Using Expert Gaze for Self-Supervised and Supervised Contrastive Learning of Glaucoma from OCT Data

Wai Tak Lau MICHAEL.LAU2@COLUMBIA.EDU
Ye Tian YT2793@COLUMBIA.EDU
Roshan Kenia RK3291@COLUMBIA.EDU
Saanvi Aima SA4166@COLUMBIA.EDU
Kaveri A. Thakoor K.THAKOOR@COLUMBIA.EDU
Columbia University, United States

Abstract

In this work, we address the challenge of limited data availability common in healthcare settings by using clinician (ophthalmologist) gaze data on optical coherence tomography (OCT) report images as they diagnose glaucoma, a top cause of irreversible blindness worldwide. We directly learn gaze representations with our 'GazeFormerMD' model to generate pseudo-labels using a novel multi-task objective, combining triplet and cross-entropy losses. We use these pseudo-labels for weakly supervised contrastive learning (WSupCon) to detect glaucoma from a partially-labeled dataset of OCT report images. Our natural-language-inspired region-based-encoding GazeFormerMD model pseudo-labels, trained using our multi-task objective, enable downstream glaucoma detection accuracy via WSupCon exceeding 91% even with only 70% labeled training data. Furthermore, a model pre-trained with GazeFormerMD-generated pseudo-labels and used for linear evaluation on an unseen OCT-report dataset achieved comparable performance to a fully-supervised, trained-from-scratch model while using only 25% labeled data.

Data and Code Availability This study utilizes an internal dataset of optical coherence tomography (OCT) reports obtained from our institution that is not yet publicly available. It also utilizes a dataset of OCT reports obtained from our industry partners, Topcon, that is not yet publicly available. The corroboration study presented in Appendix D utilizes a publicly available data of chest x-rays and corresponding gaze data Karargyris et al. (2020) Goldberger et al. Our code is available here: https://github.com/AI4VSLab/Expert-Gaze-4-Supervised-Contrastive-Learning

Institutional Review Board (IRB) This study, AAAU4079, was approved by the Columbia University Irving Medical Center Institutional Review Board on 12/22/2022 and is in accordance with the tenets set forth by the Declaration of Helsinki. Informed consent was obtained from all study participants.

1. Introduction

One of the biggest challenges in artificial intelligence (AI) for healthcare lies in the acquisition of large and accurately-labeled datasets, essential for deep learning (DL) model training.

The scarcity of such labeled data hinders the development of generalizable models that can perform well on unseen data Xiao et al. (2018). Additionally, disparities in expert opinions, for example disagreement even among clinicians on the definition of blindness-causing eye diseases like glaucoma, can impede establishment of reliable ground truths for training. To address these issues, we propose to extract information from other data modalities and use them to create labels to help learn better representations for downstream tasks on different datasets. Specifically, we consider the use of gaze data of medical experts as they view medical images to extract pseudo-labels that can be used for the task of optical coherence tomography (OCT) report classification. The use of pseudo-labels is then aided by contrastive loss to help learn representations. In doing so, we showcase the robust training of DL models via self-supervision and contrastive learning derived from clinician gaze patterns.

Self-supervised learning (SSL) aims to learn robust representations of a data distribution and allows efficient training for downstream tasks. SSL is also robust in situations with smaller datasets with limited labels, as it enables learning of latent features that are common between different views of the same image. Medical image data is rich in patterns that may not be discernable by the human eye but that can be elucidated by the power of SSL algorithms.

Eye tracking data offers a wealth of information regarding the focus of attention and the expertise level of individuals examining medical reports. Gaze data from domain experts viewing images and videos abounds especially in medicine. Spatial gaze information encodes regions of importance, while temporal information encodes image-region order of importance for diagnostic decision-making. In contrast to supervised algorithms which require large quantities of hand-annotated or labeled data, weakly-supervised learning relies on 'inexact', coarse-grained labels (e.g., human eye-tracking) that can be more easily collected in bulk from which the label and ground truth can be inferred in place of costly expert labeling.

While a few methods use spatial gaze data for diagnostic AI models Stember et al. (2019), very few have been proposed to capture both temporal and spatial relationships explicitly to supervise downstream DL tasks. Deep neural networks were utilized to transfer the eye fixation coordinate system for low-cost eye tracking but not for further processing and understanding of fixations Rakhmatulin and Duchowski (2020). There have been only structured methodologies available for machine learning (ML) techniques used for analysis in different types of eye-tracking studies Kuang et al. (2023), making our DL exploration pioneering. Previous eye tracking features such as pupil size, rotating velocity, and saccades, used for biometric AI applications, have been mostly static and positional Lim et al. (2022). The successful integration of real-time gaze tracking in human neuroscience domains such as psychophysics and neuromarketing Zdarsky et al. (2021), lays the foundation for further eye-tracking exploration in ophthalmology.

