Predicting the Visual Attention of Pathologists Evaluating Whole Slide Images of Cancer

Souradeep Chakraborty[1(✉)], Rajarsi Gupta[2], Ke Ma[9], Darshana Govind[5], Pinaki Sarder[6], Won-Tak Choi[8], Waqas Mahmud[2], Eric Yee[7], Felicia Allard[7], Beatrice Knudsen[3], Gregory Zelinsky[1,4], Joel Saltz[2], and Dimitris Samaras[1]

[1] Department of Computer Science, Stony Brook University, Stony Brook, NY, USA
souchakrabor@cs.stonybrook.edu
[2] Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA
[3] Department of Pathology, University of Utah School of Medicine, Salt Lake City, UT, USA
[4] Department of Psychology, Stony Brook University, Stony Brook, NY, USA
[5] Department of Pathology and Anatomical Sciences, University at Buffalo, Buffalo, NY, USA
[6] Department of Medicine, University of Florida at Gainesville, Gainesville, FL, USA
[7] Department of Pathology, University of Arkansas for Medical Sciences, Little Rock, AR, USA
[8] Department of Pathology, University of California San Francisco, San Francisco, CA, USA
[9] Snap Inc., Santa Monica, USA

Abstract. This work presents PathAttFormer, a deep learning model that predicts the visual attention of pathologists viewing whole slide images (WSIs) while evaluating cancer. This model has two main components: (1) a patch-wise attention prediction module using a Swin transformer backbone and (2) a self-attention based attention refinement module to compute pairwise-similarity between patches to predict spatially consistent attention heatmaps. We observed a high level of agreement between model predictions and actual viewing behavior, collected by capturing panning and zooming movements using a digital microscope interface. Visual attention was analyzed in the evaluation of prostate cancer and gastrointestinal neuroendocrine tumors (GI-NETs), which differ greatly in terms of diagnostic paradigms and the demands on attention. Prostate cancer involves examining WSIs stained with Hematoxylin and Eosin (H&E) to identify distinct growth patterns for Gleason grading. In contrast, GI-NETs require a multi-step approach of identifying tumor regions in H&E WSIs and grading by quantifying the number of Ki-67 positive tumor cells highlighted with immunohistochemistry (IHC) in a separate image. We collected attention data from pathologists viewing

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-16961-8_2.

Y. Huo et al. (Eds.): MOVI 2022, LNCS 13578, pp. 11–21, 2022.
https://doi.org/10.1007/978-3-031-16961-8_2

prostate cancer H&E WSIs from The Cancer Genome Atlas (TCGA) and 21 H&E WSIs of GI-NETs with corresponding Ki-67 IHC WSIs. This is the first work that utilizes the Swin transformer architecture to predict visual attention in histopathology images of GI-NETs, which is generalizable to predicting attention in the evaluation of multiple sequential images in real world diagnostic pathology and IHC applications.

Keywords: Visual attention · Digital microscopy · Cognitive pathology

1 Introduction

Attention tracking in digital histopathology images has been an evolving topic of research in medical imaging [5–7]. The development of techniques to analyze and predict the visual attention of pathologists during the examination of WSIs is critical for developing computer-assisted training and clinical decision support systems [4]. Interpretation of the attention behavior of pathologists has been considered in the early works of [8] that conducted eye tracking studies on grading tumor architecture in prostate cancer images and [9] capturing mouse movement as a reliable indicator of attention behavior. Other works [7,10] used eye tracking to explore the complexity of diagnostic decision-making of pathologists viewing WSIs. The attention behavior of pathologists has also been collected using a web-based digital microscope and analyzed to reveal distinct scanning and drilling diagnostic search patterns [10]. Recently, the work in [4] presented ProstAttNet, a fine-tuned ResNet34 model [14], for analyzing and predicting visual attention heatmaps of WSIs from prostate cancer.

