,K,V) = \text{Concat}(A_1,\ldots,A_h)W^O, \\ &\text{where } A_i := \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \\ &\qquad = \text{softmax}\left(\frac{(QW_i^Q)(KW_i^K)^\top}{\sqrt{d_k}}\right)VW_i^V, \end{aligned} \tag{1}$$

where Q, K, V are input query, key, value vectors, $W_i^Q, W_i^K \in \mathbb{R}^{d_{\text{model}},d_k}, W_i^V \in \mathbb{R}^{d_{\text{model}},d_v}$ are trainable query, key, value weights at the i-th attention head, d_{model}

the model dimension , h number of heads, d_k, d_v key and value dimension. By only using self-attention for processing sequential data, Transformer has shown state-of-the-art performance in numerous natural language processing tasks such as machine translation and sequence generation. Since its conception, Transformer architecture has been adapted to other domains.

Vision Transformer Vision Transformer (ViT)[7] closely adapts Transformer Architecture for natural language processing for image classification. The concept consists of dividing the whole image into small chunks analogous to the sequence of tokens in machine translation. There are a few modifications from the original Transformer architecture for image domains:

1. In order to circumvent the quadratic memory bottleneck, ViT takes in patches of images instead of pixels that are flattened then embedded via linear layer, essentially taking patches as tokens.

2. Because of the sequential nature, ViT can take higher resolution inputs. However, position embeddings need to be interpolated for finetuning. This has been shown to be emprically effective in the subsequent downstream tasks.

3. As in the pretrained language model BERT, a classification token is appended before the patch tokens.

Through extensive experiments, it has been proved that the ViT outperforms the existing state-of-the-art models in image classification tasks, thereby several variants of ViT have been proposed[27, 32]. Furthermore, ViT model can be a good feature extractor or encoder given that it has been trained on various type of datasets. Therefore, we apply this promising architecture to histological images segmentation and breast cancer classification.

3. Method

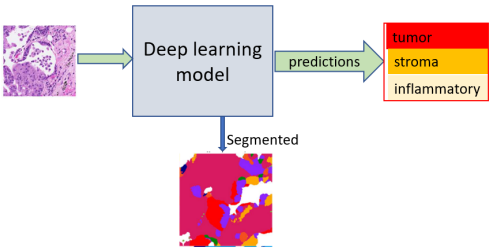


Figure 1: Block diagram of the model for WSIs segmentation and classification

Figure 1 depicts the block diagram of the proposed method. Our approach can be decomposed into two main steps: first, we build a classification model based on ViT for breast cancer detection, then we build a model for WSI segmentation. We then combine those models to produce a multi tasks model. Another option we have is to apply domain adaptation to another organ segmentation dataset.

3.1. Classification

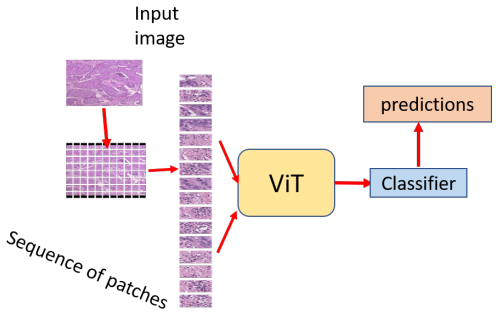


Figure 2: Block diagram of the model for WSI classification

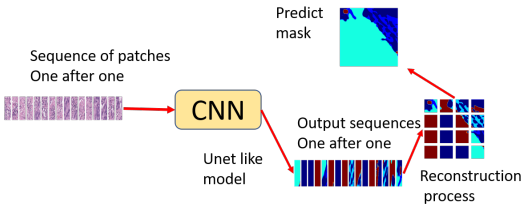
In this section we begin with the classification model. Our first goal is to use the ViT to accurately predict the class of breast cancer given a whole slide image as input. Previous works [14] decompose the images into patches then fit those discriminative and non discriminative patches to a CNN one by one or resize to a lower resolution. We resize to an appropriate resolution (not necessarily lower for CNN) then use all patches of each image at once with ViT. Given an image $X^{W\times H\times C}$ with spatial resolution $W \times H$, ViT firstly divides it into $N = \frac{WH}{P^2}$ patches where P is the patch size. This leads to each spatial dimension being a multiple of patch size. However, WSI do not always come in such manner. To deal with the design choice we propose to use rectangular patch size and also use padding strategy in case the sizes do not match. Although padding may have small effect on classification results, it is important for WSI segmentation. The new number of patches can be computed as in 2:

$$N = \left\lceil \frac{H}{P_1} \right\rceil \cdot \left\lceil \frac{W}{P_2} \right\rceil, \quad (2)$$