In this study, we propose GazeFormerMD, a region-based eye movement encoder to learn gaze representations on medical images as pseudo-labels for disease classification. These pseudo-labels are then used for Weakly-Supervised Contrastive Learning (WSupCon) based on Khosla et al. (2021) to classify OCT reports as glaucomatous or not glaucomatous. GazeFormerMD is a transformer based model Vaswani et al. (2023) and is trained with a novel multi-task objective that consists of triplet loss Schultz and Joachims (2003); Weinberger et al. (2005) and cross entropy loss for classification. This approach leveraging eye tracking 'pseudo-labels' has the potential to enhance the performance of DL models for glaucoma diagnosis from OCT reports even with few explicit labels. Our work offers the following three key contributions:

- A new encoding scheme for gaze that retains spatiotemporal relationships by modeling gaze data as words.
- **GazeFormerMD**, a transformer based encoder that is trained with a multi-task objective which learns useful representations, creating positive and negative pairs for contrastive learning. Its embeddings are used as pseudo-labels to help learn image representations for downstream tasks.

2. Related Work

2.1. Medical Expert Gaze Patterns

Expert gaze patterns have been used in different applications in machine learning and medicine. They have been found to contain useful information, such as underlying differences between expert vs. novice image viewers Akerman et al. (2023). Past work Stember et al. (2019) has also attempted to use masks generated from eye-tracking of experts while viewing radiology images vs. masks hand-annotated by experts, to compare resulting Structures of Interest (SOIs) segmented via AI. This past work showed that eye-tracking is *not significantly different* in quality to hand annotations for segmenting SOIs. This finding provided evidence that even coarse, inexact eye movements can provide the information necessary to train AI systems to achieve accurate DL-based segmentation. Other work Li et al. (2019) has attempted to specifically enhance glaucoma detection from fundus images of the retina using labeled human attention maps and region localization as well as classification via convolutional neural networks (CNNs).

