Predicting Clinician Fixations on Glaucoma OCT Reports via CNN-Based Saliency Prediction Methods

Mingyang Zang, *Student Member, IEEE*, Pooja Mukund, *Student Member, IEEE*, Britney Forsyth, *Student Member, IEEE*, Andrew F. Laine, and Kaveri A. Thakoor, *Member, IEEE*

***Abstract*—*Goal:* To predict physician fixations specifically on ophthalmology optical coherence tomography (OCT) reports from eye tracking data using CNN based saliency prediction methods in order to aid in the education of ophthalmologists and ophthalmologists-in-training. *Methods:* Fifteen ophthalmologists were recruited to each examine 20 randomly selected OCT reports and evaluate the likelihood of glaucoma for each report on a scale of 0-100. Eye movements were collected using a Pupil Labs Core eye-tracker. Fixation heat maps were generated using fixation data. *Results:* A model trained with traditional saliency mapping resulted in a correlation coefficient (CC) value of 0.208, a Normalized Scanpath Saliency (NSS) value of 0.8172, a Kullback–Leibler (KLD) value of 2.573, and a Structural Similarity Index (SSIM) of 0.169. *Conclusions*: The TranSalNet model was able to predict fixations within certain regions of the OCT report with reasonable accuracy, but more data is needed to improve model accuracy. Future steps include increasing data collection, improving quality of data, and modifying the model architecture.**

***Index Terms*—Deep learning, optical coherence tomography, saliency prediction.**

***Impact Statement*—Reliable prediction of physician fixations could aid in teaching physicians-in-training and AI systems how to most efficiently evaluate OCT reports and identify regions of interest.**

I. Introduction

In A clinical setting, eye movement analysis can be used to measure the position, location, and duration of human gaze on medical images. Many studies have used eye tracking to examine differences in how radiologists with varying levels of experience interpret medical images [1], [2], [3], [4]. For example, eye tracking has been utilized in the field of mammography to show performance of expert clinicians in diagnosing breast cancer is much higher than that of clinicians-in-training [5].

Eye tracking software can measure the eye movements of individuals to generate fixation maps. These fixation maps measure the location and duration of human gaze on regions of interest in an image. Typically, in medical image segmentation tasks, regions-of-interest are hand-labeled by physicians, which is an extremely time-consuming process to collect the amount of data needed to train segmentation models. However, gaze data from eye tracking software can also be used to generate annotated data for deep learning training [6], [7], [15].

Eye-tracking processes can further be analyzed with deep learning algorithms and technologies. Specifically, visual saliency reflects the extent to which the content in an image attracts visual attention. The field of saliency prediction relies on collecting fixation data to generate ground truth saliency maps which highlight regions that attract the most visual attention from the viewer. Predicting visual saliency has been furthered by the use of convolutional neural networks (CNNs) and the availability of large-scale saliency prediction datasets [6]. Many recent studies have utilized CNNs for saliency prediction with promising results; however, many of these studies have used public benchmark datasets with generic images of animals, road signs, and humans.

This same methodology can be applied to medical image data sets used to identify regions of interest in medical diagnostic images, but limited saliency prediction studies have actually been utilized for medical applications. One example of a medical saliency prediction application was in the field of mammography to predict radiologists' visual attention on mammograms to assist in diagnosis [7]. Four hundred scans were collected and used to indicate initial perception of cancer, and the results showed that 57% of the cancer was fixated in the first second of viewing [18]. However, similar methodologies have not been applied to the field of ophthalmology. Here, we seek to develop a medical-saliency prediction model for glaucoma diagnosis on OCT reports.

Glaucoma is among the leading causes of irreversible blindness in the world; this disease causes damage to the optic nerve which can lead to gradual vision loss and ultimately complete blindness if not diagnosed in a timely manner [8]. There are two main forms of imaging used to diagnose glaucoma: fundus

Manuscript received 5 January 2024; revised 26 January 2024 and 14 February 2024; accepted 15 February 2024. Date of publication 20 February 2024; date of current version 12 March 2024. This work was supported by the Columbia University Department of Ophthalmology from Research to Prevent Blindness, Inc., New York, NY USA. The review of this article was arranged by Editor Yasemin M. Akay. *(Mingyang Zang and Pooja Mukund contributed equally to this work.) (Corresponding author: Mingyang Zang.)*

The authors are with Columbia University, New York, NY 10027 USA (e-mail: mf20080144@gmail.com).

