Enhancing Cell Detection with Transformer-Based Architectures in Multi-Level Magnification Classification for Computational Pathology

Jarl Sondre Bringslid Sæther[0009−0002−7971−2213] ⋆, Bendik Gjermundrød Holter[0009−0000−9334−6315] ⋆, Frank Lindseth, and Gabriel Hanssen Kiss

Norwegian University of Science and Technology,
Høgskoleringen 1, 7034 Trondheim, Norway
{jssaethe, bendikgh, frankl, gabriel.kiss}@ntnu.no

Abstract. Cell detection and classification are important tasks in aiding patient prognosis and treatment planning in Computational Pathology (CPATH). Pathologists usually consider different levels of magnification when making diagnoses. Inspired by this, recent methods in Machine Learning (ML) have been proposed to utilize the cell-tissue relationship with different levels of magnification when detecting and classifying cells. In particular, a new dataset named OCELOT was released, containing overlapping cell and tissue annotations based on Hematoxylin and Eosin (H&E) stained Whole Slide Images (WSIs) of multiple organs. Although good results were reached on the OCELOT dataset initially, they were all limited to models based on Convolutional Neural Networks (CNNs) that were years behind the state-of-the-art in Computer Vision (CV) today. The OCELOT dataset was posted as a challenge online, yielding submissions with newer architectures. In this work, we explore the use of transformer-based architecture on the OCELOT dataset and propose a new model architecture specifically made to leverage the added tissue context, which reaches state-of-the-art performance with an F1 score of 72.62% on the official OCELOT test set. Additionally, we explore how the tissue context is used by the models.

Keywords: Computational Pathology · Machine Learning · Computer Vision.

1 Introduction

Computational Pathology (CPATH) is a branch of digital pathology that deals with the development of methods for the analysis of digitized patient specimens, such as Whole Slide Images (WSIs) [6]. A WSI is a digital representation of an entire histopathological glass slide, which is digitized at the resolution of a microscope and produced using slide scanners [1]. The use of WSIs offers

⋆ These authors contributed equally to this work

considerable advantages to the workflow of pathologists as they are immune to physical deterioration and cannot be physically damaged. They are stored digitally and therefore open immense opportunities for computational methods [6].

Cell detection and classification in histology images is an important task in CPATH for aiding patient prognosis and treatment planning [13]. In particular, it can be used for cell counting of breast cancer specimen [3]. Cell counting requires specialized skills and is time-consuming for pathologists, making it a critical task for automation.

Pathologists usually consider different levels of magnification when making diagnoses, considering a wide range of textures in a large Field of View (FoV) and cell morphology in a small FoV [16], but former research in cell detection and classification in histology images has mostly focused on small FoV patches [13]. In this context, Ryu et al. [13] released OCELOT, a dataset with overlapping cell and tissue annotations, together with several model architectures that serve as a baseline. The OCELOT task concerns detecting and classifying cells and is framed as a semantic segmentation task. It was also released as a Grand Challenge [12].

In this work, we explore how transformer-based architectures can be utilized to increase performance on the OCELOT task. We regard our main contribution as two-fold. First, we propose the *Additive joint pred-to-decoder* architecture, a novel architecture using a two-fold loss, a U-Net-like architecture, and overlapped patch merging of the tissue predictions. Second, we take inspiration from recent techniques within XAI to find the situations where the extra tissue information is useful and provide further evidence that it contributes to the classification of cells in the OCELOT task.

2 Related Work

The OCELOT grand challenge contained numerous submissions with different approaches for solving the OCELOT task [12]. Most entries used CNN-based approaches [5, 7, 9, 14], while some entries explored the use of transformer-based architectures [8, 11]. In particular, Millward et al. [11] used the SegFormer, a transformer-based architecture made for semantic segmentation [17], and Li et al. [8] utilized a Vision Transformer (ViT)-based U-Net approach.

Some of the entries changed the labels to better facilitate the learning process, either adapting them to the cell morphology [5] or by experimenting with softening the boundaries [14]. Additionally, a technique called Test-Time Augmentation (TTA), which entails performing augmentations at test-time and averaging the results for better results, was used by Lo and Yang [9] and Schoenpflug and Koelzer [14], both increasing their performance. Finally, Millward et al. [11] found that some of the images were of poor quality and thus omitted them, yielding an increase in performance.