Here we propose PathAttFormer, a deep learning model based on Swin transformer [13] that is able to predict visual attention for multiple cancer types. The Swin transformer model more efficiently leverages the global contextual information across cells and nuclei regions within sub-patches of a WSI patch compared to other conventional models. This paper demonstrates the application of PathAttFormer to predict the viewing behavior of pathologists in the task of evaluating and grading Gastrointestinal Neuroendocrine tumors (GI-NETs). GI-NETs require pathologists to examine H&E WSIs and corresponding salient regions in additional tissue sections stained with Ki-67 immunohistochemistry (IHC) to grade tumors by quantifying brown colored Ki-67 positive tumor cells. In contrast, pathologists assign Gleason grades to prostate cancer based on the primary and secondary patterns of tumor growth viewed in H&E images only. Given that pathologists routinely examine multiple tissue sections, stained with H&E and other IHC biomarkers, to evaluate numerous types of cancers, this work represents a generalizable methodology that can be broadly useful in digital pathology. We depict the slide examination processes for the Prostate cancer and GI-NETs WSIs in Fig. 1.

To the best of our knowledge, we are the first to analyze attention data on GI-NET WSIs and to present a generalizable methodology to predict visual

attention for multiple cancer types. Our study also represents a strong proof of concept for collecting WSI navigation data to study visual attention in pathologists evaluating cancer without the need for specialized eye-tracking equipment. While [4] only analyzed the one-stage examination process in prostate cancer, we also study attention in a two-stage examination with different sequential tasks (tumor detection and nuclei counting) for the more complex GI-NETs, where we show (Fig. 4, Table 2) that attended WSI regions at stage 1 influence attention at the next stage. Lastly, our work is important because our framework can be used to characterize and predict pathologist visual attention across cancer types involving multiple stages of examination in real world digital pathology workflows. In clinical terms, our work can be used to develop clinical applications for training residents and fellows and for reducing observer variability among pathologists via clinical decision support.

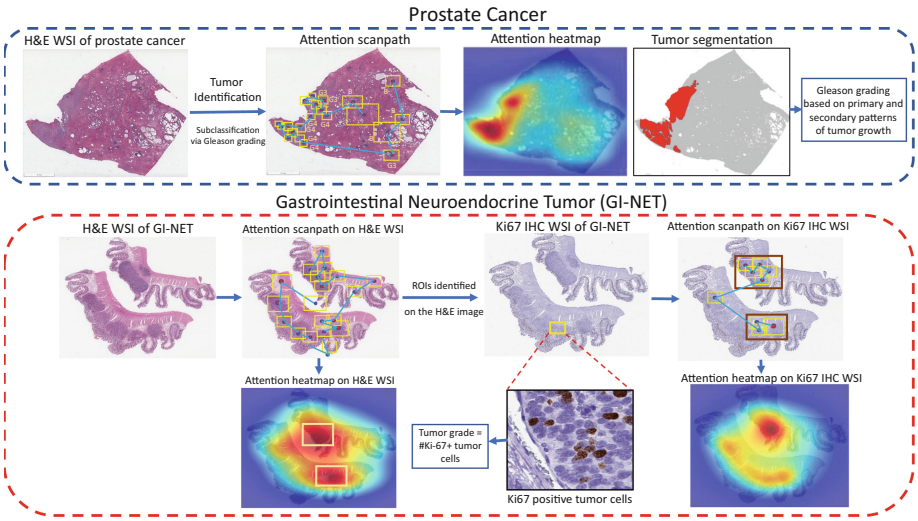

Fig. 1. Demonstration of visual attention heatmap generation by analyzing the viewing behavior of pathologists in Prostate cancer (top) and GI-NET (bottom). The yellow boxes indicate the viewport boxes and the scanpath is constructed by joining the viewport centers. Greater attention is indicated by hotter (redder) color. (Color figure online)

2 Methods

2.1 Data Collection and Processing

For Prostate cancer, we used the same dataset of 22 H&E WSIs as in [4]. Among the 22 WSIs, attention data for 5 was collected from 13 pathologists and attention data for the remaining 17 was collected from a Genitourinary (GU) specialist. The GU specialist also annotated the Gleason grades on all the 22 WSIs.

Following the procedure described in [4], we used QuIP caMicroscope [11], a web-based toolset for digital pathology, data management and visualization to record the attention data of pathologists as they viewed GI-NET WSIs [2,12]. We collected the attention data from 21 resection H&E WSIs and the corresponding 21 Ki-67 IHC WSIs. Two pathologists participated in the GI-NET attention data collection. They viewed the H&E and the Ki-67 IHC WSIs sequentially and graded the tumors. The average viewing time per slide per pathologist was 37.67 s for the H&E WSIs and 131.98 s for the Ki-67 IHC WSIs.