Digital Object Identifier 10.1109/OJEMB.2024.3367492

photography and optical coherence tomography (OCT). Fundus photography has limitations. For example, factors such as miosis, corneal opacity, and the opacity of the ocular media can impact ophthalmic imaging and are associated with unclear fundus photographs. Furthermore, optic nerve findings in fundus photography are shown in a two-dimensional manner. These combined limitations have been shown to result in decreased diagnostic efficiency in physicians with less training [5], [6]. OCT is considered a reliable alternative to fundus photography that provides 3D structural information of the retina and has improved the accuracy of glaucoma detection; however, there are currently few agreed-upon guidelines for glaucoma diagnosis using OCT, and results may differ depending on a given physicians' interpretation [9]. We believe analyzing the OCT scans can improve clinical diagnosis of glaucoma.

Current research in salient detection is focused on two tasks: saliency prediction and salient object detection. Saliency prediction predicts the possibility of the human eyes to stay in a certain position in the scene, while salient object detection detects the object as a whole, similar to an image segmentation task. We sought to determine if a saliency prediction model could be implemented to create binary masks of salient vs. non-salient regions in OCT images, such as those often utilized in image segmentation tasks. Image segmentation refers to classifying regions of interest in an image which have similar properties. In our use case, the ophthalmic OCT report used to diagnose glaucoma is composed of 7 different sub-images, so regions-of-interest are actually composed of different features depending on the sub-image. Thus, we wanted to evaluate if we could leverage saliency prediction methods to classify regions of interest, based on learning medically-salient regions fixated most by ophthalmic experts (i.e., that captured their medical visual attention) when they viewed OCT reports.

In this study, we utilized a CNN-based saliency prediction model to predict binary and continuous saliency maps on a custom dataset of optical coherence tomography (OCT) reports used for glaucoma diagnosis. After reliable prediction, these saliency maps can be used to educate clinicians-in-training to analyze OCT reports as efficiently and reliably as a trained clinician. Furthermore, predicted fixation maps can aid in the diagnosis of glaucoma by informing clinicians about various regions of interest implicated in glaucoma. Once salient OCT-report regions are predicted, these patterns could also be used to train self-supervised deep learning algorithms to distinguish between healthy and diseased image classes without the need for as many costly expert-provided labels. This model could even work in tandem with pure glaucoma classification models to provide interpretability to clinicians to encourage trust in AI algorithms. Lastly, this study contributes the novelty of using a custom medical-imaging dataset with a clinically significant application to other work in the saliency prediction field, which has predominantly used large publicly available natural-image datasets such as MIT300 and SALICON [16], [17].

II. Materials and Methods

A. Dataset

Eye tracking data was acquired from Columbia University Irving Medical Center's (CUIMC) Harkness Eye Institute. Ophthalmologists of varying expertise/tenure from CUIMC were asked to view full OCT reports. Tenure of ophthalmologists ranged from residents with 10 months of training to glaucoma fellows (3+ years of experience) and faculty (30+ years of experience). The OCT images, as seen in Fig. 4 column 1, are composed of a retinal nerve fiber layer (RNFL) probability map, a retinal ganglion cell inner plexiform layer (RGCP) probability map (both overlaid with visual field points), RGCP thickness map, a RNFL thickness map, and a circumpapillary RNFL b-scan. The image dataset consisted of 185 OCT reports collected using a Topcon Atlantis machine (from a subset of patients who visited the CUIMC Harkness Eye Institute between 2010 and 2023), with 121 reports being glaucomatous and 64 reports being healthy. The OCT reports varied in difficulty from clear glaucoma/healthy images to suspect cases (possibly exhibiting myopia or optic neuropathies mimicking glaucoma), thus ensuring our eye tracking dataset consisted of gaze on straightforward as well as challenging OCT reports. In order to collect fixation data, clinicians were fitted with a Pupil Labs Core head-mounted eye tracker while viewing the OCT reports. After viewing 20 randomly selected OCT reports, each of the fifteen clinicians was immediately asked to report whether or not the patient had glaucoma on a scale of 0-100, with 0 indicating definitely healthy and 100 indicating definitely glaucoma. This study, AAAU4079, was approved by the Columbia University Irving Medical Center Institutional Review Board on December 23, 2022 and is in accordance with the tenets set forth by the Declaration of Helsinki. Informed consent was received from all study participants.