Outside of the OCELOT grand challenge, Gildenblat et al. [4] pointed out that earlier methods did not utilize the information within each cell. They pro-

posed to combine cell-level and tile-level embedding summaries. They demonstrated that their method could boost Human Epidermal Growth Factor Receptor 2 (HER2) and Estrogen Receptor (ER) prediction tasks for breast cancer by up to 8% in Area Under The Curve (AUC).

3 Methodology

All the code used in this work can be found on GitHub[1].

3.1 OCELOT Dataset

The OCELOT dataset, see Ryu et al. [13], comprises 400 WSI samples with two levels of magnification, a smaller "cell" FoV and a larger "tissue" FoV, showing a smaller and larger tissue area, respectively. The images are taken with a digital microscope from multiple organs. Each sample contains six components,

$$\mathcal{D} = \left\{(x_s, y_s^c, x_l, y_l^t, c_x, c_y)_i\right\}_{i=1}^{N}, \tag{1}$$

where x_s and x_l are the small and large FoVs respectively, y_s^c and y_l^t are the corresponding cell and tissue annotations, and c_x and c_y are the relative center coordinates of the center of x_s in x_l.

In the annotations for the tissue patches, each pixel belongs to one of three categories: *background*, *cancer area*, and *unknown*. The annotations for the cell patches, however, are given as a list of coordinates, where each coordinate corresponds to a cell nucleus. Each cell falls into one of two categories: *background cell* and *tumor cell*. To overcome the problem of inconsistent color values in histology slides, we normalize the images using Macenko normalization [10].

3.2 Information Flows

In this work, we use different ways of passing the information to the models, i.e. different information flows, as done by Ryu et al. [13]. The cell-only information flow is used as a baseline to understand the effect of the added context from the tissue patches. This method only uses the cell image and thus disregards the tissue patch. The pred-to-input flow, however, makes use of the extra information by passing a predicted segmentation mask to the input of the model, hence the name "pred-to-input".

We first train a neural network, the *tissue-branch*, and then store the predictions to file. These predictions are then concatenated channel-wise to the cell images and used to train the second neural network, the *cell-branch*. This does incur some degree of data leakage, as the validation set is used to choose the best model for the *tissue-branch* before training the *cell-branch*.

To deal with the data leakage and to better facilitate tissue predictions that are suitable for the cell task, we used the *joint pred-to-input* flow where the loss

[1] Link: https://github.com/bendikgh/histopathology_segmentation/

from both tasks are considered jointly. The loss function is calculated as the sum of the loss of each branch, meaning that the *tissue-branch* gets penalized both for poor tissue predictions and for tissue predictions leading to poor cell predictions.

Finally, we use a *naive model*, which consists of a cell-only model and a tissue-branch, both trained separately on their respective tasks. Then, the cell-only model predicts the location and class of the cells, and then the classifications are overwritten using the predictions from the tissue-branch in the corresponding coordinates. This serves as a naive combination of the two branches.

3.3 Additive Joint Pred-to-decoder Architecture

Building upon the *joint pred-to-input* architecture, and inspired by Xie et al. [17] and Li et al. [8], we propose an architecture that leverages the SegFormer's hierarchical structure by adding the tissue outputs to different levels of the cell encodings. Thus, instead of feeding the tissue outputs to the cell encoder, we feed them to the decoder together with the cell encodings. An overview of this architecture can be seen in Figure 1.