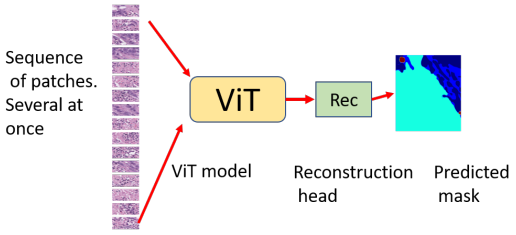
where P_1 and P_2 are the spatial dimensions of a patch. Figure 2 depicts the proposed architecture for cancer WSI classification.

3.2. Segmentation

If the image is too large we divide it into chunks of appropriate size then use those chunk as original image. Doing so we reduce the number on chunks to be treated sequentially compared to existing approaches which will



(a) Previous patch-based approach



(b) Proposed segmentation architecture

Figure 3: WSI segmentation model diagram

produce large number of chunks then treats them sequentially 3a. Most existing biomedical images segmentation approaches integrate UNet[20] or a large CNN architecture. Our method also have the advantage to allow the insertion of part of an UNet module at the output of the transformer if needed. Figure 3b depicts the proposed segmentation approach. The model can produce a sequence of images at once if the ground true is a sequence of images like in machine translation. Also, we can bypass the post processing part and produce accurate segmentation mask. The reconstruction head in the Figure 3b can be a set of convolution layers to produce the appropriate final shape for the output model.

3.3. Evaluation metric

To evaluate the proposed method, we use two different metrics for prediction and segmentation. The prediction level metric consists of: *F1*, *precision*, and the *sensitivity* as defined in 3

$$\begin{aligned} \text{F1} &= \frac{2TP}{2TP+FP+FN}, \\ \text{Sensitivity} &= \frac{TP}{TP+FN}, \\ \text{Precision} &= \frac{TP}{TP+FP}, \end{aligned} \tag{3}$$

where TP, FP and FN are true positive, false positive and false negative, respectively. For segmentation level metric, the intersection of Union metric: Jacard index or DICE's coefficient D defined in equation 4,

$$D(\hat{Y}, Y) = 2\frac{|\hat{Y} \cap Y|}{|\hat{Y} \cup Y|}, \tag{4}$$

where $\hat{Y}$ and Y are the predicted region from the model and the ground truth region of input image, respectively.

4. Preliminary experiments

At this stage of our work we evaluate the proposed model on breast cancer data compared to some recent works. The proposed classifier was trained using an RTX2080ti and 4 RTX TitanX GPUs with TensorFlow and PyTorch library. Most of the training was done on RTX2080ti desktop.

4.1. Data

Breast Cancer Histopathological Image Classification (BreakHis) dataset [4] was used to assess the performance. This dataset is composed of 9,109 microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). It contains 2,480 benign and 5,429 malignant samples (700×450 pixels, 3-channel RGB, 8-bit depth in each channel, PNG format). Benign and malignant subtypes 1 are also provided which makes this dataset an good choice for comparing models performance. According to the dataset, there are four histological distinct types of benign breast tumors (B): adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenona (TA) and four malignant breast tumors (M): carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC) and papillary carcinoma (PC).