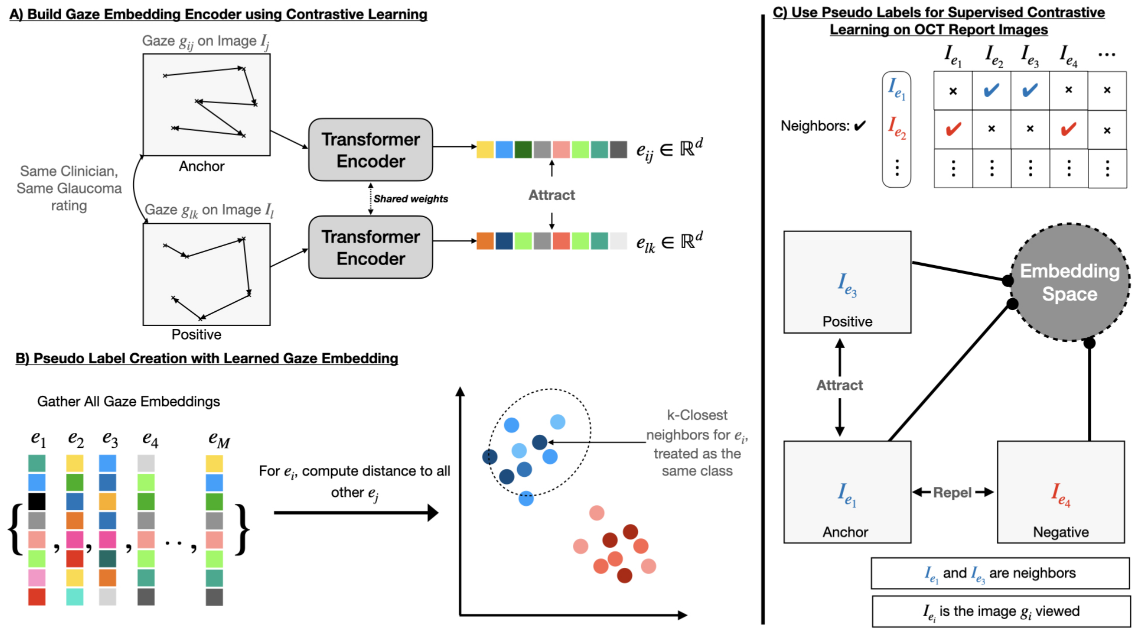

A) Build Gaze Embedding Encoder using Contrastive Learning
Gaze g_{ij} on Image I_j
Anchor
Same Clinician, Same Glaucoma rating
Gaze g_{lk} on Image I_l
Positive
Transformer Encoder
Shared weights
Transformer Encoder
$e_{ij} \in \mathbb{R}^d$
Attract
$e_{lk} \in \mathbb{R}^d$
B) Pseudo Label Creation with Learned Gaze Embedding
Gather All Gaze Embeddings
e_1 e_2 e_3 e_4 e_M
For e_i, compute distance to all other e_j
k-Closest neighbors for e_i, treated as the same class
C) Use Pseudo Labels for Supervised Contrastive Learning on OCT Report Images
I_{e_1} I_{e_2} I_{e_3} I_{e_4} …
Neighbors: ✔
I_{e_1} × ✔ ✔ × ×
I_{e_2} ✔ × × ✔ ×
Embedding Space
I_{e_3}
Positive
Attract
I_{e_1}
Anchor
Repel
I_{e_4}
Negative
I_{e_1} and I_{e_3} are neighbors
I_{e_i} is the image g_i viewed

Figure 1: Overview of method: In (A), we take gaze data as input to pretrain GazeFormerMD using our multi-task objective. Then in (B), all embeddings are gathered to generate pseudo-labels. These labels are then used for WSupCon in (C). Each e_i is assigned with neighbors E; $I_{\{e_i\}}$ now has neighbors I_E. More details about (B) are in section 3.3.4

2.2. Self-Supervised and Contrastive Learning

In contrast to supervised learning, self-supervised learning (SSL) has shown its potential to serve as an effective pre-training strategy to learn better representations Caron et al. (2021), thus enabling more robust performance than supervised learning alone especially when labeled data is limited. Balestriero and colleagues Balestriero et al. (2023) offered a detailed description of state of the art methods in SSL, including BYOL Grill et al. (2020), SimSIAM Chen and He (2020), and VICReg Bardes et al. (2022). SSL has also shown success in various medical applications, such as for medical image segmentation Chaitanya et al. (2020) and for electronic health records Krishnan et al. (2022). SimCLR Chen et al. (2020) is a contrastive learning method that attempts to maximise agreement between two views of the same image through NT-Xent loss. More recently, Khosla and colleagues Khosla et al. (2021) extended the NT-Xent loss introduced in SimCLR by leveraging labels during contrastive pre-training, showing labels can be incorporated into a contrastive learning framework, yielding supervised contrastive learning.

2.3. Weak Supervision and Gaze Data

Gaze data has been used in various tasks in machine learning for both weak supervision as well as gaze generation.

Gaze generation's goal is to create realistic gaze patterns that are similar to those of human viewers. Models such as recurrent neural networks (RNNs) and CNNs have been used to generate gaze patterns Li et al. (2022); Assens et al. (2017); Xia et al. (2019); Kümmerer et al. (2022); Yang et al. (2020), where gaze was modeled via a reward function and represented as 3D volumetric input.

Weak supervision seeks to use weak labels to supervise a model instead of the actual ground truth. Saab and colleagues Saab et al. (2021) also tried to extract gaze features on biomedical images to aid supervising image classification models. While this approach is most similar to ours, it also differs greatly: our goal is to use gaze data to aid in classification where only few labels exist. We extract features using a DL model trained on gaze instead of using gaze statistics, and we use gaze to generate weak labels to pre-train an image classification model.

3. Methods

3.1. Approach Overview

We propose a two stage approach to use gaze to aid in classifying glaucoma. First, we train GazeFormerMD with processed gaze data using a multitask loss function that includes triplet and cross entropy losses. The processed gaze data consists of a vector where each fixation is a unique word index. Positive examples for triplet learning are gaze from the same clinician given the same glaucoma classification rating. We then use the embeddings of GazeFormerMD to generate weak labels by assigning closest neighbors the same pseudo-labels. Second, these pseudo-labels are used for training an image encoder through weakly supervised contrastive learning (WSupCon). The encoder is then frozen, and we attach a linear layer for linear evaluation: classification of a report as glaucomatous or healthy. Our goal is to use gaze data to learn an informative embedding that guides contrastive learning with OCT reports, thereby learning more robust representations that can improve downstream glaucoma classification performance. We hypothesize that given the same clinician, their gaze pattern should be more similar on OCT reports of the same class than on OCT reports of a different class. Figure 1 shows the overall process of our approach.