We process attention data in terms of attention heatmap and attention scanpath as shown in Fig. 4. The aggregate spatial distribution of the pathologist's attention is captured using the attention heatmap and the temporal information is recorded in the attention scanpath. Following [4], a value of 1 is assigned at all image pixels within a viewport and the values are summed up over all viewports to construct the attention heatmap. The final attention heatmap is obtained after map normalization as follows:

$$\begin{aligned} M'^{I}_{Attn.}(x,y) &= G^{\sigma} * \sum_{v=1}^{V} (\sum_{v_x^s}^{v_x^e} \sum_{v_y^s}^{v_y^e} 1) \\ M^{I}_{Attn.} &= \frac{M'^{I}_{Attn.} - min(M'^{I}_{Attn.})}{max(M'^{I}_{Attn.}) - min(M'^{I}_{Attn.})} \end{aligned} \tag{1}$$

where, $M'^{I}_{Attn.}$ is the intermediate attention heatmap, $M^{I}_{Attn.}$ is the final normalized attention heatmap, V is the number of viewports on a WSI I, and v_x^s, v_x^e, v_y^s, v_y^e are the starting and the ending x and y coordinates of the viewport v respectively, and G^{σ} is a 2D gaussian ($\sigma = 16$ pixels) for map smoothing. For constructing the attention scanpath, we stack the viewport centers of every viewport, v in the WSI I following [4].

2.2 Predicting Attention Heatmaps

For the Prostate cancer WSIs, we follow a two-step process for predicting the final attention heatmap. In the first step, we produce the patch-wise attention labels and assemble the patch-wise predictions to construct the attention heatmap on the WSI. In the next step, we refine the patch-wise attention predictions using a self-attention based visual attention refinement module that considers pairwise similarities between the patches to update the patch-wise attention labels. For the GI-NET WSIs, we first predict the attention heatmap on the H&E WSI similar to Prostate cancer. Next, we cascade the attention prediction modules for the H&E and Ki-67 IHC WSIs by using the patch-wise attention predictions on the H&E WSI (from Stage 1) and the Ki-67 positive nuclei detection map (discretized patch-wise similar to the H&E attention heatmap) on the Ki-67 IHC WSI as inputs (each input encoded to a 20-dimensional feature vector) to the model for predicting attention on Ki-67 IHC WSIs. We depict our attention prediction model, PathAttFormer for the two cancer types in Fig. 2 and Fig. 3.

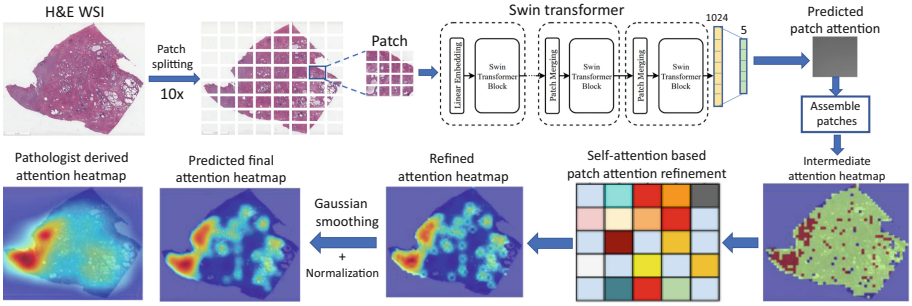

Fig. 2. Our model, PathAttFormer, for predicting visual attention in Prostate grading.

Step 1: Patch-Wise Attention Prediction. Similar to [4], we formulate attention prediction as a classification task where the aim is to classify a WSI patch into one of the N attention bins. $N = 5$ in our study, which best reconstructs the attention heatmaps. $N < 5$ leads to inaccurate reconstruction of the attention heatmap and $N > 5$ provides us minimal improvement in the reconstructed attention heatmap while reducing the accuracy of patch classification performance. During training, we discretize the average pixel intensity of every heatmap patch into an attention bin and at inference we assign the average pixel intensity of a predicted bin to the image patch to construct the intermediate patch-wise heatmap. Our model comprises of a Swin Transformer (a Swin-Base model using patch size $p = 4$, window size $w = 7$, operating on 224×224 images) [13] pre-trained on the ImageNet 1K dataset as the backbone feature extractor, followed by a fully connected layer $fc(1024, 5)$ as the classifier.