Class	Type	Magnification 40X	100X	200X	400X	Total	Sum
B	A	114	113	111	106	444	2480
	F	253	260	264	237	1014	
	TA	109	121	108	115	453	
	PT	149	150	140	130	569	
M	DC	864	903	896	788	3451	5429
	LC	156	170	163	137	626	
	MC	205	222	196	169	792	
	PC	145	142	135	138	560	
Total		1995	2081	2013	1820	7909	7909

Table 1: BreakHis dataset description

4.2. Breast cancer classification using BreakHis dataset

We consider two approaches to evaluate our model which are magnification specific binary(MSB) classification and multi classes classification. For the binary classification, we only consider benign and malignant to be the classes. For multi classes classification, we only consider the subclasses. The results of the experiment for binary classification are showed in Table 2. In this experiment, for fair comparison we consider two state-of-the-art

methods[9, 3] With similar setting on the same dataset. Data augmentation techniques consisting to increase the number of samples where not used. Also, we do not use any pre-trained techniques.

Works	Year	Magnification 40X	100X	200X	400X	Mean
[9]	'19	97.90	96.88	96.88	96.88	97.13
Ours	'21	**98.36**	96.19	95.86	**97.98**	97.10
VGG[3]	'18	96.82	96.96	96.36	95.97	96.52
Ens[3]	'18	98.33	**97.12**	**97.85**	96.15	**97.36**

Table 2: Comparison of MSB classification results. VGG stands for VGG-16 architecture and Ens stands for Ensemble method proposed in [3]. Boldface means the maximum within the respective column.

In this preliminary experiment we trained the proposed model from scratch. For comparison, the BreakHis dataset is randomly split into 70% for training and 30% for testing. We only applied horizontal flip and normalization. All methods are within in the same range although those method used complex convolution architectures. Based on the results presented in [4], our approach achieves competitive compared to existing breast cancer classification method tested on BreakHis. However, further improvement are still needed. Therefore, we will perform further experiments, parameter tuning to outperform the state-of-the-art on BreakHis and also move to tissues segmentation. The results presented in this paper are preliminary results and the final results may different significantly. Therefore we have not yet include the evaluation metrics results.

4.3. Additional experiments with pretrained ViT

Some additional experiments were conducted with publicly available pretrained ViT model. This was to investigate the effect of pretraining on a drastically different image domains to lay basis for further investigation. For this model, the image was resized to 224-by-224 for preliminary experiment. However, this part will be modified in the coming end-to-end implementation. For experiment, a base model with patch size 16 and input size 224 was used. The experiments can be viewed here.

The results are provided Table 3. As expected, it was found that pretrained models can absolutely leverage information obtained during pretraining. Even without data augmentation, the model already achieves the state-of-the-art result. When coupled with data augmentation, the model performance increases at low magnification, leading to the mean accuracy across magnifications levels of 98.52, which is remarkable. The effect of higher input resolution for downstream task, however, still needs to be investigated.

Works	Magnification 40X	100X	200X	400X	Mean
PT	97.66	**98.72**	**98.01**	**98.35**	98.19
PT+AG	**99.00**	**98.72**	**98.01**	**98.35**	**98.52**

Table 3: Experiments on pretraining. All models were trained under the same hyperparameters. PT stands for pre-training and AG stands for augmentation.

5. Conclusion

In this stage of work we proposed a classification model for histological images using ViT architecture. The proposed method is fast to train and achieved good performance compared to previous works. In the coming stage, we will evaluate the model performance using the proposed metrics. Also, we will conduct more classification experiments to include the multi classes classification results. After that we will apply the proposed architecture for WSI segmentation. Finally we will investigate transfer learning and domain adaptation.

References

[1] Yinan Kong Abdullah-Al Nahid, Mohamad Ali Mehrabi. Histopathological breast cancer image classification by deep neural network techniques guided by local clustering. *BioMed Research International*, 2018:20, 2018. 2

[2] Md Zahangir Alom, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network, 2018. 2

[3] Dalal Bardou, Kun Zhang, and Sayed Mohammad Ahmad. Classification of breast cancer based on histology images using convolutional neural networks. *IEEE Access*, 6:24680–24693, 2018. 5

[4] Yassir Benhammou, Boujemâa Achchab, Francisco Herrera, and Siham Tabik. Breakhis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. *Neurocomputing*, 375:9–24, 2020. 2, 4, 5

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation, 2021. 1

[6] Neofytos Dimitriou, Ognjen Arandjelović, and Peter D. Caie. Deep learning for whole slide image analysis: An overview. *Frontiers in Medicine*, 6:264, 2019. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. 3