B. Data Pre-Processing

For the saliency prediction, a TranSalNet model was utilized [10]. The model takes in three inputs: original stimuli image, fixation map, and saliency map. The original stimuli was the full OCT report. To produce the fixation map, we plotted the fixations as white scatter plot points using corresponding (x, y) fixation coordinates on a black background sized identically to the original OCT image shown.

For generating the saliency map input to the model, we attempted 2 approaches: the first was a traditional saliency map via a gaussian-kernel with a mean of 63 pixels and a standard deviation of 7 pixels applied at all fixation points weighted by fixation duration. The second approach employed the Pygaze software to overlay a heatmap on the original stimuli (OCT report) showing regions where clinicians fixated most by convolving those x, y locations with a gaussian kernel proportional to the length of fixation duration with a mean of 200 pixels and a standard deviation of 33 pixels [19]. Of note, the parameters were chosen based on the default value of the Pygaze software [19], and no

significant change of result was found if they were changed. From there, the binary saliency map was created by setting all pixels with duration greater than 0 to 1 and all pixels with no corresponding duration data set to 0. Upon visual inspection, two participants with extremely noisy data were excluded. In addition, fixations that appeared in the whitespaces of the original stimuli image and not on any of the 7 sub-images were removed, as they were determined not to be meaningful. After following this exclusion criteria, we utilized data from 12 participants, resulting in a total of 216 saliency maps and stimuli images which were then split into a 5-fold cross-validation with 20% holdout. Data augmentation was performed to add variability to the sample; augmentations included horizontal image flipping, image rotation with a maximum angle of 30 degrees, and color changes. Additionally, input images were resized to 384 × 288 to as required by the TranSalNet model architecture.

C. Model Architecture & Training

The model chosen for saliency prediction was the TranSalNet Res-Net model (architecture shown in Fig. 1). TranSalNet was chosen because of its superior performance on public saliency datasets, compared to other benchmark models. It achieves the best performance on all perception-based metrics in both MIT1003 and CAT2000. Its increased performance is attributed to the addition of transformer-encoders in the typical CNN architecture for saliency prediction. The TranSalNet utilizes a basic convolutional neural network (CNN) backbone (ResNet-50 pretrained on ImageNet [20]) with additional transformers to generate saliency predictions from original stimuli and fixation data. The TranSalNet model also utilizes a linear combination of four loss metrics typically used for saliency prediction and found to increase performance. These loss functions include Kullback–Leibler (KLDiv) divergence, Normalized Scanpath Saliency (NSS), Correlation Coefficient (CC), and a Structural Similarity Metric. The model was trained using a learning rate of 1e-5 and Adam optimizer. The epoch with the lowest validation loss was kept and used to generate predictions. Predictions obtained were produced as grayscale and thresholded to produce the binary saliency mask.

III. Results

A. Objective Function/Medically-Salient Region Predictions – Quantitative

The model was trained using a linear combination of saliency prediction loss metrics as defined in the TranSalNet paper. The training loss for the binary mask approach was found to decrease over epochs, but validation loss started increasing after 2 epochs. While training loss improved, validation loss exhibited severe overfitting.

Fig. 3 shows that the model using a traditional saliency map started to overfit after 20 epochs. Although training and validation loss decreased over more epochs, absolute training loss achieved a lower value (12) for the binary saliency map, shown in Fig. 2.