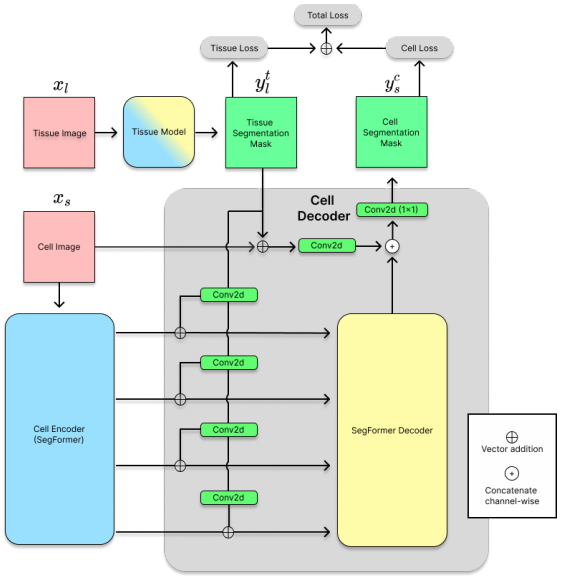


Fig. 1. Visualization of the *additive joint pred-to-decoder* architecture. The outputs from the tissue model are passed to the cell decoder together with the outputs from the cell encoder and the cell image itself. The inputs to the cell decoder are transformed in various ways, using additions and convolutions. A convolutional layer produces the final output of the cell decoder, y_s^c. Vector addition, i.e. element-wise addition, is denoted using the $\oplus$ operator, and channel-wise concatenation is denoted using a small $+$ inside the white circle.

In this architecture, the predictions from the tissue model, y_l^t, are softmaxed and then passed to the cell decoder. There, they are transformed four times, yielding four results, y_1, y_2, y_3 and y_4. This transformation is done using convolutional layers, with $y_1 = \text{Conv2D}_1(y_l^t)$ and $y_i = \text{Conv2D}_i(y_{i-1})$ for $i \in \{2, 3, 4\}$. Conv2D_1 has a kernel size of 7, a padding of 3 and a stride of 4, and $\text{Conv2D}_i, i \in \{2, 3, 4\}$ has a kernel size of 3, padding of 1, and a stride of 2. These sizes are chosen so that they match those used in the overlapped patch merging procedure in Xie et al. [17]. These transformed versions of the tissue output are then added element-wise to the outputs from the cell encoder, giving us the inputs to the SegFormer decoder, $h_i = z_i \oplus y_i$ for $i \in \{1, 2, 3, 4\}$, where z_i is the ith output from the cell encoder.

In addition to this, the tissue segmentation mask is added element-wise to the cell image and then passed through a convolutional layer, $h_0 = \text{Conv2D}_0(x_s \oplus y_l^t)$, which is then concatenated channel-wise with the outputs from the SegFormer decoder.

We choose to add the tissue segmentation mask to the cell image directly to give the model spatial understanding between the cell image and tissue mask. Finally, the concatenated outputs are passed through a 1×1 convolutional layer, like by Li et al. [8], serving as a weighting mechanism for each channel when outputting the final class logits. When performing backpropagation, we use the joint loss of both models. That is, we run backpropagation both with regard to the loss from the tissue target mask and the cell target mask.

3.4 Layer Weight Analysis

When calculating the output of a CNN, a single filter with six channels will create a single output scalar as a weighted sum. Thus, using the associativity of addition, we can split this into a sum of the first and last three channels. Let X be the part of the input being considered at this step of the computation, let W be the weights of the filter and let m and n be the height and width of the filter, respectively. For simplicity, we will disregard the bias in this computation. Then, we can express the single scalar output of the current step as:

$$y = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{6} X_{ijk} \cdot W_{ijk} \tag{2}$$

$$= \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{3} X_{ijk} \cdot W_{ijk} \right) + \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=4}^{6} X_{ijk} \cdot W_{ijk} \right) \tag{3}$$

where y is the scalar output. If we name the first term y_{cell} and the second term y_{tissue}, we can use a simple proxy where we consider the sum of the amplitude of each weight in the filters. That is, we can express an approximation of the emphasis on the cell image, which corresponds to the emphasis on the first three

channels, as:

$$\hat{y}_{\text{cell}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{3} |W_{ijk}| \tag{4}$$

$$\hat{y}_{\text{tissue}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=4}^{6} |W_{ijk}| \tag{5}$$

$$\hat{z} = \frac{\hat{y}_{\text{cell}}}{\hat{y}_{\text{cell}} + \hat{y}_{\text{tissue}}}, \tag{6}$$

where $\hat{z}$ is the proxy for the cell contribution fraction, $\hat{y}_{\text{cell}}$ and $\hat{y}_{\text{tissue}}$ are the proxies for the cell and tissue contributions, respectively. Notably, this proxy suffers from the very implausible assumption that $X_{ijk} = \text{sgn}(W_{ijk})\ \forall i, j, k$, where $\text{sgn}(\cdot)$ is the sign function, and that the input image is the same size as the filter itself, thus requiring only a single step for calculation. $\hat{z}$ only represents a single filter, so if the current layer has multiple filters we must aggregate these values somehow. We have chosen to do this by calculating the mean. The final cell contribution score, ranging from 0 to 1, can thus be calculated as the average over all the filters in the layer. We also keep track of each individual filter value, allowing us to visualize a histogram of the filter contributions for further analysis. This yields some insight into how much weight is put on the cell image versus the tissue image.