3.2. Problem Setting and Datasets

We are given two datasets: a dataset of OCT report images with their corresponding labels $\mathcal{D}_{OCT} = \{(x_i, y_i)\}_{i=1}^{N}$ and a gaze dataset of clinicians' gaze on OCT reports $\mathcal{D}_{gaze} = \{(g_i, \tilde{y_i}, y_i^e, c_i)\}_{i=1}^{M}$. g_i, $\tilde{y_i}$, y_i^e, and c_i are the gaze time series data, clinician's diagnosis (glaucoma or healthy), expertise of the given clinician, and the corresponding clinician, respectively. $\mathcal{D}_{OCT}$ could be incomplete or very small; in our setting, we will consider a complete but small dataset. In total, we have 177 Topcon Maestro (Topcon Healthcare, Tokyo, Japan) OCT reports (LabSet) and 467 eye-tracking fixation sequences (LabGazeSet) from 10 glaucoma experts. Eye-tracking fixations were collected with Pupil Labs Core (200 Hz) and Tobii Pro Fusion (250 Hz) eye-trackers while ophthalmologists viewed OCT images. Clinician experience level varied from resident to faculty; clinicians were asked to rate each OCT report from 0 (healthy) to 100 (glaucoma). Each OCT report image x_i has gaze sequences $G_i = \{g_{i1}, g_{i2}, ...\}$ and $\sum_{i=1}^{N} |G_i| = M$. Each gaze g_{ij} also has embedding e_{ij}. $I_{\{g_{ij}\}}$ (or $I_{\{e_{ij}\}}$) is the image that was looked at by g_{ij}. We will use g_i and e_i only when discussing gaze data alone for brevity. We will also call the learned pseudo-labels $\hat{y}$.

3.3. Gaze Representation Learning

3.3.1. Gaze-to-Word Region Encoding

We take inspiration from the recent success of BERT Devlin et al. (2019) and Sentence-BERT (SBERT) Reimers and Gurevych (2019) in learning representations in natural language. Particularly, SBERT takes sequences/paragraphs of text and learns representations that can meaningfully compare sequences of text. Gaze data is a time-series from a participant's viewing of a given image. Gaze contains spatiotemporal information; the order of eye fixations and duration of each fixation contain information about relative importance that leads to the participant's diagnosis decision.

Our goal is to learn a good representation $e_i \in \mathbf{R}^d$ of these sequences so we can distinguish similarities between different gaze sequences by modelling them as words. In order to capture information from gaze data for our language-inspired DL approach, for each gaze time-series, we encode gaze in the following ways (depicted pictorially in Figure 2, and a toy example is given in A.1):

1. region-based: g_i^{region}, we convert each fixation to the letter that corresponds to the current region in which it falls (A, B, C, D, etc.).

2. region-based count vector: we convert g_i^{region} into a count vector $g_i^{region\_cv}$, where each element is the count of fixations in that region; total length of this vector is equal to the total number of regions, which is 7 (A-G).

3. region-grid: g_i^{rg}, we divide each region (A-G) into sub-regions/patches (depicted by red, green, blue, violet, yellow, magenta, and black grids in Figure 2). Each fixation is then quantized to an integer (0-109) corresponding to a patch. Since each fixation has different duration lengths, we also bin fixations into $100ms$ bins such that each

element in our fixation sequence corresponds to the same amount of time. Fixations longer than 100ms are split into different bins and averaged.

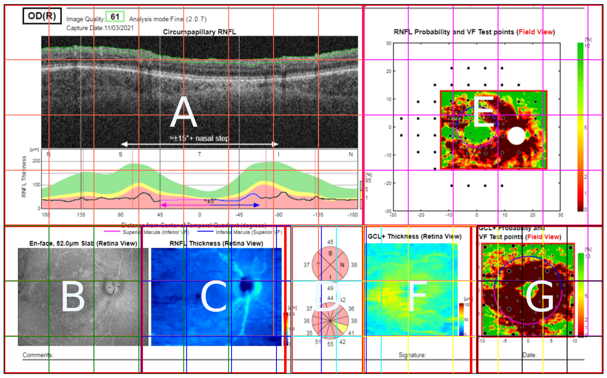

OD(R)
Circumpapillary RNFL
A
RNFL Probability and VF Test points (Field View)
E
En-face, 52.0µm Slab (Retina View)
B
RNFL Thickness (Retina View)
C
GCL+ Thickness (Retina View)
F
GCL+ Probability and VF Test points (Field View)
G

Figure 2: OCT report is split into regions, from A to G. Then each region is split into grids.

