Segmentation of Whole-Slide Images with Context-Aware Vision Transformers

Michal Franczel*
Supervised by: Lukáš Hudec†

Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Bratislava / Slovak Republic

Abstract

Histological examination is a crucial component of breast cancer diagnostics. Analysis of whole-slide images (WSI) is a time-consuming process due to their hierarchical nature and size, resulting both in slower diagnostics and a lack of annotations. Recent advances in vision transformers have demonstrated potential within the field of computer vision. However, their properties with hierarchical gigapixel images, where contextual information is crucial, remains underexplored. In this paper, we propose a solution employing semi-supervised learning based on a self-supervised pretraining and supervised fine-tuning paradigm, utilizing these advancements. Our approach modifies vision transformer encoders within the segmentation network to incorporate contextual information from lower magnification levels through late feature fusion. The multi-scale model variant outperforms its single-scale counterpart, improving the dice score by 6.2%. Furthermore, we examine the properties of features learned by masked image modeling (MIM) and establish that vision transformers trained with MIM can effectively learn morphological phenotypes from unlabeled histopathological images, thereby validating its use as a pretraining technique in this domain.

Keywords: Whole-Slide Images, Breast Cancer, Deep Learning, Segmentation, Semi-Supervised Learning, Vision Transformers, Medical Imaging

1 Introduction

Breast cancer is one of the leading preventable causes of death, accounting for more than 13 percent of all new cancer cases and 28.7 percent of all cancer discoveries in women in the European Union as of 2020. While the number of new cases has increased over time, the number of deaths has decreased [9]. This can be explained not just by increasing the quality of treatment but also by increasing the rate of early disease diagnosis [7]. Histological analysis is a vital, yet time-consuming and difficult, component of breast cancer diagnosis. As part of the pathological examination of the breast, a biopsy is performed. The extracted tissue is sliced, and the slice is then stained most commonly with hematoxylin and eosin (H&E). Subsequently, it is placed on a glass slide, which is scanned with a motorized microscope. Slides are scanned at multiple magnification levels, resulting in z-stacks. This enables pathologists to switch between these magnification levels, simulating classical microscopy. Pathologists analyze a variety of structures, not only at the level of cellular morphology but also at the level of larger breast structures, utilizing both contextual information from lower levels and detailed information from higher levels of magnification.

Deep learning has been an essential part of computer vision, and convolutional models have been successfully applied to the field of computational pathology. However, in recent years, transformer-based vision models have gained prominence, obtaining state-of-the-art performance across numerous general vision tasks. The properties of these transformers have, however, not yet been thoroughly examined within the field of histopathology, especially when dealing with segmentation of hierarchical images. Due to their size, Whole-Slide Images (WSI) must be split into smaller patches. These patches, at the highest magnification level, may not contain the necessary information for segmentation models to make accurate predictions, as they lack coarser-grained tissue features and their spatial organization. Additionally, the complex hierarchical nature of WSI results in a lack of annotated datasets in terms of both quantity and quality, especially when dealing with pixel-level annotations.

In this work, we first assess the utility of vision transformers on the BCSS dataset and then propose modifications to the segmentation network so that it uses multiple magnifications and passes contextual information in a top-down manner. We split methods of feature fusion into three categories: early, intermediate, and late fusion. Prior work on convolutional encoders used variations of early and late feature fusion with linear and LSTM layers. We examine early and late feature fusion with the use of

*xfranczel@stuba.sk
†xhudecl@stuba.sk

the Swin Transformer [8] encoder and a modified Upernet [15] architecture, using both linear layers and cross-attention mechanisms. Additionally, we expand these two types of fusion through intermediate fusion, merging features between individual blocks of the encoder. We find that vision transformers, even though they lack the inherent biases of convolutional networks, can achieve similar accuracy within the domain of histopathology. With the introduction of late feature fusion, we surpass the accuracy of single-scale single architectures, increasing the dice score by 6.2%.

Additionally, to address the issue of the lack of annotations, we explore the use of self-supervised pretraining methods based on masked image modeling (MIM) utilizing the BRACS [3] dataset. Through qualitative analysis of learned features, we find that transformer-based models pretrained with MIM on the BRACS dataset can learn useful representations of various types of tissues, confirming that it can be used as a pretraining step within the domain of histopathology.