3.5 Input-Corrected Analysis

Due to the implausible assumption of the previous method, we extend our method to include actual input images. It is worth noting, however, that it is quite unlikely that the chosen input image effectively represents the entire dataset.

Using an actual input image for X, we no longer require the assumption that $X_{ijk} = \text{sgn}(W_{ijk})\ \forall i, j, k$ or that the input image is the same size as the filter. Thus, we end up with multiple scalar outputs for each filter, i.e. the elements that make up the corresponding channel in the output tensor after convolving the whole input image. To create a single number for each filter, we sum up the absolute values of each element in the output tensor. Similar to the previous section, we calculate the mean and also keep track of the individual filter values, allowing us to create a histogram.

4 Results

When evaluating the cell model's outputs in the experiments, we used the cell-wise mF1 score as done by Ryu et al. [13]. During training, we calculated the Dice Score in a pixel-wise manner, to make it differentiable.

The performance of the different models can be seen in Table 1, where we can see an increase in performance when adding the tissue context. Additionally,

we notice that the SegFormer performs particularly well on the cell only task on the test set, compared to the CNN-based models.

Model	Val Score (%)	Test score (%)
Cell Only		
DeepLabv3+ (Ours)	68.96 ± 1.84	65.91 ± 2.42
DeepLabv3+ (Ryu et al. [13])	68.87 ± 1.76	64.44 ± 1.82
SegFormer	71.67	69.39
Using Cell and Tissue		
DeepLabv3+ (Ours)	70.92 ± 0.35	69.41 ± 0.30
DeepLabv3+ (Ryu et al. [13])	73.36 ± 0.59	69.65 ± 3.93
SegFormer (short)	71.00	69.26
SegFormer (long)	**73.53**	69.47
SegFormer Joint Pred-to-input	72.96	70.16
Joint Additive Pred-to-decoder	73.06	**70.66**

Table 1. The results of the different models when using only the cell information and the cell information together with the tissue information. The "long" and "short" in the parenthesis signifies the training duration for the tissue branch. The scores are reported as mean cell-wise F1 scores, where the models that were recreated from Ryu et al. [13] also include a 95% standard deviation from five trials. The **best scores** are written in bold.

In Table 2, we present the result of using different performance enhancing techniques on the *joint additive pred-to-decoder* model. Notably, we see that both changing the emphasis of the loss function towards the cell training and that utilizing TTA increases the validation and test scores. In particular, the test scores are better than any other results seen on the challenge, to the best of our knowledge.

Method	Val F1 Score (%)	Test F1 Score (%)
Exclusion	73.37	71.66
Exclusion + Cell Emphasis	73.91	72.20
Exclusion + Cell Emphasis + TTA	**74.63**	**72.62**

Table 2. Mean cell-wise F1 scores for the *joint additive pred-to-decoder* with different performance-enhancing techniques. The **best scores** are written in bold.

In Table 3, we can see the empirical results of the naive model. We can see that naively adding the tissue class to the cell-only locations yields a negative impact on the performance, while adding the actual labels increase the performance, but only slightly. We also notice that the cell-only model has a higher recall and lower precision on background cells, while the inverse is true for the tumor cells.

Model	mF1	Background Cells			Tumor Cells		
		Prec.	Rec.	F1	Prec.	Rec.	F1
Cell-only	**71.67**	63.47	**67.07**	**65.22**	**84.82**	72.39	**78.12**
Tissue + Cell	69.91	**65.77**	58.37	61.85	80.43	**75.66**	77.97
Labels + Cell	72.54	71.62	60.38	65.52	81.08	78.12	79.57

Table 3. Different statistics on the validation set for the *cell-only* model and the naive models using tissue predictions and tissue labels to classify cells. "Prec." refers to precision, "Rec." refers to recall and "mF1" refers to the mean cell-wise F1 Score. With the exception of the model using leaked labels, the **best scores** are written in bold.