Training Details: During training PathAttFormer on the Prostate cancer WSIs, we froze the swin transformer and updated the last fully-connected layer only. We used 500×500 image patches (resized to 224×224 for training) extracted from 15 WSIs at $10\times$ magnification (the most frequent magnification used by pathologists per our analysis) for training while using 2 WSIs for validation. We performed data augmentation [16] by introducing color jitter and random horizontal and vertical image flips during training. We used the weighted Cross-Entropy loss between the predicted and the pathologist-derived attention bins. The class weight for a bin was inversely proportional to the number of training instances for the class. We only processed patches with tissue area >30% of the total patch area, which provided us with 11K H&E patches for training.

For the GI-NET WSIs, we trained separate models for the H&E and the Ki-67 IHC WSIs corresponding to patches extracted at 4X and 40X magnification (image sizes 1250×1250 and 125×125 respectively). These correspond to the most frequent magnification levels for the two slide types per our analysis. We used 9K H&E patches and 267K Ki-67 patches for training following a similar training method as Prostate WSIs. We used the AdamW optimizer [15] with an initial learning rate of 0.01. Training converged within 16 epochs with a training time of approximately 8 h on a Nvidia Titan-Xp GPU for both cancer types.

Step 2: Self-attention Based Visual Attention Refinement. We introduce a dense method for attention refinement that eliminates spatial discontinuities in the prediction caused by patch-based processing. We refine the patch-wise predictions from PathAttFormer using a self-attention (SA) based visual attention refinement module. This step enforces the spatial continuity in the predicted attention heatmap, thereby avoiding abrupt variations in the predicted attention labels caused by the absence of the contextual information. We compute the contribution of an image patch q to an image patch p as:

$$w_{q:d_{p,q}\leq d^t}(p) = \frac{\exp(-\frac{||F_p - F_q||^2}{2\hat{\alpha_1}^2} - \frac{||l_p - l_q||^2}{2\hat{\alpha_2}^2})}{\sum_{r:d_{p,r}\leq d^t} \exp(-\frac{||F_p - F_r||^2}{2\hat{\alpha_1}^2} - \frac{||l_p - l_r||^2}{2\hat{\alpha_2}^2})} \tag{2}$$

where F_p denotes the 1024-dim. feature vector encoded by the Swin Transformer and l_p denotes the location of the patch p, $d_{p,r}$ is the euclidean distance between the patch p and r, and d^t is the threshold distance. The Gaussian kernel parameters are selected as: $\hat{\alpha}_1, \hat{\alpha}_2 = \arg\min_{\alpha_1,\alpha_2} ||L^{Val}_{refined} - L^{Val}_{GT}||^2$, where $L^{Val}_{refined}$ and L^{Val}_{GT} denote the refined and the ground truth patch labels. We used grid search on our validation set [17] to find the optimal kernel parameters $\hat{\alpha_1} = 1.6$, $\hat{\alpha_2} = 1$ and threshold distance factor $d_t = 0.3 \times I_{Dg}$ (I_{Dg} = image diagonal length), respectively. Next, we update the label of the patch p based on the weights $w_q(p)$ obtained in Eq. 2 as: $L^I_{refined}{}'(p) = \sum \sum_q w_q(p) L_{coarse}(q)$, where $L^I_{refined}{}'$ is the