2 Related Work

The segmentation task can be interpreted either as a pixel-wise segmentation or a patch-wise segmentation, where an image is split into patches, which are then classified and combined to create a coarse segmentation mask. Numerous approaches to pixel-wise segmentation have been proposed, with the most prevalent being the convolutional U-Net [12], featuring an encoder-decoder architecture with a bottleneck and skip connections. Variants of this architecture have also emerged, including U-Net++ [18], which employs a densely connected decoder subnetworks, and R2U-Net [1], which incorporates recurrent modules within both its encoder and decoder stages.

These architectures, frequently employed in cell and organ segmentation tasks, do not take advantage of the hierarchical structure of whole-slide images. To address this limitation, a number of context-aware methods have been introduced for the classification and segmentation of histopathological images. Sirinukunwattana et al. [13] studied the impact of providing contextual information to the prediction algorithm. They approached the problem of image segmentation as a patch-level classification and compared three types of architectures: single-scale architecture, which operates at a single image resolution; early fusion, which fuses information from multiple resolutions before passing it through a neural network; and late fusion, which uses separate networks for different magnifications and combines the output to make a prediction. Out of the three groups described, a single-scale design performed significantly worse than architectures that used contextual information. Feng et al. [6] proposed an end-to-end framework that generates predictions at multiple magnification levels and combines them using a voting process, adopting the late fusion approach. This approach was also used by the multi-scale classification model proposed by Wetteland et al. [14] to classify small patches, combining them into segmentation mask of an entire WSI. One major advantage of the late fusion approach is its ability to utilize contextual information from multiple resolutions, enhancing prediction accuracy. However, the main disadvantage is the increased computational complexity compared to single-scale architectures.

Chen et al. [5] proposed the Hierarchical Image Pyramid Transformer (HIPT), a three-stage architecture that performs bottom-up aggregation for slide-level representation, akin to hierarchical attention networks in long document modeling. The model allows for self-supervised pretraining methods to pretrain each aggregation layer separately, which can then be fine-tuned with slide-level labels for cancer subtyping and survival prediction tasks in the TCGA. Since the attention is computed only within local windows, learning long-range dependencies is tractable. Even though this method of bottom-up aggregation is not useful for image segmentation, it may be useful as a pretraining step.

3 Data

Breast Cancer Semantic Segmentation (BCSS) [2] dataset contains regions of interest derived from 151 WSIs stained with H&E, collected from histologically confirmed cases of breast cancer. Pathologists graded regions from 21 difficult slides that were annotated by trained non-pathologist research participants. Masks, pixel-level annotations with 21 classes, were the resulting annotations. Within the scope of our work, we opted for the use of the modified version of this dataset with labels reduced according to TIGER Challenge.

BReAst Carcinoma Subtyping (BRACS) [3] dataset is a breast carcinoma subtyping dataset containing 547 H&E-stained whole-slide images and 4539 extracted regions of interest from these WSIs. Both WSIs and ROIs were annotated with lesion categories by the consensus of three pathologists. Benign, malignant, and atypical lesions are further subtyped into seven distinct categories. Even though this dataset is one of the largest in its category, it does not include annotations at the pixel level. However, it can be used for unsupervised training.

4 Method

One of our objectives is to make use of current developments in transformer-based models. We decided to exploit current breakthroughs in semi-supervised learning, focusing on the paradigm of self-supervised pretraining and fully-supervised finetuning, as training these transformer models requires vast quantities of data. Thus, as shown on Figure 1, training is divided into two stages, with each of these stages being performed at the slide level using a

patch generator. This generator generates a tissue mask and patch stacks for the entire WSI. The resulting data is subsequently used for training. Within the first stage of our experiments, we compare various architectures that either utilize input from a single magnification level or use custom architectures that make use of three separate magnifications. In the second stage of our experiments, we focus on self-supervised pretraining with the BRACS dataset. Since the dataset contains 547 WSIs, training is difficult given the available computational resources. Therefore, we have chosen to sample WSIs, and from these images, we have chosen to sample 150,000 patches. In this phase, we will evaluate method of masked image modeling.