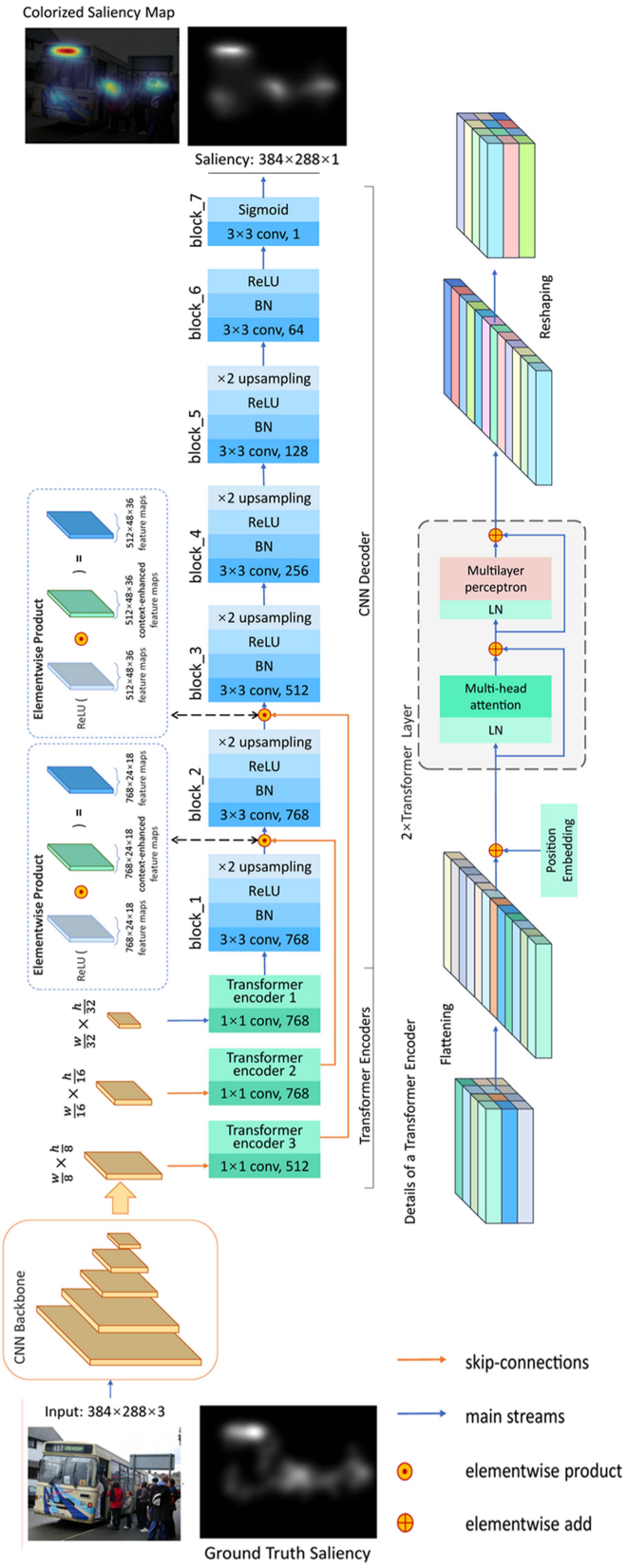

Fig. 1. TransalNet architecture.

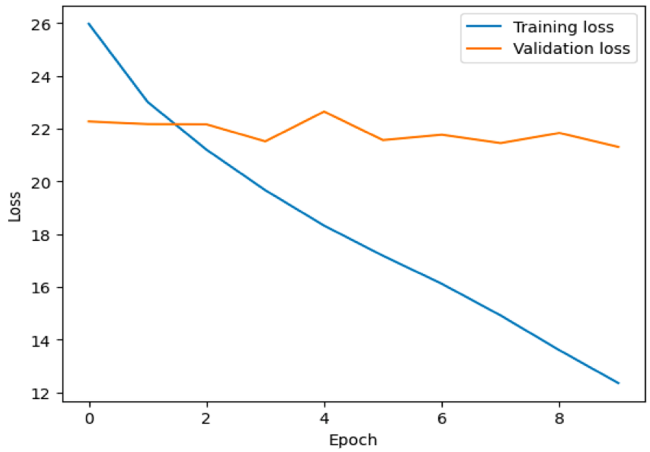

Fig. 2. Loss curve for binary mask approach.

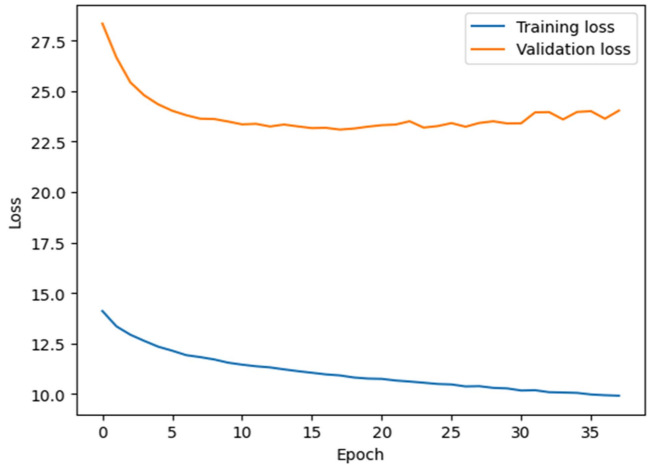

Fig. 3. Loss curve for traditional saliency map approach.

