Hierarchical Vision Transformers for Context-Aware Prostate Cancer Grading in Whole Slide Images

Clément Grisi*, Geert Litjens, Jeroen van der Laak

Computational Pathology Group, Radboudumc, Netherlands

Abstract

Vision Transformers (ViTs) have ushered in a new era in computer vision, showcasing unparalleled performance in many challenging tasks. However, their practical deployment in computational pathology has largely been constrained by the sheer size of whole slide images (WSIs), which result in lengthy input sequences. Transformers faced a similar limitation when applied to long documents, and Hierarchical Transformers were introduced to circumvent it. Given the analogous challenge with WSIs and their inherent hierarchical structure, Hierarchical Vision Transformers (H-ViTs) emerge as a promising solution in computational pathology. This work delves into the capabilities of H-ViTs, evaluating their efficiency for prostate cancer grading in WSIs. Our results show that they achieve competitive performance against existing state-of-the-art solutions.

1 Introduction

With the advent of whole-slide imaging technology, large numbers of tissue slides are being scanned and archived digitally. Digitization of histological slides has contributed to the growing availability of large datasets, fostering tremendous computer vision research opportunities. Recent progress in deep learning has set new performance standards in various clinical applications, especially with the recent surge of attention-based models (Dosovitskiy et al. [2021]).

However, conventional deep learning methods are ill-equipped to handle the enormous sizes of whole-slide images (WSIs), which do not fit into the memory of graph processing units (GPUs). Innovative strategies have emerged to overcome this memory bottleneck. A prevailing approach consists of partitioning these massive images into smaller patches. These patches often serve as the input units, with labels derived from pixel-level annotations. Because the traditional requirement of input-label pairs can be prohibitive – or even impossible as pathologists can only annotate what they know – recent research has explored techniques beyond full supervision, focusing on more flexible training paradigms, such as weakly supervised learning. Multiple Instance Learning (MIL) has recently emerged as a powerful weakly supervised approach in several computational pathology challenges, where it has shown remarkable performance (Hou et al. [2015], Campanella et al. [2019]).

Despite showcasing impressive results, most MIL methods neglect the spatial relationships among patches, thereby overlooking valuable contextual information. To address this limitation, recent research focused on developing methods capable of integrating context (Lerousseau et al. [2021], Pinckaers et al. [2021], Shao et al. [2021]). By leveraging the hierarchical structure inherent to WSIs, Hierarchical Vision Transformers (H-ViTs) have proven successful at learning context-aware image representations from whole slide images, achieving state-of-the-art results in cancer subtyping and survival prediction (Chen et al. [2022]). We explore their application in the

*Corresponding author: clement.grisi@radboudumc.nl

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

multi-class classification setting of prostate cancer grading. We additionally provide enhanced model interpretability by introducing an innovative approach for computing factorized attention heatmaps. Our code is available at `https://github.com/computationalpathologygroup/hvit`.

2 Proposed Method

Hierarchical Vision Transformer. The inherent hierarchical structure within whole-slide images spans across various scales, from tiny cell-centric patches containing fine-grained information, all the way up to the entire slide, which exhibits overall intra-tumoral heterogeneity of the tissue microenvironment. Along this spectrum, various patch sizes may capture cell-to-cell interactions or macro-scale interactions. Drawing inspiration from this layered structure, our H-ViT model mimicks of the multi-stage architecture introduced in Chen et al. [2022], which incorporates multiple Vision Transformers. Each performs bottom-up aggregation, effectively mapping the token sequence at one scale into a single representation at the subsequent scale, eventually resulting in a slide-level feature vector (Appendix A). A first Vision Transformer (referred to as patch-level Transformer) performs cell-level aggregation by breaking down 256×256 patches into 16×16 mini-patches. Then, a second Transformer (referred to as region-level Transformer) builds context-aware embeddings by aggregating non-overlapping 256×256 patches within larger 2048×2048 regions. Finally, a third Transformer (referred to as slide-level Transformer) pools the resulting region-level tokens into a single slide-level representation that can be used for downstream prediction tasks.

We experimented with two variants of this model. The first, deemed Global H-ViT, replicates HIPT (Chen et al. [2022]) : both patch-level and region-level Transformers are pretrained and kept frozen. Only the slide-level Transformer is finetuned on the downstream task. The second one, deemed Local H-ViT, consists of a pretrained and frozen patch-level Transformer, with both the region-level and slide-level Transformers finetuned on the downstream task.