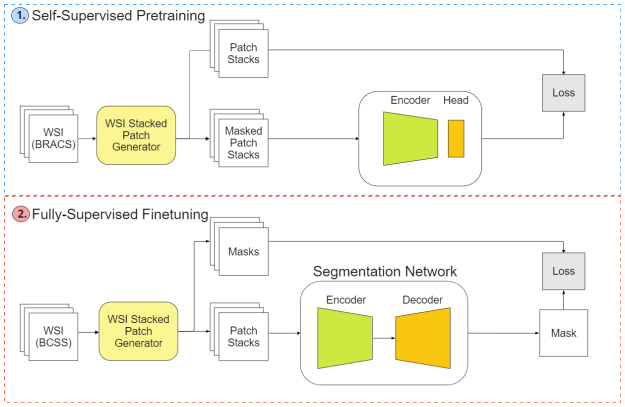

Figure 1: Overall two-stage architecture of proposed framework with self-supervised pretraining using MIM and fully-supervised context-aware finetuning

4.1 Data Preparation

Upon reviewing slides included in the dataset, we have discovered that a large part of all WSIs consist of background material without any histopathological relevance. That is why, firstly, a tissue mask is constructed employing simple thresholding and morphological operations. As some of the slides were labeled and contained artifacts, we created a mask for artifacts we observed by applying thresholding to the converted LAB image and deleting them from the tissue mask.

Patches are generated based on the pixels per micron (PPM) parameter of the slide, so that the generator can be used on varying datasets. Patches at the greatest magnification level containing tissue proportions below the threshold are discarded. When dealing with the BCSS dataset, patches with masks that include background levels above threshold in their annotation at the greatest magnification level are eliminated. When multiscale patch stacks are required, the location of the patch at the highest magnification level is determined first, and then the locations of patches at lower magnification levels are calculated, with higher magnification being at the central position. Stain normalization is the final phase in the preparation process, and we have decided to use the Macenko method of normalization [10]. This method is frequently employed as a preprocessing step, as it estimates hematoxilin and eosin concentrations from color space distributions and normalizes input images based on these concentrations, given some target image. Preprocessing flow is illustrated on Figure 2.

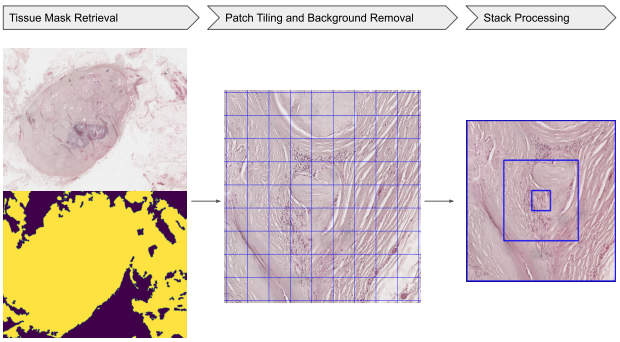

Figure 2: Three stages of data preparation: tissue mask retrieval with tresholding, patch tiling and background removal, and the addition of contextual patches from lower magnifications

4.2 Training Configuration

As per established and recommended training parameters, the final training configuration for fully supervised training uses the ReduceLRonPlaeau scheduler for convolutional networks and the cosine scheduler for transformer-based networks. As for the optimizer, AdamW is used, with betas set to 0.90 and 0.999, epsilon 1e-8, and a base learning rate of 5e-4. Since we have observed that this dataset is unbalanced and comprises disproportionately greater stroma and tumor types, we used Dice Cross-Entropy Loss, which includes squared versions of targets and predictions in the denominator,

$$\text{Loss} = \left(1 - \frac{2 * \sum_{c=1}^{C} p_c \hat{p}_c}{\sum_{c=1}^{C} p_c^2 + \sum_{c=1}^{C} \hat{p}_c^2}\right) - 0.5 * \sum_{c=1}^{C} p_c \log \hat{p}_c$$

where, C is the number of classes or categories, p_c is the true probability of class c, and $\hat{p}_c$ is the predicted probability of class c. Since the background class label in this dataset reflects unannotated regions and offers no semantic relevance in terms of training, it is not included in the calculation of loss.