We also include qualitative results from the *tissue + cell* configuration: In Figure 2, we can see that the *cell-only* model is mostly correct in its predictions. However, during the adjustment of labels with the tissue predictions, we see that many of the cells in the bottom left corner, that were correctly predicted as background cells, were changed to tumor cells as a result of the tissue classification of the tissue model. This tissue classification is indeed correct, but it does not immediately follow that the encompassed cells are cancerous.

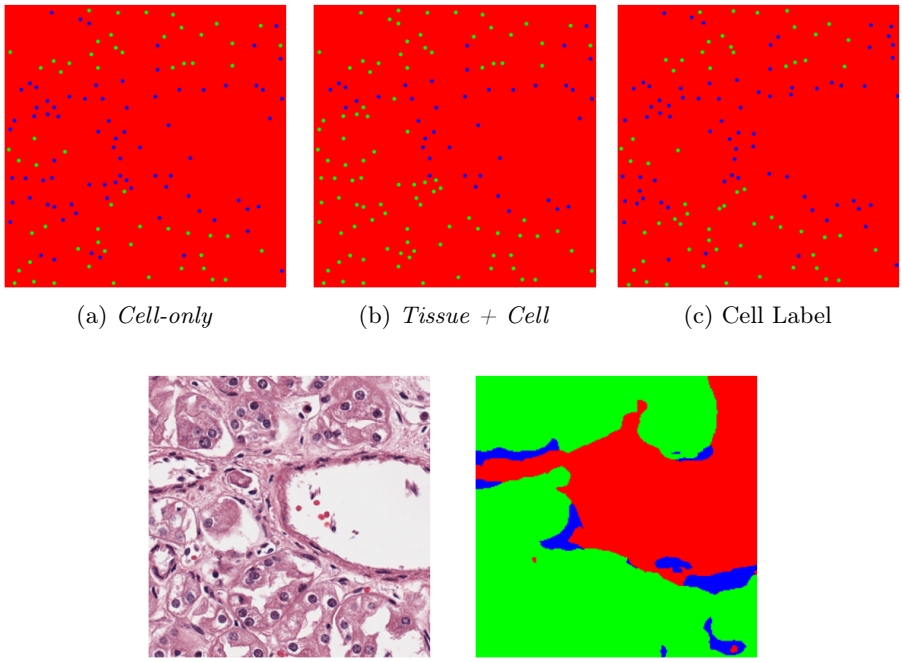
In Figure 3, we can see the results of the weight analysis of the SegFormer trained for a short duration and a long duration. We can see that most of the weights are centered around the mean, with a couple of outliers. Additionally, we can see that the means of the two models are quite close, but with the mode of scores slightly left of the mean for the long tissue training, while directly on the mean for the short tissue training.

In Figure 5, we can see the results of the input-corrected analysis of the same models. The input images are shown in 4. We see that the corrected histograms are much flatter than in Figure 3, and that the model with long training has slightly more values on the left side of the histogram, indicating a higher emphasis on the tissue weights. Additionally, we see that the mean is now considerably lower for the model with long training.

5 Discussion

As seen in Table 1 and Table 2, we achieve the highest performance with our *joint additive pred-to-decoder* model with TTA and *exclusion*. We note that the SegFormer model indeed achieved a higher validation score in Table 1, but this is both with slight data leakage and the inability to train using the joint loss. Our best score is better than any we have seen in any of the submissions to the challenge and suggests that our novel architecture is successful at effectively combining the cell and tissue images in a useful way.

We believe that the high performance of the *additive joint pred-to-decoder* model could be due to several reasons. First, due to the joint loss, we believe it has a better ability to tailor the tissue probits specifically towards the cell task and take advantage of the uncertainty reflected by the probits compared to



(a) *Cell-only* (b) *Tissue + Cell* (c) Cell Label

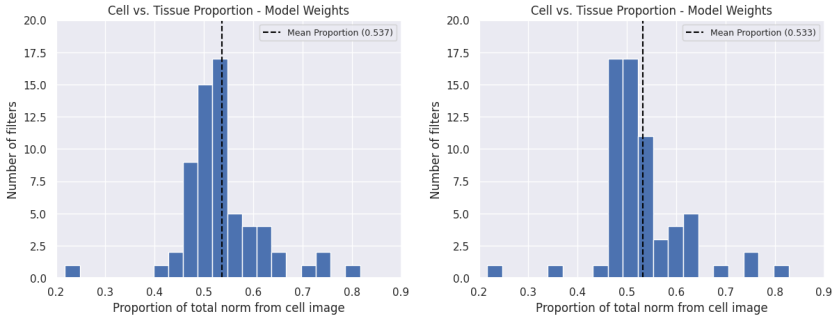
(d) Input Image (e) Tissue Pred.