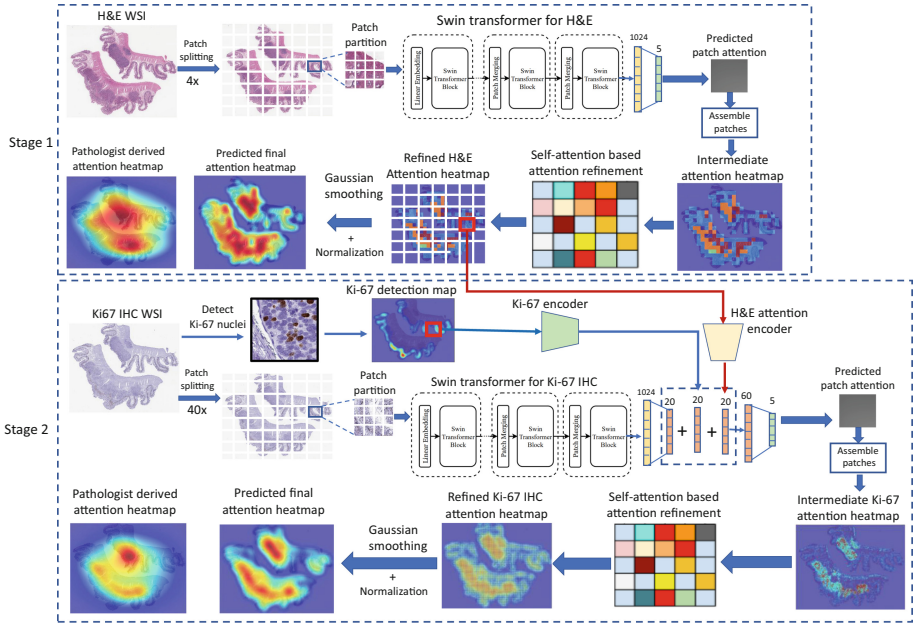

Fig. 3. Cascaded PathAttFormers for predicting visual attention in GI-NET.

refined patch label and L_{coarse} is the patch label predicted using PathAttFormer. Finally we construct the attention map $M^I_{refined}{}'$ by assembling $L^I_{refined}{}'$. The map $M^I_{refined}{}'$ is further smoothed and normalized.

In order to reduce the compute time for GI-NET WSIs, which involved a higher number of patches per WSI at 40× magnification, we computed the attention weights for alternate pixels in the image and applied bilinear interpolation to the intermediate refined attention heatmap, $M^I_{refined}{}'$ for computing the final attention heatmap, $M_{refined}$. This step reduced compute time 16 times although model complexity remained O(N^2). Average number of patches per slide for prostate and GI-NET test sets were 834 (10×) and 16.8K (40×) respectively.

3 Results

3.1 Qualitative Evaluation

Figure 4(a), row 1 shows the visual scanpath of a pathologist with the magnification at each viewport center and the attention heatmap computed from the viewport boxes for a test H&E WSI instance. We see that the pathologist mostly viewed the WSI at 10× magnification. We also compare the attention data with the tumor annotation we obtained from the GU specialist. The attention heatmap correlates well with the tumor locations in the ground truth tumor annotation.

We compare the attention heatmaps predicted by our model with 4 baseline models: (1) ResNet34 [14] and (2) Vision Transformer (ViT) [22], as the backbone feature extractor, (3) ProstAttNet [4], (4) DA-MIL [3], in Fig. 4(a). The multiple instance learning model (DA-MIL) [3] was trained on the WSIs with the primary Gleason grades as the bag labels. We also compare the predictions obtained using the proposed self-attention (SA) based attention refinement module to the predictions using Dense Conditional Random Fields [17] (CRF) as an alternative method to refine the attention heatmap. PathAttFormer produces more accurate attention heatmap compared to the baselines. Moreover, the SA module improves the overall spatial consistency in the predicted attention heatmaps compared to the patch-wise predictions from the baseline models.

In Fig. 4(b), we show the attention scanpaths with the magnification at each viewport center and the computed attention heatmaps for a H&E and Ki-67 IHC WSI instance from our GI-NET dataset. While the pathologist viewed all image regions in the H&E WSI to detect the tumor regions, as seen in the H&E scanpath, their attention on the Ki-67 IHC WSI was mostly confined within the tumor regions detected on the H&E WSI. Also, the regions examined on the Ki-67 IHC WSI are well correlated with the Ki-67 positive nuclei detection map obtained using [1]. We also compare the attention data with the tumor segmentation (on the H&E WSI) obtained from [1]. The attention heatmap correlates well with the tumor locations in the tumor segmentation map. We also compare the attention heatmaps predicted by PathAttFormer to the other models in row 3 in Fig. 4(b) for the same test Ki-67 IHC WSI instance. We observe that