To improve model resilience and reduce susceptibility to color pertubations, we employed augmentations. Vertical and horizontal flips, random rotations, and Contrast Limited Adaptive Histogram Equalization (CLAHE), random brightness, and contrast were employed. Additionally, we attempted to address the issue of class imbalance by oversampling and undersampling patches based on the classes present within its segmentation mask.

As for patch extraction, we used patch size 224 and overlap 112, as images of this scale contained sufficiently complex structures both within images and masks. Using our patch extractor, we used scale 1 as an input for single scale experiments and 1, 4, and 8 for multi-scale experiments. Within the scope of this work, the term *scale* refers to the downsampling factor relative to the highest level of magnification.

In the case of fully supervised training, the BCSS dataset was divided into 133 training samples and 18 validation samples. To prevent training on validation data that can be caused by overlap, patches were generated after this partitioning. During the training and validation process, metrics were computed patch-wise.

4.3 Single-Scale Architectures

First, we experimented with architectures based on U-Net [12] in order to establish a baseline and test configurations for data generation, as well as the properties of transformer-based networks and their usefulness within the domain of histopathology. For these U-Net-based designs, both encoder and decoder blocks based on transformers and convolutional blocks were utilized.

After these experiments, we focused on more complex and recent architectures, which generally perform better in multi-class semantic segmentation, with one of them being SegFormer [16] and the other being Upernet. Experiments on the SegFormer architecture were conducted with only small deviations from the original publication. The Upernet architecture was changed from multi-task to single-task with a Swin Transformer backbone. Backbone, which we used as an encoder for Upernet, remained the same. All the previously mentioned single-scale architectures maintain their original parameters, with alterations limited to their training parameters and minor implementation-related deviations. The transformer variants of these architectures employ the base size of the Swin encoder.

4.4 Multi-Scale Architectures

After the first iteration of experiments with a single scale, we focus on experiments combining multiple scales. Vision transformers have some different properties than convolutional neural networks, with the most important example being that they lack their intrinsic biases. Additionally, features of these two methods vary significantly [11], with global features being present at much earlier network stages. Simultaneously, various new layers and architectural elements were introduced in vision transformers and transformers in general, some of which do not have their counterparts within convolutional architectures. We try to improve the prediction accuracy of single-scale models evaluated in previous iterations using various methods, which we have divided into three categories:

1. *Early fusion*, where a single encoder is used and images from three different scales are combined before the first encoder stages
2. *Intermediate fusion*, where either three encoder branches are used and features are combined between stages from top to bottom, or a single encoder is used sequentially
3. *Late fusion*, where three images are passed through branches separately and fusion is performed on features passed before passing them to the encoder

We chose the Upernet-based architecture for these experiments involving multiple magnifications since it performed significantly better than other segmentation networks.

4.4.1 Early Fusion

The first method involves using a single encoder and merging its input before its first stage, concatenating channel dimensions. The number of blocks used was the same as for the single-scale encoder, but we increased the number of heads in the first two stages to 9 to match the complexity of the network to the increased complexity of the input.

Our second method for early fusion, Upernet T3 is comprised of three encoder blocks, which process input triplets sequentially in a top-down manner, passing contextual information to higher magnifications. The patch at the lowest magnification level is processed by the first encoder. The input for the first stage of the next encoder is the second magnification level, and the first stage of this encoder produces a feature map that is prepended to the feature maps produced by the first encoder. This information is then fed into the feature pyramid network (FPN), which produces a single, combined feature map. This map is then fed into the remaining stages of the second encoder. After that, the same operation is carried out with the second and third encoders, respectively. The feature maps produced by the third encoder are subjected to processing with pyramid pooling module (PPM) and FPN Fuse blocks, which ultimately produce a segmentation mask.