Refined Attention Factorization. It's important to underscore that the attention scores from the different Transformers hold varying degrees of relevance in highlighting the areas critical for the model's predictions. Indeed, some are pretrained and frozen, while others are finetuned for the downstream classification task. To address this disparity, we introduce a parameter $\gamma \in [0, 1]$ to balance the influence of attention scores between frozen and finetuned Transformers.

Let N be the total number of Transformers involved in the pretraining or the finetuning processes. Let n be the number of frozen Transformers. We denote by $a^i_{(x,y)}$ the attention score of the i-th Transformer T_i for the pixel with (x, y) coordinates in the slide. The factorized attention score for that pixel, $a_{(x,y)}$, is then computed as:

$$a_{(x,y)} = \frac{1}{\beta} \sum_{i=0}^{N-1} a^i_{(x,y)} \big(\gamma \mathbb{1}_{\mathrm{F}}(\mathrm{T}_i) + (1-\gamma)(1 - \mathbb{1}_{\mathrm{F}}(\mathrm{T}_i))\big) \tag{1}$$

with $\beta = n \cdot (1-\gamma) + (N-n) \cdot \gamma$ and $\mathbb{1}_{\mathrm{F}}(\mathrm{T}_i)$ the indicator function defined on the set of Transformers $\{\mathrm{T}_i\}_{i=0}^{N-1}$ and equal to 1 if T_i is finetuned, 0 otherwise. This flexible approach allows for fine-grained control over how attention scores are combined, making it possible to emphasize either pretrained or task-specific features in the model's attention mechanism.

3 Experimental Results

Dataset. To assess the robustness of the proposed method, we use the PANDA dataset, introduced in Bulten et al. [2022]. It is the largest publicly available dataset of H&E stained prostate WSIs to date, with $11,554$ prostate biopsies curated from two different sites. All slides are provided at a pixel spacing close to 0.50 μm, with their corresponding ISUP score (Appendix B).

Data Preprocessing & Evaluation Metric. We adapted Lu et al. [2021] preprocessing pipeline to automatically segment tissue in each slide, from which we extract non-overlapping 2048×2048 regions at a resolution of 0.50 micron per pixel (Appendix C). We split the development set into

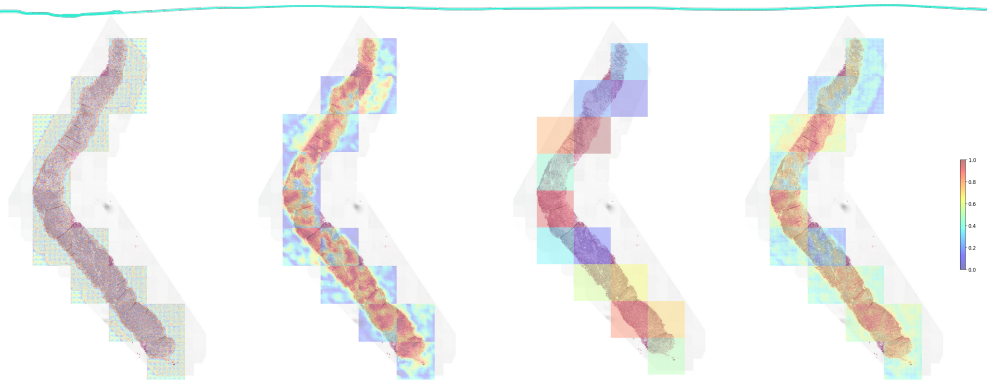
5 cross-validation folds, stratifying on the ISUP score. To evaluate the model's classification performance, we report quadratic weighted kappa scores on the tuning set, the public test set and the private test set, averaged across the 5 folds.

Prostate Cancer Grading. We pretrain the patch-level and region-level Transformers on the PANDA dataset via the student-teacher knowledge distillation framework DINO (Caron et al. [2021]). To leverage the ordinal nature of the ISUP scores, we formulated the classification problem as a regression task and used the Mean Squared Error (MSE) loss. Classification results are summarized in Table 1. Local H-ViT achieves higher macro-averaged performance than Global H-ViT and comparable performance with the winning solutions to the PANDA challenge (Appendix D). Allowing gradients to flow through the region-level Transformer grants the model more freedom to refine the pretrained features, such that they better fit the classification task at hand.