Fig. 2. Illustration of predictions from the different models. (a) shows the predictions of the *cell-only* model, (b) shows the predictions of the *tissue + cell* model and (c) shows the target values. Red pixels signify background tissue, i.e. no cell present, blue pixels signify that there is a cell classified as a background cell and green pixels signify that there is a cell classified as cancerous. (d) shows the input image that the models received and (e) shows the tissue predictions on the input image.

earlier models. Also, the joint loss may push the model to generalize more, as it has to optimize multiple tasks at once, possibly making it harder to overfit towards one.

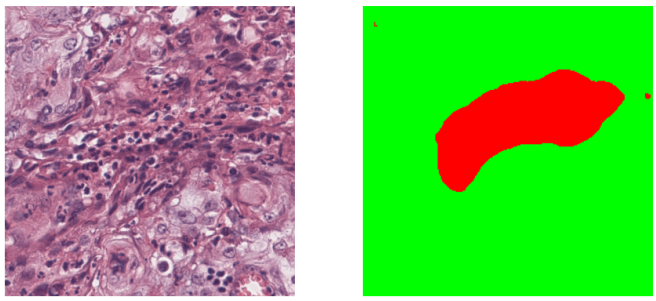
Second, the model has a U-Net-like structure between the encoder and the decoder that adds a spatial bias at all intermediate steps of the model. We do this by matching the receptive fields of each convolutional layer with those of the SegFormer model. This structure also utilizes the overlapping patch merging mechanism of the SegFormer [17], which we believe allows it to make use of the tissue predictions more effectively, as each patch also considers parts of surrounding patches.

From the results in Table 3 we see that the naive model using the tissue predictions performed worse than the cell-only model. In particular, we notice that the naive model has a higher precision and lower recall for the background cells, but lower precision and higher recall for the tumor cells. This tells us that a potential deficiency of the cell-only model is that it misclassified tumor cells



(a) *SegFormer, Short (mean: 53.69%)* (b) *SegFormer, Long (mean: 53.27%)*

Fig. 3. Histograms of the cell proportions for the SegFormer with a tissue-branch trained (a) for a short time and (b) for a long time, using the layer weight analysis. The vertical axis show the number of filters and the horizontal axis shows the proportion of weights belonging to the cell image, as a score from 0 to 1.



(a) *Cell Input Image* (b) *Tissue Input Image*

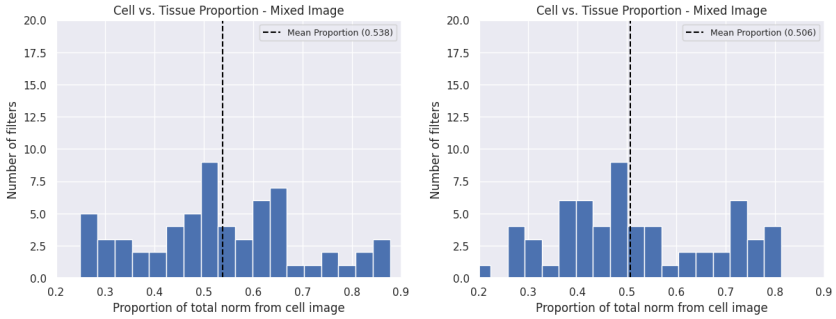
Fig. 4. The (a) cell input image and (b) tissue input image for the input-corrected analysis. In (b), red means background tissue and green means cancerous tissue.

as background cells, while the naive model fails as the prevalence of background cells in cancerous areas is large.

When considering the weight analysis, we see from Figure 3 and Figure 5 that although the results between the models are quite similar with the naive weight analysis, they start differing considerably when performing the input correction. The latter substantiates our thought that the better models are able to more effectively use the tissue information, as this shows that the better performing models put more emphasis on the tissue images.