4.4.2 Intermediate Fusion

The first model architecture, which we implemented with respect to intermediate fusion, marked as Upernet T6, was intermediate fusion with the use of cross-attention, visualized on Figure 3. We used three separate Swin encoders. Images taken at different scales are passed through encoder stages, with each stage being followed by a cross-attention stage. The cross-attention stage is composed of two cross-attention blocks followed by layer normalization, where the first cross-attention blocks takes an input low smallest magnification as context and intermediate magnification as an input and the second one takes highest magnification as an input and result of previous cross-attention as context. The intuition behind this idea was

that, using purely attention-based mechanisms, we could pass information between the encoder stages of these three branches in a top-down manner.

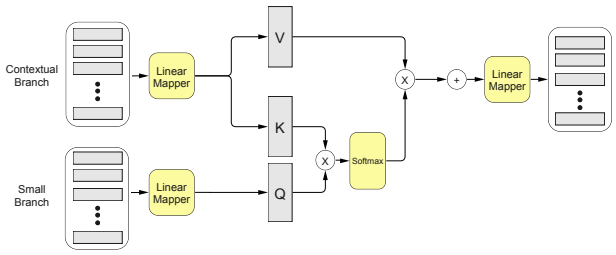

Figure 3: Cross-attention module

The second architecture utilizing intermediate fusion, with a designated label of Upernet T7, functioned in a similar manner, but instead of cross-attention, we fused class (CLS) token with patch tokens of higher magnification using convolutional layers. First, since Swin does not contain an explicit learnable CLS token, we compute it using an embedding layer, which takes input patch tokens as input, passes them through layer normalization followed by adaptive average pooling and a linear layer, producing a single class token representing all patches combined. This token is concatenated with patch tokens from higher magnification, and dimensions are reduced back to their original size using point-wise convolution. This way, we attempted to fuse features between three encoder branches by passing an aggrieved CLS token to lower magnifications.

4.4.3 Late Fusion

As for late fusion, we experimented with two methods as well. Our first method, named Upernet T2, was composed of three swin encoders, each used for different magnification level, all of them of same size. Since there are three encoder branches, the outputs of these stages need to be merged. For this, three outputs are first rearranged so that the height and width dimensions are reshaped into a single dimension, representing all tokens. Then we concatenate these tokens and pass them through an MLP block with the GELU activation function, which reduces their number to their original number. Finally, they are rearranged back to their original shape, and the resulting feature map is passed through the same PPM and FPN Fuse blocks. This architecture is shown on Figure 4.

The second method, named Upernet T5, utilizes feature merging instead of linear layers with a single cross-attention block, merging feature maps that are outputs of the last encoder stages. First, images from three selected scales are passed through all three encoders. Subsequently, the feature maps from the last encoder stages are passed through a self-attention block. Cross-attention is done twice, with the objective of passing contextual information from the lowest to the highest magnification level. The feature map produced by the second cross-attention module is then passed through the FPN Fuse block, result of which is then passed through the PPM block together with other features from the lowest magnification level. We hypothesized that the application of the self-attention mechanism in this manner may prove useful for passing high-level features at lower magnifications.

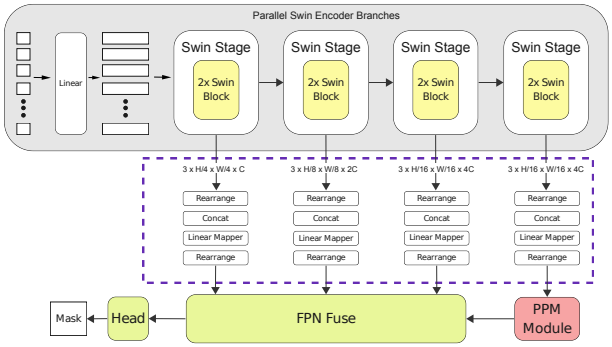

Figure 4: Architecture of Upernet T2 with three encoder branches and late future fusion

4.5 Masked Image Modeling

The effects of masked image modeling within the domain of histopathology have not yet been thoroughly studied. Thus, we focus on self-supervised training using masked-image modeling with iBot [17], which obtained state-of-the-art performance on various vision tasks outside of the medical domain.