Table 1: Classification results

Model	# parameters	Tune Score	Public Test Score	Private Test Score
Global H-ViT	$594,818$	0.888 ± 0.019	0.799 ± 0.018	0.807 ± 0.016
Local H-ViT	$3,348,098$	$\mathbf{0.950 \pm 0.002}$	$\mathbf{0.904 \pm 0.011}$	$\mathbf{0.904 \pm 0.011}$

Model Interpretability. Attention heatmaps offer a streamlined form of model interpretability, revealing the specific image features that the model has learned to associate with each class. To that end, factorized heatmaps generated using our approach offer a more comprehensive view of model attention than heatmaps from single-level Transformers (Figure 1). Using $\gamma = 0.5$ in Equation 1 equally weights the contributions of the frozen and finetuned Transformers. We advocate for γ values greater than 0.5, as these upweight the contributions of finetuned attention scores, hence highlighting features that are more directly related to the prediction task at hand. Among the various values tested, $\gamma = 0.7$ yielded visually satisfying and informative heatmaps, balancing fine-grained and coarse-grained attention features (Appendix E). We only show heatmaps for Local H-ViT ($n = 1$) as it is our best-performing model.



(a) patch-level (b) region-level (c) slide-level (d) factorized ($\gamma = 0.7$)

Figure 1: Local H-ViT individual and factorized attention heatmaps

4 Conclusion

We presented Hierarchical Vision Transformer (H-ViT), a promising method to address the unique challenges associated with the analysis of whole-slide images in computational pathology. By leveraging the inherent hierarchical structure of WSIs, H-ViT efficiently captures context-aware representations, providing a comprehensive view of the tissue microenvironment. Our experiments showcased the efficacy of this method in prostate cancer grading, with Local H-ViT outperforming Global H-ViT and achieving performance comparable to the winning solutions of the PANDA

challenge. Our proposed attention factorization method offers a nuanced and controllable mechanism for model interpretability, providing practical insights that could be valuable for clinical interpretation.

Acknowledgments

We express our sincere gratitude to Richard Chen for his pioneering work, "Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning". The model adopted in this paper is inspired by his research. We are especially thankful for his generous support and timely answers, which were instrumental in enhancing our understanding and successful implementation of the model.

Potential Negative Societal Impact

To the best of our knowledge, we did not find any potential negative societal impact of our research. All clinical data used in our experiments were sourced from anonymized, publicly available datasets, ensuring patient privacy. Our method could help doctors make faster and more consistent decisions, in the interest of patients. It is not meant to conduct any direct diagnosis.

References

Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F. Steiner, Hester van Boven, Robert Vink, Christina Hulsbergen-van de Kaa, Jeroen van der Laak, Mahul B. Amin, Andrew J. Evans, Theodorus van der Kwast, Robert Allan, Peter A. Humphrey, Henrik Grönberg, Hemamali Samaratunga, Brett Delahunt, Toyonori Tsuzuki, Tomi Häkkinen, Lars Egevad, Maggie Demkin, Sohier Dane, Fraser Tan, Masi Valkonen, Greg S. Corrado, Lily Peng, Craig H. Mermel, Pekka Ruusuvuori, Geert Litjens, Martin Eklund, Américo Brilhante, Aslı Çakır, Xavier Farré, Katerina Geronatsiou, Vincent Molinié, Guilherme Pereira, Paromita Roy, Günter Saile, Paulo G. O. Salles, Ewout Schaafsma, Joëlle Tschui, Jorge Billoch-Lima, Emíio M. Pereira, Ming Zhou, Shujun He, Sejun Song, Qing Sun, Hiroshi Yoshihara, Taiki Yamaguchi, Kosaku Ono, Tao Shen, Jianyi Ji, Arnaud Roussel, Kairong Zhou, Tianrui Chai, Nina Weng, Dmitry Grechka, Maxim V. Shugaev, Raphael Kiminya, Vassili Kovalev, Dmitry Voynov, Valery Malyshev, Elizabeth Lapo, Manuel Campos, Noriaki Ota, Shinsuke Yamaoka, Yusuke Fujimoto, Kentaro Yoshioka, Joni Juvonen, Mikko Tukiainen, Antti Karlsson, Rui Guo, Chia-Lun Hsieh, Igor Zubarev, Habib S. T. Bukhar, Wenyuan Li, Jiayun Li, William Speier, Corey Arnold, Kyungdoc Kim, Byeonguk Bae, Yeong Won Kim, Hong-Seok Lee, Jeonghyuk Park, and the PANDA challenge consortium. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature Medicine*, 28(1):154–163, Jan 2022. ISSN 1546-170X. doi: 10.1038/s41591-021-01620-2. URL `https://doi.org/10.1038/s41591-021-01620-2`.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, Aug 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0508-1. URL `https://doi.org/10.1038/s41591-019-0508-1`.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16144–16155, June 2022.