B. Medically-Salient Region Predictions – Qualitative

1) Binary Mask: The predictions from the binary mask output model, shown in Fig. 4, show that fixations are being predicted in the correct OCT report regions when compared to ground truth fixation/saliency maps. However, the specificity of shape and exact location of fixations still needs to be improved.

2) Traditional Saliency Mask: The predictions in Fig. 4(c) show examples of good predictions via use of the traditional saliency map. The locations of different fixations are well-captured. However, Fig. 4(d) shows a suboptimal prediction in which the boundaries of each predicted region of interest are poorly-defined.

C. Saliency Predictions – Quantitative

Several evaluation metrics were calculated to evaluate image similarity between model predictions and ground truth saliency masks. Table I shows the results of the validation metrics on test predictions when compared to the ground truth masks for both binary saliency mask and traditional saliency mask predictions. The continuous (traditional) saliency mask outperforms the binary saliency mask model predictions for all four metrics.

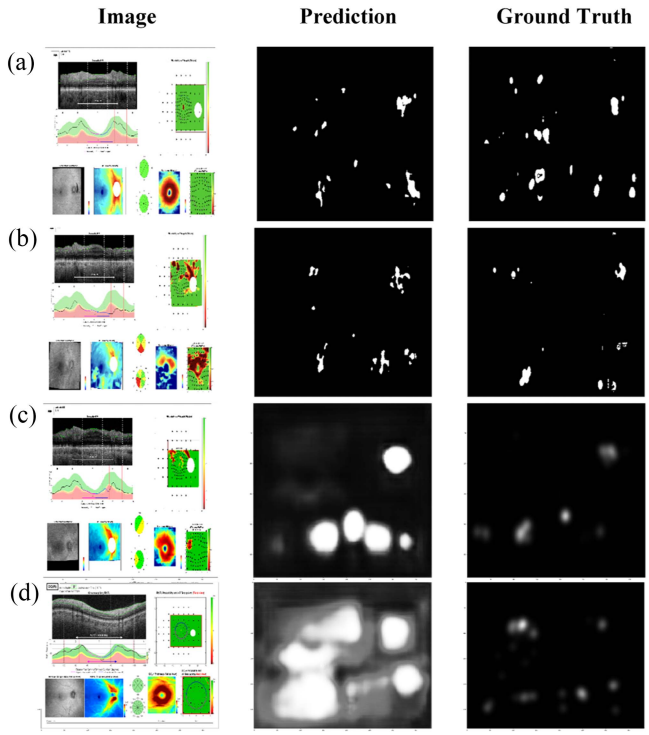

Fig. 4. (a) and (b) Predictions from the binary mask model vs. Ground Truth (c) Good predictions via use of the traditional saliency map (d) Suboptimal predictions via use of the traditional saliency map.

TABLE I
PERFORMANCE METRICS OF 5-FOLD CROSS VALIDATION OF OUR MODEL

Metric	Traditional Saliency Mask Score	Binary Saliency Mask Score
Correlation Coefficient (CC) ↑	0.410(0.153)|0.208	0.22(0.098)| 0.0518
Normalized Scanpath Saliency (NSS) ↑	2.118(0.767)|0.8172	2.095(1.721)| 0.4489
KL Divergence (KLD) ↓	2.312(0.241)|2.573	21.096(4.193)|25.989
Structural Similarity Index (SSIM) ↑	0.218(0.03)|0.169	0.16(0.075)| 0.0396

Note: Values shown are Validation mean (Validation standard deviation) | Test result

Metrics are significantly worse for the binary mask model. However, the TranSalNet model significantly outperforms our saliency mask model in regards to all four metrics with a CC value of 0.901, an NSS value of 1.998, a KLD value of 0.414 and an SSIM of 0.796. Note: The metrics achieved by the original TranSalNet model are merely for value reference as the original problem is totally different (predicting saliency on natural images rather than medical images). We believe such a difference could be a source of the large gap between their results and ours. The original problem of TranSalNet is testing the function of the transformer encoder via training on public datasets

of natural images, and they compare their results with other models. Hence, the ground-truth saliency maps of the images are well-defined, and most of these regions are object-based. In the results on a previous retinal dataset of fundus images [21], CC value between the visualized heat maps and ground-truth ophthalmologist attention maps was 0.33 and 0.14 for correct and incorrect glaucoma classification, respectively [21]. Thus, our work is the first to our knowledge that investigates expert attention vs. AI-segmented heatmaps on OCT reports instead of retinal fundus images.