iBot, similarly to DINO [4], employs two views created by augmenting the input image. Since it was not originally trained on medical images, we changed several parameters of these augmentations in order to accommodate the medical domain. iBot first applies local and global transforms, which produce local and global crops from the input image. After a visual evaluation of augmented crops, we changed the range of global crops to be between 0.7 and 1, from the original 0.14 and 1, and the range of local crops to be between 0.2 and 0.4, increasing the original values of 0.05 and 0.4, since such small patches did not contain sufficient information. Additionally, we removed color jitter and random grayscale from both local and global transforms and reduced the probability of solarization and Gaussian blur to 0.1. Experiments were done on ViT backbones on three different scales. For highest magnification, we used patch size 16, since we found it generally delivered better self-supervised results than with patch size 8. On intermediate and lowest magnification, we used patch size of 4, since we found, that smaller patch sizes on lower magnification levels provided necessary increase in resolution of attention maps.

5 Results

Within our single-scale experiments, we have found convolutional U-Net to be worse performing in comparison with transformer-based U-Net with respect to dice score. However, it obtained better IoU and per-class accuracy. Thus, we evaluated, that Swin-based U-Net performs on par with its convolutional counterpart, but with better inference and training speeds. Even though Segformer is a model with high throughput and efficiency, it performed significantly worse than Upernet, which outperformed all other evaluated models, as is shown within Table 1.

Table 1: Quantitative single-scale model evaluation

Models	mIoU	mDice	Per-Class Acc
Conv U-Net	29.44	36.57	36.27
Swin U-Net	28.58	36.91	35.72
Segformer	36.80	44.15	43.73
Upernet	46.03	56.05	55.1

Similarily to single-scale models, we have obtained several interesting results with multi-scale models. First, attention-based feature merging performs better when implemented as intermediate feature fusion rather than late fusion. Second, it was overcome by a method combining linear layers with late fusion. However, despite the fact that Upernet T6 with three encoder stages performed worse than Upernet T2 of the same size, its counterpart with a single encoder outperformed Upernet T6 of the same size. Thirdly, both methods of early fusion performed worse than their late and intermediate counterparts. Lastly and most importantly, all three best performing models outperformed their best performing single-scale counterpart. This result is attributable to the use of multiple contextual magnifications, which allowed the model to be trained on a specific dataset. Table 2 displays the respective results of the five architectures with the highest performance.

Table 2: Quantitative multi-scale model evaluation

Models	mIoU	mDice	Per-Class Acc
Upernet T3	40.8	47.92	48.24
Upernet T2*	46.12	54.67	53.96
Upernet T6*	48.15	56.52	55.68
Upernet T6	49.42	58.34	57.96
Upernet T2	52.13	62.25	61.51

* Single encoder used sequentially instead of three parallel encoders

5.1 Masked Image Modeling

Our qualitative evaluation involved the analysis of attention maps within the last transformer block. We used three different scales to measure how well these attentions worked on all three transformers that were trained with iBot. These were the same scales that were previously used for multi-scale experiments: 1, 4, and 8.

We started our experimentation at the highest magnification. Since the pretrained model was focusing only on white regions and ignoring regions with tissue compartments, we found the results to be underwhelming after training with patch size 8 and examining the attention maps of the model. However, when we increased the patch size to 16, we noticed that the attention heads began to concentrate more on the different kinds of tissue and small structures within the images, particularly cells. Following a qualitative analysis of attention, we discovered that heads 4, 7, and 12 acquired the ability to tell surrounding tissue from cells, which is particularly evident in images of tissue devoid of bubbles or other white regions. Within Figure 5 are visualized attention maps from heads 4, 3 and 2 of the last block of the encoder network, attending cells, stroma and fatty tissue, respectively.

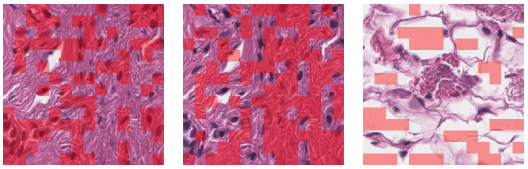

Figure 5: Visualized attention maps of cells (left), stroma (center) and fatty tissue (right) from the last encoder block with the highest magnification used