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, 2021.

Le Hou, Dimitris Samaras, Tahsin M. Kurç, Yi Gao, James E. Davis, and J. Saltz. Efficient multiple instance convolutional neural networks for gigapixel resolution image classification. *ArXiv*, abs/1504.07947, 2015. URL `https://api.semanticscholar.org/CorpusID:16405142`.

Marvin Lerousseau, Maria Vakalopoulou, Eric Deutsch, and Nikos Paragios. Sparseconvmil: Sparse convolutional context-aware multiple instance learning for whole slide image classification, 2021.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, Jun 2021. ISSN 2157-846X. doi: 10.1038/s41551-020-00682-w. URL https://doi.org/10.1038/s41551-020-00682-w.

Hans Pinckaers, Wouter Bulten, Jeroen Van der Laak, and Geert Litjens. Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels. *IEEE Trans Med Imaging*, PP, March 2021. ISSN 1558-254X. doi: 10.1109/TMI.2021.3066295.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification, 2021.

A Architecture Overview

Figure 2 shows the multi-stage H-ViT architecture we use in this work. It features three Vision Transformers, followed by a simple linear classifier that projects the resulting slide embedding onto the desired number of classes.

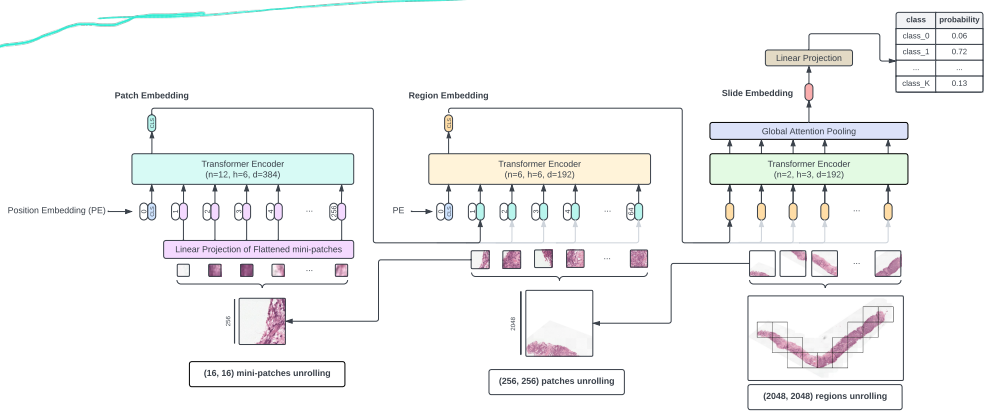


Figure 2: H-ViT overview

B PANDA Dataset Details

In Table 2, we provide a summary of the main characteristics of the PANDA dataset.

Table 2: PANDA dataset summary

Site	Scanner	Spacing (μm)	# dev	# public test	# private test
Radboud	3DHistech	0.48	5160	195	333
Karolinska	Leica	0.50	2193	97	150
Karolinska	Hamamatsu	0.45	3263	101	62

Pathologists classify tumors into different growth patterns by analyzing the histological architecture of the tumor tissue. Tissue specimens are then categorized into one of five groups based on the distribution of these patterns in the tumor. Figure 3 shows the grade group distribution for the development set, the public test set and the private test set.