[8] Han Z. et al. Breast cancer multi-classification from histopathological images with structured deep learning model. *Sci Rep*, 7, 2017. 2

[9] Xie Juanying et al. Deep learning based analysis of histopathological images of breast cancer. *JFrontiers in genetics*, 10(80), 2019. 5

[10] Jonathan Folmsbee, Margaret Brandwein-Weber, and Scott Doyle. Whole slide semantic segmentation: large scale active learning for digital pathology. In John E. Tomaszewski and Aaron D. Ward, editors, *Medical Imaging 2021: Digital Pathology*, volume 11603, pages 83 – 93. International Society for Optics and Photonics, SPIE, 2021. 2

[11] Xiaohang Fu, Tong Liu, Zhaohan Xiong, Bruce H. Smaill, Martin K. Stiles, and Jichao Zhao. Segmentation of histological images and fibrosis identification with a convolutional neural network. *Computers in Biology and Medicine*, 98:147–158, Jul 2018. 2

[12] Zichao et al. Guo. A fast and refined cancer regions segmentation framework in whole-slide breast pathological images. *Sci Rep.*, 10(1), 2020. 2

[13] Zabit Hameed, Sofia Zahia, Begonya Garcia-Zapirain, José Javier Aguirre, and Ana María Vanegas. Breast cancer histopathology image classification using an ensemble of deep learning models. *Sensors*, 20(16), 2020. 2

[14] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2424–2433, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society. 1, 3

[15] Mahendra Khened, Avinash Kori, Haran Rajkumar, Balaji Srinivasan, and Ganapathy Krishnamurthi. A generalized deep learning framework for whole-slide image segmentation and analysis, 2020. 2

[16] Elizabeth A Krupinski, Allison A. Tillack, Lynne Richter, Jeffrey T. Henderson, Achyut K. Bhattacharyya, Katherine M. Scott, Anna R. Graham, Michael R. Descour, John R. Davis, and Ronald S. Weinstein. Eye-movement study and human performance using telepathology virtual slides. implications for medical education and differences with experience. *Human Pathology*, 37(12):1543–1556, 2006. 1

[17] Tao Lei, Risheng Wang, Yong Wan, Bingtao Zhang, Hongying Meng, and Asoke K. Nandi. Medical image segmentation using deep learning: A survey, 2020. 1

[18] Blanca Maria Priego-Torres, Daniel Sanchez-Morillo, Miguel Angel Fernandez-Granero, and Marcial Garcia-Rojo. Automatic segmentation of whole-slide h&e stained breast histopathology images using a deep convolutional neural network architecture. *Expert Systems with Applications*, 151:113387, 2020. 2

[19] Rhys Thomas & Darren Treanor Rebecca Randell, Roy A. Ruddle. Diagnosis at the microscope: a workplace study of histopathology. *Cognition, Technology & Work*, 14:319–335, 2014. 1

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 4

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. 1

[22] F. Shahidi, S. Mohd Daud, H. Abas, N. A. Ahmad, and N. Maarop. Breast cancer classification using deep learning approaches and histopathology image: A comparison study. *IEEE Access*, 8:187531–187552, 2020. 2

[23] Fabio A. Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016. 1

[24] Chetan L. Srinidhi, Ozan Ciga, and Anne L. Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021. 1

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 2

[26] Quoc Dang et al. Vu. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology*, 7, 2019. 2

[27] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers, 2021. 3

[28] Yan Xu, Jun-Yan Zhu, Eric I-Chao Chang, Maode Lai, and Zhuowen Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical Image Analysis*, 18(3):591–604, 2014. 2

[29] Wang LB. et al. Xu Y., Jia Z. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinformatics*, 18, 2017. 2

[30] Jiang Yun, Chen Li, Zhang Hai, and Xiao Xiao. Breast cancer histopathological image classification using convolutional neural networks with small se-resnet module. *PLOS ONE*, 14(3):1–21, 03 2019. 2

[31] C. Zhang, Y. Song, D. Zhang, S. Liu, M. Chen, and W. Cai. Whole slide image classification via iterative patch labelling. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1408–1412, 2018. 1

[32] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer, 2021. 3