Secondly, we trained the same model with a scale of 4, with the model patch size set to 8. Even though the model did not focus on cellular structures as much as with higher magnification, upon analyzing various tissue types, we observed that it rather focused on differentiating between tissue compartments. Within fatty tissue, we can see that the attention map of heads 0, 1, 3, and 9 focused on membranes, and heads 2, 4, 5, 7, and 8 focused more on the fat itself, which resides within these membranes. Structures recognized by head 6 were not as apparent. However, it seems like model focused more on darker structures, including cells and darker tissue material where membranes meet. Within the darker patches that do not contain fat, we observed that attention within heads 0, 1, 3, 6 and 9 focused on stroma tissue connecting various other compartments, and attention within head 11 focused on darker regions of an image. On Figure 6, we can observe attention maps attending connective tissue and darker regions within fatty patches, as well as darker regions and stroma within patches containing tissue.

Finally, we trained a Swin transformer, which could be utilized for additional fine-tuning experiments, following the same approach as with ViT. We processed randomly sampled images and extracted features from the final layer of the encoder network. After applying T-SNE dimensionality reduction, we clustered the points using K-Means

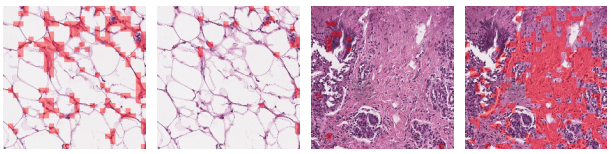

Figure 6: Visualized attention maps of connective tissue, dark regions and stroma from the last encoder block with patches from lower level of magnification

and visualized two components in Figure 7, with images representing the cluster centers. Figure 8 presents points accompanied by their respective images. The formation of clusters containing visually similar images is observed, with the primary basis for similarity being their visual characteristics.

6 Conclusions

First, we focused on the development of a full data processing pipeline that prepares whole-slide images for either training or analysis. Following a preliminary analysis of available datasets, we focused on evaluating single-scale models on the BCSS dataset and comparing convolutional networks with transformer-based networks, which have gained popularity in recent years. For the subsequent experiments involving multiple contextual magnifications, we chose the top-performing architecture based on an evaluation of these models. Experimenting with various variations of Upernet, we discovered that our multi-scale modification, Upernet T2, which is based on the late fusion of features between three backbones, outperforms its single-scale counterpart. We conclude, based on these results, that contextual usage improves the performance of the model.

Finally, we examined the effects of self-supervised learning with masked image modeling and its applicability in the medical field. After analyzing vision transformers pretrained with iBot, we conclude that masked image modeling is applicable to this domain and that models trained with masked image modeling may prove useful in future experiments.

Further research is required in the area of self-supervised pretraining. We found that this pretraining method is not appropriate for lower magnifications, but we hypothesize that this pretraining process could be generalized to multi-scale image processing by changing the training process so that all three encoders would learn representations together rather than separately. This requires changing the iBot head to accommodate the use of three backbones and changing the loss function so that all three encoders can be trained. Furthermore, we used only small models and a limited number of pretraining samples, since our evaluation of the self-supervised algorithm concentrated primarily on the analysis of feature maps. In order to evaluate larger Swin models, which should perform better with self-supervised pretraining, more training samples are required.

Weakly supervised learning could be incorporated into the fully supervised stage of our training pipeline. Since the BCSS dataset contains whole-slide images, but only ROI-level annotation, it is possible to use surrounding regions progressively as a method of weakly-supervised learning. As the segmentation model is trained from the labeled data during training, it can also produce segmentation masks for nearby regions, resulting in their weak labels, which can subsequently be added to the training set. Since the areas around labeled patches have some commonalities, we hypothesize that the model could use this shared information to generate reliable weak predictions. However, as training goes on, the factor of expansion must decrease because patches farther away do not benefit from proximity to the labeled patches.

References

[1] Md. Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *CoRR*, abs/1802.06955, 2018.

[2] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai A T Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem S E Salem, Ahmed F Ismail, Anas M Saad, Joumana Ahmed, Maha A T Elsebaie, Mustafijur Rahman, Inas A Ruhban, Nada M Elgazar, Yahya Alagha, Mohamed H Osman, Ahmed M Alhusseiny, Mariam M Khalaf, Abo-Alela F Younes, Ali Abdulkarim, Duaa M Younes, Ahmed M Gadallah, Ahmad M Elkashash, Salma Y Fala, Basma M Zaki, Jonathan Beezley, Deepak R Chittajallu, David Manthey, David A Gutman, and Lee A D Cooper. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18):3461–3467, 02 2019.