IV. Discussion

A. Qualitative Analysis

Saliency predictions were generated using the TranSalNet model for fixation prediction on glaucoma OCT reports with reasonable qualitative results but suboptimal quantitative results. The continuous (traditional) saliency mask model outperformed the binary mask model in terms of quantitative results. The slightly better qualitative results for the binary saliency mask compared to the traditional saliency mask suggest that the binary saliency mask may be a better starting input for medically-salient region prediction compared to the traditional (continuous) saliency mask typically used for saliency prediction tasks for natural images.

Qualitative results for the binary mask showed that the predictions were able to adequately predict fixations on the correct sub-images. Very few predictions had fixations on sub-images that were not fixated on in the ground-truth image. Most of the predicted fixations for the good traditional and binary saliency map predictions (as shown in Fig. 4(a)–(c) above) were focused on the RNFL and RGCP probability maps and visual field (VF) test points overlaid on the probability maps as well as the RNFL and RGCP thickness maps, which is in line with the physicians' written responses (regarding what features they were using to make their decisions) we obtained during the experiment. This indicates that the model is generally predicting fixations on the correct regions. The predicted fixations also seem to more closely match the ground truth predictions in the circumpapillary RNFL and RNFL probability map and VF test point sub-images. This is likely due to the fact that physicians spent more time fixating on these regions, so the model was trained better for those regions

B. Quantitative Analysis

Quantitative results showed that both the binary mask model and continuous (traditional) saliency mask model performed poorly when compared to results of public saliency dataset such as SALICON. Of note, these datasets consist of natural objects and relatively well-defined ground-truth while defining the ground-truth for our task could be difficult since every participant had a unique pattern of observing the OCT report, meaning that the process is stochastic. However, our results of traditional saliency map were comparable to results of other medical AI studies exploring expert gaze on retinal images, such as glaucoma detection using CNN models in terms of CC value. Medical images are different from natural images in several aspects: (1) regions of interest can be smaller when compared to the salient objects in natural images [22]; (2) local details, such as lesion/tissue density and patterns can be important for feature extraction [23]; (3) the original resolution is generally much higher in medical images; thus, the aggressive downsampling needed to fit the TranSalNet input size constraints could result in lost pixel-level information.

The low scores for both approaches may indicate that the structure of the predictions matches the ground truth, but they are not capturing specific details such as exact location and shape of the original fixations. For example, when we compare the ground truth with the predictions, we can see that fixations that are being predicted are being predicted in the correct subimages in the OCT report. The predictions using the traditional saliency map capture the circular shapes in the OCT image accurately, especially if they are in RGB color. However, when we compare the predictions from the ground truth at a pixel-by-pixel level, we are still not accurately capturing the shape and exact location of the fixations, which explains the low quantitative metrics. This indicates the model is capturing the general locations of physician gaze but not yet the specific regions of interest within a sub-image. This motivates the development of new 'medical-saliency' metrics that differ from conventional natural image saliency metrics.

Another possible explanation for our lower quantitative performance could be that OCT reports are evaluated differently than the natural images typically seen in visual saliency datasets. For instance, natural objects are given different saliency or fixation weights via shape, contour and color; however, shape and contour may not be the key components of determining regions of interest in medical images. The exact contour or shape might not be essential for the clinical use case; rather, variations in anatomy within similar OCT-report sub-images across patients may be more salient. Additionally, a very high KLD could be the result of mis-detections, as KLD is very sensitive to that [24]. We had a comparable CC value to the result of the glaucoma classification task in past work [21], which suggests that our model prediction distribution had a reasonably good linear correlation with the ground truth distribution. NSS can be interpreted as a discrete approximation of CC since it operated on the fixations instead of the continuous saliency mapping [24]. SSIM directly compares the histogram of prediction and ground truth [24]. Our model generally predicts more positive regions (white region) in both methods. Hence, the histogram of our prediction can be more skewed towards the low pixel intensity.