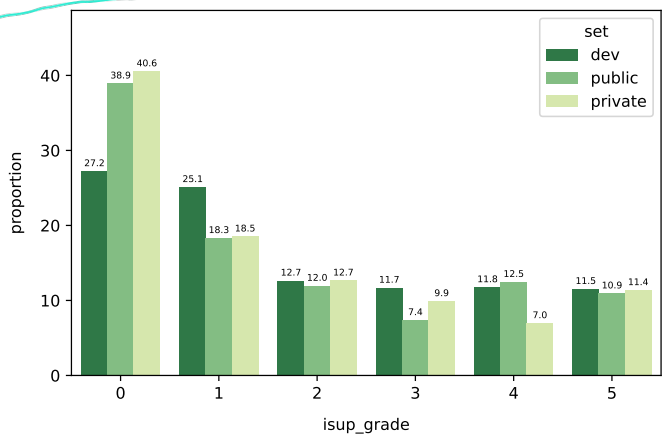
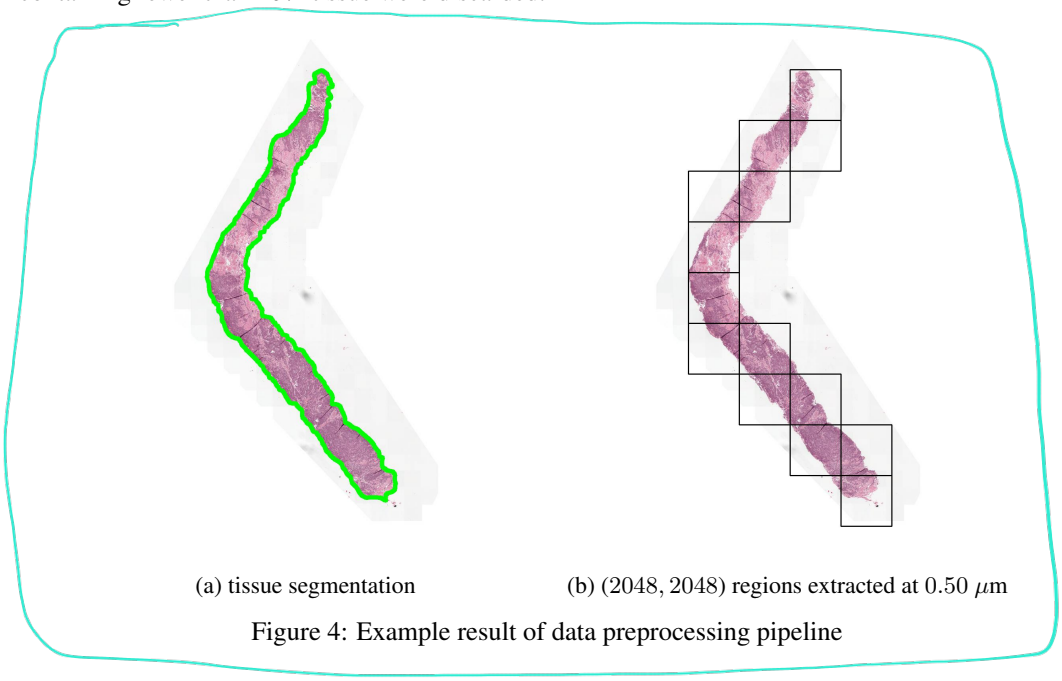


Figure 3: PANDA label distribution

C Data Preprocessing

Figure 4 shows an example result of our tissue segmentation and region extraction algorithm. Regions containing fewer than 10% tissue were discarded.



(a) tissue segmentation (b) (2048, 2048) regions extracted at 0.50 μm

Figure 4: Example result of data preprocessing pipeline

D Winning Solutions to the PANDA Challenge

Top-performing methods demonstrated remarkable accuracy. On the public test set, the best submissions achieved quadratic weighted kappa scores ranging from 0.89 to 0.92. Similarly, on the private test set, the leading entries scored between 0.92 and 0.94. Table 3 shows the results of the 14[2] teams invited to join the PANDA consortium, as well as an additional entry for our best performing model (obtained by ensembling the predictions of each fold).