[3] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, Maria Gabrani, Florinda Feroce, and Maria Frucci. BRACS: A Dataset for BReAst Carcinoma Subtyping in H
amp;E Histology Images. *Database*, 2022, 10 2022. baac093.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294, 2021.

[5] Jia-Mei Chen, Yan Li, Jun Xu, Lei Gong, Lin-Wei Wang, Wen-Lou Liu, and Juan Liu. Computer-

aided prognosis on breast cancer with hematoxylin and eosin histopathology images: A review. *Tumour Biol.*, 39(3):1010428317694550, March 2017.

[6] Yanbo Feng, Adel Hafiane, and Hélène Laurent. A deep learning based multiscale approach to segment cancer area in liver whole slide image, 2020.

[7] S. W. Fletcher, W. Black, R. Harris, B. K. Rimer, and S. Shapiro. Report of the International Workshop on Screening for Breast Cancer. *J Natl Cancer Inst*, 85(20):1644–1656, Oct 1993.

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021.

[9] Bettio M, Negrao De Carvalho R, Dimitrova N, Dyba TA, Giusti F, Martos Jimenez MDC, Neamtiu L, Nicholson N, Randi G, Rooney R, Voti L, Crocetti E, and Voithenberg L. Dataset collection: European cancer information system. 2023.

[10] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110, 2009.

[11] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *CoRR*, abs/2108.08810, 2021.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[13] Korsuk Sirinukunwattana, Nasullah Khalid Alham, Clare Verrill, and Jens Rittscher. Improving whole slide segmentation through visual context - a systematic study, 2018.

[14] Rune Wetteland, Kjersti Engan, Trygve Eftestøl, Vebjørn Kvikstad, and Emiel A. M. Janssen. A multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides. *Technology in Cancer Research & Treatment*, 19:1533033820946787, 2020.

[15] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding, 2018.

[16] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *CoRR*, abs/2105.15203, 2021.

[17] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan L. Yuille, and Tao Kong. ibot: Image BERT pre-training with online tokenizer. *CoRR*, abs/2111.07832, 2021.

[18] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *CoRR*, abs/1807.10165, 2018.

7 Appendix

Figure 7: T-SNE projection with K-Means clustering and patches from cluster centers

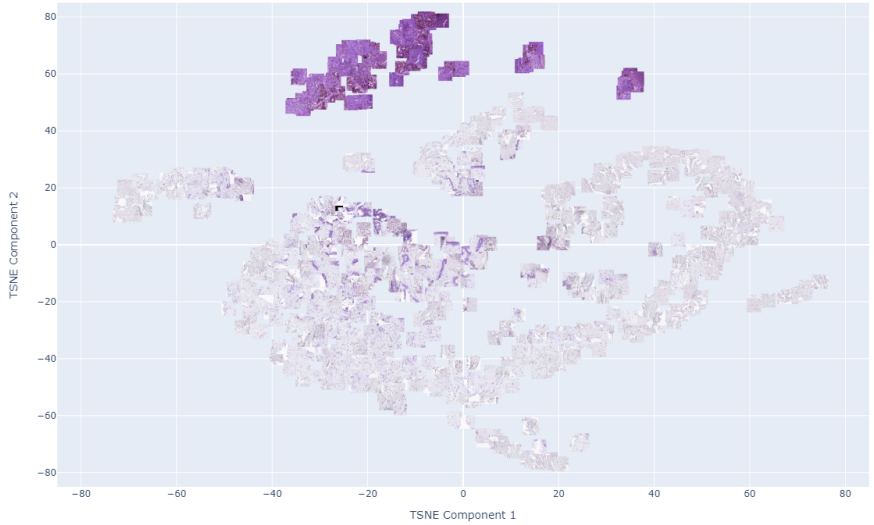

Figure 8: T-SNE projection of features from 10,000 images, with images as points