C. Future Directions

Due to the fact that our approach is one of the first that attempts to predict clinician fixations on glaucoma OCT reports via CNN-based deep learning models, many aspects of the study could be improved and enhanced. Further work needs to be done to improve the quantity and quality of dataset, model capacity, generalization capability, and to mitigate overfitting. The TranSalNet paper, as well as many other state-of-the-art

CNN-based saliency prediction models, were trained and evaluated using the SALICON dataset which contains over 10,000 training, 5,000 validation, and 5,000 testing natural images, which is significantly greater than the sample size of our custom dataset. The sheer size of the dataset could introduce variability to the model and prevent overfitting. Our data set does not compare in size and variability to benchmark datasets. Collecting more data will likely help with increasing model generalization performance. At the time the model was trained, only a total of 12 participants were used, and many were qualitatively evaluated to show a significant amount of noise. This is likely due to calibration issues faced when fitting clinicians to the eye tracker. Alternatively, in future work, we could try to use a TranSalNet with its transformer encoder pretrained on a large corpus of eye movements on natural images (SALICON, etc.) followed by fine-tuning on our medical eye movement dataset to potentially improve performance.

The PupilCore eye tracker is limited in its ability to adjust for different face structures. The world camera could only be pushed towards the face and away from the face a couple centimeters which led to calibration issues for many individuals. Future work will involve collecting eye tracking data with the 250 Hz Tobii Pro Fusion eye tracker, which has higher precision and higher frequency than the 200 Hz Pupil Core device. Our dataset contained eye movements from 9 residents and 6 faculty/fellows. Of note, we trained a separate model only on the data from the 6 faculty/fellows and found that performance was comparable. As we collect more data, we will be able to evaluate the impact on performance of using more data from just the more senior expertise level. Another approach which may help to improve the quantitative pixel-by-pixel accuracy of our model could be to ask each clinician to evaluate the 7 OCT sub-images in sequence instead of the whole report, which can increase the resolution of each figure significantly and provide more detailed information for model training.

In terms of model selection, CNN models pretrained on a large number of OCT reports could speed our training and improve generalizability [13]. Models like the TranSalNet aim to capture/learn saliency predominantly from spatial features, such as sharp-edges or blob-like structures from the encoder. However, this might not fit well with our dataset, and methods utilizing pixel-wise regression directly have been shown to produce improved performance over state-of-the-art saliency methods via a customized objective function [14]. Probability distribution-based loss functions could be more ideal to train our dataset as the stochastic nature of evaluating OCT reports would be emphasized more than the saliency of the exact object contour or shape [14]. Hence, a simpler architecture consisting of a CNN backbone and a few convolutional layers (without transformers and a self-attention mechanism) could be sufficient to model the visual attention process in our clinical setting.

V. Conclusion

This paper presents a novel approach for aiding ophthalmologists and ophthalmologists-in-training in efficiently evaluating OCT reports and identifying regions of interest via using salient regions produced by CNNs to replace the need for many hand labels for similar tasks. The training dataset segmentation labels were created via two different approaches and trained separately. The traditional saliency map was generated via a gaussian-kernel applied at all fixation points weighted by fixation duration. The binary saliency map set image pixels (after gaussian convolution) larger than 0 to 1. The former approach has a higher score of CC, NSS, KLS, and SSIM. However, both results suggest that the locations of different fixations were well-captured, and few predictions had fixations that were not fixated-on in the ground-truth image. Better qualitative results, but lower metrics score of the binary saliency approach could be a result of the chosen objective function and metrics. To our best knowledge, no prior work has been done on utilizing saliency mapping to assist glaucoma OCT report prediction. We believe a larger OCT report dataset, customized objective function, pretrained models being trained and well-tuned in similar dataset, and reducing intra and inter subject noise can finalize the model to reach a level of a trained ophthalmologist reliably; at which point the model predictions can be used to generate self-supervised input data for other glaucoma prediction AI tools that can detect disease using inherent differences in predicted eye movement patterns, without requiring physicians to spend time looking at images or providing labels.

Author Contributions

Mingyang Zang: idea conception, writing – original draft and revision, data preprocessing pipeline, coding. **Pooja Mukund**: idea conception, writing – original draft, data preprocessing pipeline, coding, experiments. **Britney Forsyth**: data preprocessing pipeline, coding, experiments. **Andrew F. Laine**: supervision, feedback. **Kaveri A. Thakoor**: idea conception, supervision, feedback, revision. All authors provided critical feedback and helped shape the manuscript.

Conflicts of Interest

Kaveri A. Thakoor receives research funding from Topcon Healthcare.

Acknowledgment

The authors are grateful to George A. Cioffi, Jeffrey M. Liebmann, and Royce W.S. Chen for valuable clinical insights. The authors also wish to thank Ari Leshno for help with acquiring the OCT datasets and ground truth labels used in this study and Ryan Zukerman and Sophie Gu for help with recruiting resident participants for this study.