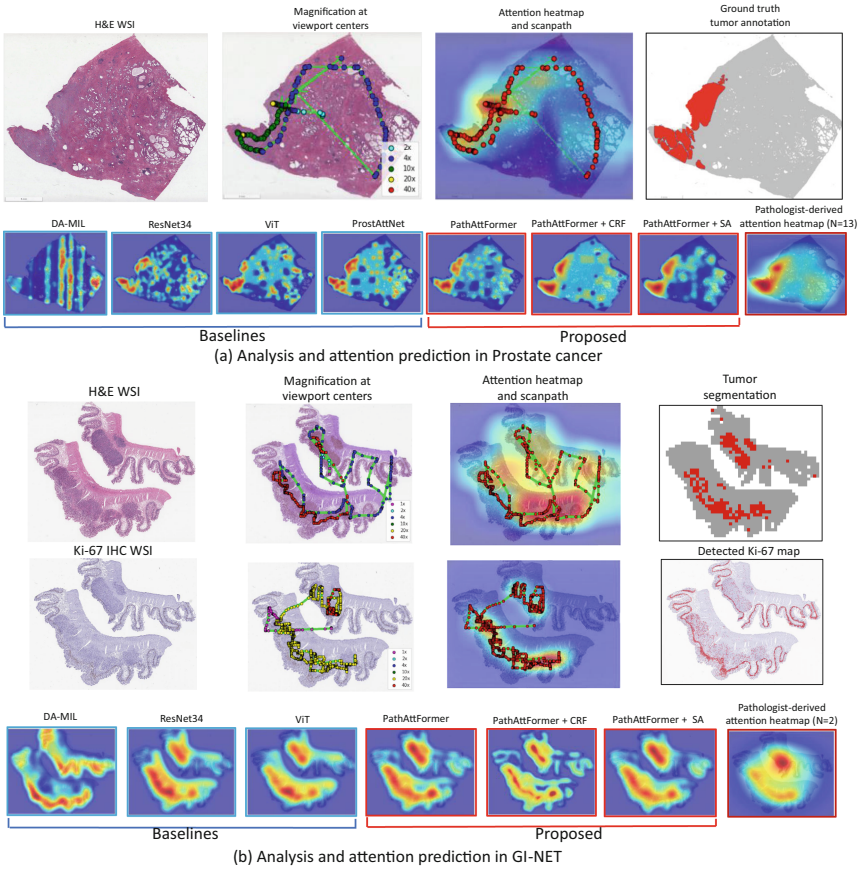

Fig. 4. Visualization of observed attention data on a test WSI of (a) Prostate cancer, and (b) GI-NET. We also compare the predicted attention heatmaps to the pathologist-derived attention heatmap (row 2 in (a) and row 3 in (b)). PathAttFormer+SA best predicts the attention data. More results in the supplementary.

PathAttFormer produces more accurate attention heatmaps compared to the baselines using ResNet34 and ViT as the backbone feature extractors, as well as the DA-MIL method. Also, the predicted attention heatmap correlates well with the corresponding tumor segmentation map.

3.2 Quantitative Evaluation

We quantitatively evaluate the model performance using four metrics: weighted F1-score of attention intensity classification, Cross Correlation (CC), Normalized Scanpath Saliency (NSS), and Information Gain (IG) [18]. A higher weighted F1-score indicates a better class-balanced classification performance. A high CC value indicates a higher correlation between the map intensities. NSS measures

the average normalized attention intensity at the viewport centers and IG measures the average information gain of the predicted attention heatmap over a center prior map at viewport centers [18,19]. To ensure that the distributions of the predicted and pathologist-derived attention heatmaps are similar, we perform histogram matching [20] of the two maps as a pre-processing step [4,21]. In Tables 1 and 2, we report the 4-fold cross-validation scores CC_{Attn}, NSS_{Attn} and IG_{Attn} between the predicted and the pathologist-derived attention heatmaps and the corresponding weighted F1-score. We also show the CC_{Seg} score between the predicted attention heatmap and the ground truth tumor segmentation map.