6 Conclusion

In this work, we propose the *additive joint pred-to-decoder*, a U-Net-like structure utilizing a joint loss. We show that the new architecture is competitive with



(a) *SegFormer, Short (mean: 53.85)* (b) *SegFormer, Long (mean: 50.63)*

Fig. 5. Histograms of the cell proportions for the SegFormer with a tissue-branch trained (a) for a short time and (b) for a long time, using the input-corrected analysis. The vertical axis show the number of filters and the horizontal axis shows the proportion of weights belonging to the cell image, as a score from 0 to 1.

recent architectures, and by optimizing the training procedure, we show that it outperforms the current SOTA on the OCELOT task. Inspired by recent developments in the field of XAI, we develop a method to approximate the importance of the surrounding tissue classifications. We find that the models emphasizing tissue structures perform better.

7 Shortcomings and Future Work

We had a case of data leakage in some of the earlier experiments. In particular, we trained the tissue branch separately and then chose the model based on the performance of the validation set. This meant that the weights of the cell branch were influenced indirectly by the validation set, which means that the scores attained on the validation set with these models are not entirely valid. Still, this is not the case for the *additive joint pred-to-decoder*.

We consider model explainability to be an interesting area for future work. Although there exists frameworks for LRP [2] and GradCAM [15], these are not compatible with the dual-input modality of this dataset. Our approach is quite rudimentary, so focusing on adapting more complex techniques to this task could yield even better understandability and thus more trust in the techniques, facilitating clinical use.

Acknowledgments. The experiments of this study were conducted as a part of a master's thesis at the Norwegian University of Science and Technology.

Disclosure of Interests. The authors have no competing interests relevant to this article's content.

Bibliography

[1] Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M.D., Van Der Laak, J., Bui, M.M., Vemuri, V.N., Parwani, A.V., Gibbs, J., Agosto-Arroyo, E., Beck, A.H., Kozlowski, C.: Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. The Journal of Pathology **249**(3), 286–294 (Nov 2019), ISSN 0022-3417, 1096-9896, https://doi.org/10.1002/path.5331, URL https://pathsocjournals.onlinelibrary.wiley.com/doi/10.1002/path.5331

[2] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE **10**(7), e0130140 (Jul 2015), ISSN 1932-6203, https://doi.org/10.1371/journal.pone.0130140, URL https://dx.plos.org/10.1371/journal.pone.0130140

[3] Chen, L., Bao, J., Huang, Q., Sun, H.: A robust and automated cell counting method in quantification of digital breast cancer immunohistochemistry images. Polish Journal of Pathology **70**(3), 162–173 (2019), ISSN 1233-9687, https://doi.org/10.5114/pjp.2019.90392

[4] Gildenblat, J., Yüce, A., Abbasi-Sureshjani, S., Korski, K.: Deep Cellular Embeddings: An Explainable Plug and Play Improvement for Feature Representation in Histopathology, Lecture Notes in Computer Science, vol. 14225, p. 776–785. Springer Nature Switzerland, Cham (2023), ISBN 978-3-031-43986-5, https://doi.org/10.1007/978-3-031-43987-2_75, URL https://link.springer.com/10.1007/978-3-031-43987-2_75

[5] Ha, S.M., Ko, Y.S., Park, Y.: Generating BlobCell Label from Weak Annotations for Precise Cell Segmentation. In: Ahmadi, S.A., Pereira, S. (eds.) Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology, vol. 14373, pp. 161–170, Springer Nature Switzerland, Cham (2024), ISBN 978-3-031-55087-4 978-3-031-55088-1, https://doi.org/10.1007/978-3-031-55088-1_15, URL https://link.springer.com/10.1007/978-3-031-55088-1_15, series Title: Lecture Notes in Computer Science

[6] Hosseini, M.S., Bejnordi, B.E., Trinh, V.Q.H., Chan, L., Hasan, D., Li, X., Yang, S., Kim, T., Zhang, H., Wu, T., Chinniah, K., Maghsoudlou, S., Zhang, R., Zhu, J., Khaki, S., Buin, A., Chaji, F., Salehi, A., Nguyen, B.N., Samaras, D., Plataniotis, K.N.: Computational Pathology: A Survey Review and The Way Forward. Journal of Pathology Informatics p. 100357 (Jan 2024), ISSN 21533539, https://doi.org/10.1016/j.jpi.2023.100357, URL https://linkinghub.elsevier.com/retrieve/pii/S2153353923001712