References

[1] C. F. Nodine et al., "Nature of expertise in searching mammograms for breast masses," *Academic Radiol.*, vol. 3, no. 12, pp. 1000–1006, 1996.

[2] C. F. Nodine and E. A. Krupinski, "Perceptual skill, radiology expertise, and visual test performance with NINA and WALDO," *Academic Radiol.*, vol. 5, no. 9, pp. 603–612, 1998.

[3] E. A. Krupinski et al., "Eye-movement study and human performance using telepathology virtual slides: Implications for medical education and differences with experience," *Hum. Pathol.*, vol. 37, no. 12, pp. 1543–1556, Dec. 2006.

[4] S.Voisin G.Tourassi, V Paquit, and E Krupinski, "Investigating the link between radiologists' gaze, diagnostic decision, and image content," *J. Amer. Med. Inform. Assoc.*, vol. 20, no. 6, pp. 1067–1075, Nov. 2013.

[5] C. F. Nodine, C. Mello-Thoms, H. L. Kundel, and S. P. Weinstein, "Time course of perception and decision making during mammographic interpretation," *Amer. J. roentgenol.*, vol. 179, no. 4, pp. 917–923, 2002.

[6] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1072–1080.

[7] L. Lévêque, H. Bosmans, L. Cockmartin, and H. Liu, "State of the art: Eye-tracking studies in medical imaging," *IEEE Access*, vol. 6, pp. 37023–37034, Jun. 2018.

[8] W. Lu, Y. Tong, Y. Yu, Y. Xing, C. Chen, and Y. Shen, "Applications of artificial intelligence in ophthalmology: General overview," *J. Ophthalmol.*, vol. 2018, 2018, Art. no. 5278196.

[9] L. S. Abrams, I. U. Scott, G. L. Spaeth, H.A. Quigley, and R. Varma, "Agreement among optometrists, ophthalmologists, and residents in evaluating the optic disc for glaucoma," *Ophthalmology*, vol. 101, pp. 1662–1667, 1994.

[10] J. D. Rossetto, L. A. S. Melo Jr., M. S. Campos, and I. M Tavares, "Agreement on the evaluation of glaucomatous optic nerve head findings by ophthalmology residents and a glaucoma specialist," *Clin. Ophthalmol.*, vol. 11, pp. 1281–1284, 2017.

[11] A.Y. Maa et al, "The impact of OCT on diagnostic accuracy of the technology-based eye care Services protocol: Part II of the technology-based eye care Services Compare trial," *Ophthalmology*, vol. 127, pp. 544–549, 2020.

[12] J. Lou, H. Lin, D. Marshall, D. Saupe, and H. Liu, "TranSalNet: Towards perceptually relevant visual saliency prediction," *Neurocomputing*, vol. 494, pp. 455–467, 2022.

[13] N. Motozawa et al, "Optical coherence tomography-based deep-learning models for classifying normal and age-related macular degeneration and exudative and non-exudative age-related macular degeneration changes," *Ophthalmol. Ther.*, vol. 8, pp. 527–539, 2019.

[14] S. Jetley, N. Murray, and E. Vig, "End-to-end saliency mapping via probability distribution prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5753–5761.

[15] J. N. Stember et al., "Eye tracking for deep learning segmentation using convolutional neural networks," *J. Digit. Imag.*, vol. 32, no. 4, pp. 597–604, 2019.

[16] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, Mar. 2019.

[17] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1072–1080.

[18] H. L. Kundel, C. F. Nodine, E. A. Krupinski, and C. Mello-Thoms, "Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms," *Academic Radiol.*, vol. 15, no. 7, pp. 881–886, 2008.

[19] E. S. Dalmaijer, S. Mathôt, and S V. d. Stigchel, "PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eye-tracking experiments," *Behav. Res. methods*, vol. 46, pp. 913–921, 2014.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[21] L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu, "Attention based glaucoma detection: A large-scale database and CNN model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10563–10572.

[22] Y. Shen et al., "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," *Med. Image Anal.*, vol. 68, pp. 101908, 2021.

[23] J. Wei et al., "Association of computerized mammographic parenchymal pattern measure with breast cancer risk: A pilot case-control study," *Radiology*, vol. 260, no. 1, pp. 42–49, 2011.

[24] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, Mar. 2019.