Table 3: Comparison with PANDA consortium teams

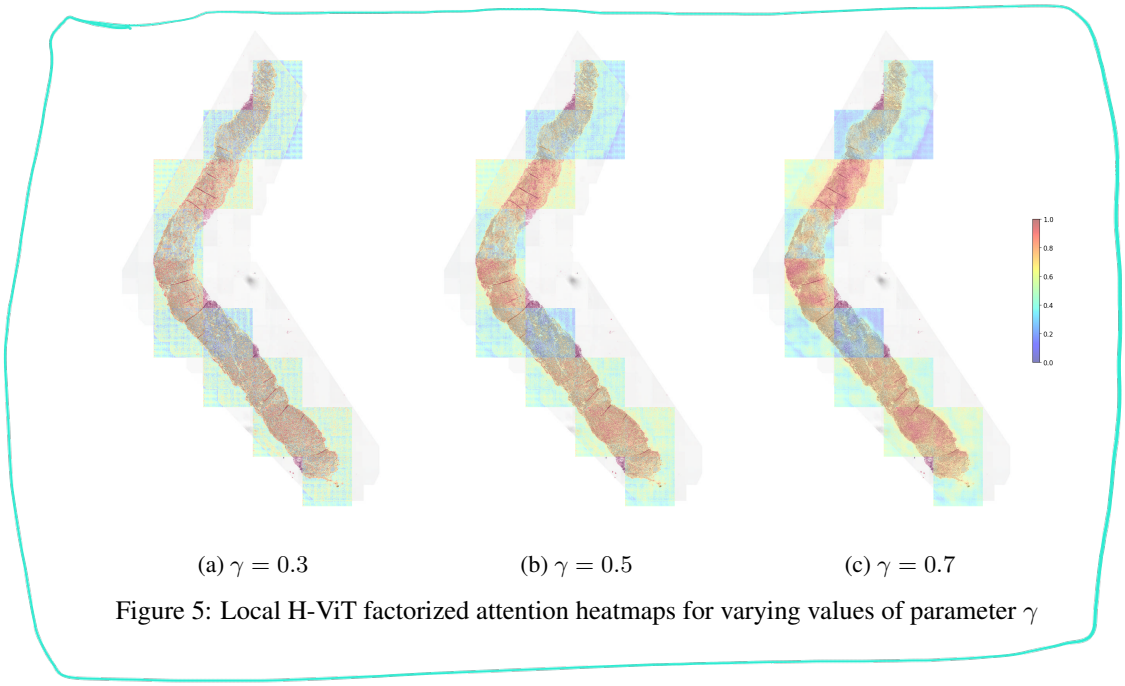
Team Name	Public Test Score	Private Test Score	Combined Test Score
Save The Prostate	0.9209	0.9377	**0.9280**
NS Pathology	0.9180	0.9340	0.9272
PND	0.9108	**0.9408**	0.9252
iafoss	0.9179	0.9301	0.9250
Aksell	0.9212	0.9274	0.9247
vanda	**0.9219**	0.9303	0.9215
ChienYiChi	0.9086	0.9324	0.9214
BarelyBears	0.9118	0.9326	0.9204
rähmä.ai	0.9096	0.9262	0.9182
ctrasd123	0.8948	0.9324	0.9165
Kiminya	0.9007	0.9328	0.9164
Manuel Campos	0.8935	0.9296	0.9142
Dmitry A. Grechka	0.8861	0.9283	0.9105
KovaLOVE v2	0.8889	0.9277	0.9099
Local H-ViT	0.9149	0.9170	0.9161

[2]there were 15 teams invited in total, but the scores for the team "UCLA Computational Diagnostics Lab" couldn't be retrieved

In our comparative analysis, a two-sided permutation test on the combined test set between our model and the best performing team (`Save The Prostate`) shows the difference in performance is not statistically significant ($p = 0.1235$).

E Factorized Attention Heatmaps

Figure 5 shows an example of factorized heatmaps for varying values of parameter γ. The choice of γ indirectly controls the granularity of the factorized attention heatmaps: values closer to 0 produce finer-grained heatmaps focused on patch-level attention, whereas values closer to 1 yield coarse-grained heatmaps focused on region-level and slide-level attention. Further analysis with expert pathologists is needed to better understand the morphological patterns identified by the model. This collaboration will also help determine the most suitable value for γ.



(a) $\gamma = 0.3$ (b) $\gamma = 0.5$ (c) $\gamma = 0.7$

Figure 5: Local H-ViT factorized attention heatmaps for varying values of parameter γ