[7] Lafarge, M.W., Koelzer, V.H.: Detecting Cells in Histopathology Images with a ResNet Ensemble Model. In: Ahmadi, S.A., Pereira, S. (eds.) Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology, vol. 14373, pp. 123–129,

Springer Nature Switzerland, Cham (2024), ISBN 978-3-031-55087-4 978-3-031-55088-1, https://doi.org/10.1007/978-3-031-55088-1_11, URL https://link.springer.com/10.1007/978-3-031-55088-1_11, series Title: Lecture Notes in Computer Science

[8] Li, Z., Li, W., Mai, H., Zhang, T., Xiong, Z.: Enhancing Cell Detection in Histopathology Images: A ViT-Based U-Net Approach, Lecture Notes in Computer Science, vol. 14373, p. 150–160. Springer Nature Switzerland, Cham (2024), ISBN 978-3-031-55087-4, https://doi.org/10.1007/978-3-031-55088-1_14, URL https://link.springer.com/10.1007/978-3-031-55088-1_14

[9] Lo, Y.W., Yang, C.H.: Enhancing Cell Detection via FC-HarDNet and Tissue Segmentation: OCELOT 2023 Challenge Approach. In: Ahmadi, S.A., Pereira, S. (eds.) Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology, vol. 14373, pp. 130–137, Springer Nature Switzerland, Cham (2024), ISBN 978-3-031-55087-4 978-3-031-55088-1, https://doi.org/10.1007/978-3-031-55088-1_12, URL https://link.springer.com/10.1007/978-3-031-55088-1_12, series Title: Lecture Notes in Computer Science

[10] Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Xiaojun Guan, Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1107–1110, IEEE, Boston, MA, USA (Jun 2009), ISBN 978-1-4244-3931-7, https://doi.org/10.1109/ISBI.2009.5193250, URL http://ieeexplore.ieee.org/document/5193250/

[11] Millward, J., He, Z., Nibali, A.: Dense Prediction of Cell Centroids Using Tissue Context and Cell Refinement. In: Ahmadi, S.A., Pereira, S. (eds.) Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology, vol. 14373, pp. 138–149, Springer Nature Switzerland, Cham (2024), ISBN 978-3-031-55087-4 978-3-031-55088-1, https://doi.org/10.1007/978-3-031-55088-1_13, URL https://link.springer.com/10.1007/978-3-031-55088-1_13, series Title: Lecture Notes in Computer Science

[12] Ryu, J., Puche, A.V., Shin, J., Park, S., Brattoli, B., Lee, J., Jung, W., Cho, S.I., Paeng, K., Ock, C.Y., Yoo, D., Pereira, S.: Home - Grand Challenge (????), URL https://ocelot2023.grand-challenge.org/ocelot2023/

[13] Ryu, J., Puche, A.V., Shin, J., Park, S., Brattoli, B., Lee, J., Jung, W., Cho, S.I., Paeng, K., Ock, C.Y., Yoo, D., Pereira, S.: Ocelot: Overlapped cell on tissue dataset for histopathology (2023)

[14] Schoenpflug, L.A., Koelzer, V.H.: SoftCTM: Cell Detection by Soft Instance Segmentation and Consideration of Cell-Tissue Interaction. In: Ahmadi, S.A., Pereira, S. (eds.) Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology, vol. 14373, pp. 109–122, Springer Nature Switzerland, Cham (2024), ISBN 978-3-031-55087-4 978-3-031-55088-1, https://doi.org/10.1007/978-3-031-55088-1_10, URL

https://link.springer.com/10.1007/978-3-031-55088-1_10, series Title: Lecture Notes in Computer Science

[15] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. International Journal of Computer Vision **128**(2), 336–359 (Feb 2020), ISSN 0920-5691, 1573-1405, https://doi.org/10.1007/s11263-019-01228-7, URL http://arxiv.org/abs/1610.02391, arXiv:1610.02391 [cs]

[16] Tokunaga, H., Teramoto, Y., Yoshizawa, A., Bise, R.: Adaptive Weighting Multi-Field-of-View CNN for Semantic Segmentation in Pathology (Apr 2019), URL http://arxiv.org/abs/1904.06040, arXiv:1904.06040 [cs]

[17] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers (Oct 2021), URL http://arxiv.org/abs/2105.15203, arXiv:2105.15203 [cs]