For Prostate WSIs (Table 1), PathAttFormer with attention refinement (SA) best predicts the attention heatmap in terms of the CC and IG metrics compared to pathologist-derived attention data, while the PathAttFormer model has the best NSS score. PathAttFormer + SA also best predicts the tumor segmentation in terms of CC. The proposed model improves performance by first using a Swin Transformer backbone instead of a convolutional model (e.g. ResNet34) for predicting attention on a WSI patch, followed by refining patch-wise predictions using an attention refinement module. PathAttFormer also outperforms ProstAttNet [4] on our test set. For GI-NET Ki-67 WSIs (Table 2), our PathAttFormer + SA model best predicts the attention heatmap in terms of all metrics compared to the pathologist-derived attention data, while the PathAttFormer model with the H&E and Ki-67 detection map as inputs best predicts the ground truth tumor segmentation in terms of the CC metric.

Table 1. Comparison of the 4-fold cross validation performance on the baseline models (blue) and the PathAttFormer models (red) for five test H&E WSIs of prostate cancer.

Model	Weighted-F1	CC_{Attn}	NSS_{Attn}	IG_{Attn}	CC_{Seg}
ResNet34 [14]	0.327 ± 0.01	0.710 ± 0.03	0.382 ± 0.01	0.978 ± 0.04	0.675 ± 0.07
ViT [22]	0.321 ± 0.01	0.706 ± 0.03	0.441 ± 0.02	0.241 ± 0.01	0.682 ± 0.07
ProstAttNet [4]	0.329 ± 0.01	0.712 ± 0.02	0.408 ± 0.01	1.046 ± 0.04	0.678 ± 0.07
DA-MIL [3]	-	0.504 ± 0.03	0.275 ± 0.03	0.042 ± 0.02	0.303 ± 0.08
PathAttFormer	0.348 ± 0.01	0.737 ± 0.02	$\mathbf{0.584 \pm 0.02}$	1.032 ± 0.04	0.681 ± 0.07
PathAttFormer+CRF	0.348 ± 0.01	0.743 ± 0.02	0.526 ± 0.02	0.702 ± 0.04	0.684 ± 0.07
PathAttFormer+SA	$\mathbf{0.348 \pm 0.01}$	$\mathbf{0.751 \pm 0.02}$	0.580 ± 0.02	$\mathbf{1.087 \pm 0.04}$	$\mathbf{0.689 \pm 0.07}$

Table 2. Comparison of the 4-fold cross validation performance on the baseline models (blue) and the PathAttFormer models (red) for five Ki-67 IHC WSIs of GI-NET.

Model	Weighted-F1	CC_{Attn}	NSS_{Attn}	IG_{Attn}	CC_{Seg}
ResNet34 [14]	0.273 ± 0.01	0.728 ± 0.05	0.514 ± 0.01	0.718 ± 0.03	0.820 ± 0.05
ViT [22]	0.270 ± 0.01	0.726 ± 0.06	0.526 ± 0.02	0.764 ± 0.03	0.809 ± 0.05
DA-MIL [3]	-	0.521 ± 0.07	0.383 ± 0.03	0.104 ± 0.02	0.692 ± 0.06
PathAttFormer	0.291 ± 0.01	0.732 ± 0.05	0.566 ± 0.02	0.758 ± 0.03	0.819 ± 0.04
PathAttFormer (w/ H&E attn.)	0.291 ± 0.01	0.741 ± 0.06	0.568 ± 0.02	0.763 ± 0.03	0.827 ± 0.04
PathAttFormer (w/ H&E attn.+Ki-67)	0.291 ± 0.01	0.744 ± 0.06	0.562 ± 0.02	0.771 ± 0.04	$\mathbf{0.835 \pm 0.04}$
PathAttFormer+CRF (w/ H&E attn.+Ki-67)	0.291 ± 0.01	0.758 ± 0.06	0.565 ± 0.02	0.479 ± 0.03	0.826 ± 0.04
PathAttFormer+SA (w/ H&E attn.+Ki-67)	$\mathbf{0.291 \pm 0.01}$	$\mathbf{0.762 \pm 0.06}$	$\mathbf{0.573 \pm 0.02}$	$\mathbf{0.802 \pm 0.04}$	0.834 ± 0.04

4 Conclusion

We have shown how pathologists allocate attention while viewing prostate cancer and GI-NET WSIs for tumor grading and presented a generalizable deep learning model that predicts visual attention on WSIs. Our work forms the foundation for research on tracking and analysing the attention behavior of pathologists viewing multiple stained images in sequence in order to grade the tumor type. In the future, we will collect attention data in a larger study with more WSIs in order to improve our attention prediction model. Additionally, we aim at predicting the attention scanpaths of pathologists that can reveal insights about the spatio-temporal dynamics of viewing behavior.

References

1. Govind, D., et al.: Improving the accuracy of gastrointestinal neuroendocrine tumor grading with deep learning. Sci. Rep. **10**(1), 1–12 (2020)
2. Matsukuma, K., Olson, K.A., Gui, D., Gandour-Edwards, R., Li, Y., Beckett, L.: Synaptophysin-Ki-67 double stain: a novel technique that improves interobserver agreement in the grading of well-differentiated gastrointestinal neuroendocrine tumors. Mod. Pathol. **30**(4), 620–629 (2017)
3. Hashimoto, N., et al.: Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3852–3861 (2020)
4. Chakraborty, S., et al.: Visual attention analysis of pathologists examining whole slide images of Prostate cancer. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE, March 2022
5. Brunyé, T.T., Drew, T., Kerr, K.F., Shucard, H., Weaver, D.L., Elmore, J.G.: Eye tracking reveals expertise-related differences in the time-course of medical image inspection and diagnosis. J. Med. Imaging **7**(5), 051203 (2020)
6. Sudin, E., et al.: Eye tracking in digital pathology: identifying expert and novice patterns in visual search behaviour. In: Medical Imaging 2021: Digital Pathology, vol. 11603, p. 116030Z. International Society for Optics and Photonics, February 2021
7. Brunyé, T.T., Mercan, E., Weaver, D.L., Elmore, J.G.: Accuracy is in the eyes of the pathologist: the visual interpretive process and diagnostic accuracy with digital whole slide images. J. Biomed. Inform. **66**, 171–179 (2017)
8. Bombari, D., Mora, B., Schaefer, S.C., Mast, F.W., Lehr, H.A.: What was I thinking? Eye-tracking experiments underscore the bias that architecture exerts on nuclear grading in prostate cancer. PLoS ONE **7**(5), e38023 (2012)
9. Raghunath, V., et al.: Mouse cursor movement and eye tracking data as an indicator of pathologists' attention when viewing digital whole slide images. J. Pathol. Inform. **3**, 43 (2012)
10. Mercan, E., Shapiro, L.G., Brunyé, T.T., Weaver, D.L., Elmore, J.G.: Characterizing diagnostic search patterns in digital breast pathology: scanners and drillers. J. Digit. Imaging **31**(1), 32–41 (2018)
11. Saltz, J., et al.: A containerized software system for generation, management, and exploration of features from whole slide tissue images. Can. Res. **77**(21), e79–e82 (2017)

12. Govind, D., et al.: Improving the accuracy of gastrointestinal neuroendocrine tumor grading with deep learning. Sci. Rep. **10**(1), 1–12 (2020)
13. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
16. Tellez, D., et al.: Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. IEEE Trans. Med. Imaging **37**(9), 2126–2136 (2018)
17. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFS with gaussian edge potentials. In: Advances in Neural Information Processing Systems, vol. 24 (2011)
18. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Trans. Pattern Anal. Mach. Intell. **41**(3), 740–757 (2018)
19. Kümmerer, M., Theis, L., Bethge, M.: Deep gaze I: boosting saliency prediction with feature maps trained on imagenet. arXiv preprint arXiv:1411.1045 (2014)
20. Gonzales, R.C., Fittes, B.A.: Gray-level transformations for interactive image enhancement. Mech. Mach. Theory **12**(1), 111–122 (1977)
21. Peacock, C.E., Hayes, T.R., Henderson, J.M.: Center bias does not account for the advantage of meaning over salience in attentional guidance during scene viewing. Front. Psychol. **11**, 1877 (2020)
22. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)