3.3.2. Baselines

We established baselines with g_i^{region}, $g_i^{region\_cv}$, and g_i^{rg} via Principal Component Analysis (PCA), Multilayer Perceptron (MLP), and logistic regression (LogReg). SGT Ranjan et al. (2021) was used to convert variable-length gaze data into a fixed-length vector, and PCA was used to reduce SGT's output since it has a dimension of 9410. For supervised baselines, the models were trained to classify experts vs. novices or glaucoma vs. healthy. We evaluated the performance of our baselines using model accuracy on glaucoma or expertise classification and visualized using t-SNE van der Maaten and Hinton (2008). PCA was used here to reduce the dimensionality of SGT's outputs. Table 1 shows input data format and machine learning (ML) pipelines used as baseline models.

For multi-layer perceptron (MLP) and logistic regression using g_i^{region} or g_i^{rg}, the learned representations were the activations before the final MLP layer and the sigmoid function, respectively.

Baseline	Method	Data Used
1	SGT→PCA→MLP	g_i^{region} / g_i^{rg}
2	SGT→PCA→LogReg	g_i^{region} / g_i^{rg}
3	MLP	$g_i^{region\_cv}$
4	LogReg	$g_i^{region\_cv}$

Table 1: Different configurations (data input format and ML pipelines) for baseline models.

3.3.3. GazeFormerMD

Architecture We use a vanilla transformer encoder with 4 layers and 4 heads, with max sequence length $l = 1024$, hidden dimension $h = 256$, and dropout rate 0.1. We first zero-pad our inputs $pad(g_i^{rg})$ to length l. The output $f(pad(g_i^{rg})) \in \mathbb{R}^{l \times h}$ of the transformer is a collection of embeddings. However, since each input gaze has variable length, we need to create a single embedding to compare them. Similar to SBERT, we experiment with the MEAN pooling strategy, which takes the output embeddings and averages them to create MEAN embedding e_i after masking outputs corresponding to zero-padding. Additionally, outputs corresponding to zero-padding, CLS, and SOS tokens are masked before MEAN pooling for training and inference.

Optimization GazeFormerMD is trained with a multi-task (MTL) objective. The first task is contrastive triplet loss, and the second task is classification for either expertise or glaucoma status. We experiment with expertise as targets since past work Akerman et al. (2023) has shown that gaze can be useful for expertise classification. The combined loss to minimize is:

$$\mathcal{L}_{MTL} = \mathcal{L}_{triplet} + \mathcal{L}_{CE} \tag{1}$$

Where $\mathcal{L}_{CE}$ is the cross entropy (CE) loss, for predicting between expert vs. novice or between glaucoma vs. healthy. Ground truth labels used here are individual clinician diagnoses $\tilde{y}$. Based on the hypothesis presented in Section 3.1, for g_i, its positives are the other g_j that were viewed by the same clinician given the same label. Triplet loss helps minimize the distance between an anchor and its positives, while maximizing distance to negatives.

Our objective is to minimize the following loss with x_i^{rg} Schroff et al. (2015):

$$\begin{aligned}\mathcal{L}_{triplet} = \sum_{i}^{N_{triplets}} [\|f(x_i^a) - f(x_i^p)\|_2 - \\ \|f(x_i^a) - f(x_i^n)\|_2 + \alpha]_+ \end{aligned} \tag{2}$$

Where $[.]_+ = max(.,0)$ is used to ensure that when positive is closer than negative, we don't penalize the model. x_i^a, x_i^p and x_i^n are anchor, positive and negative examples, respectively. α is a hyper-parameter used to avoid collapse, when $f(.)$ learns to map everything to $\mathbf{0}$. There are multiple ways to select triplet pairs; we select triplets using a **batch-all** strategy: all valid triplets are selected and averaged with only hard and semi-hard triplets. Easy triplets are those with loss less than 0; hence averaging with them would result in a very small loss. Hermans et al. (2017) provides a more in-depth discussion about triplet selections.

We use the Adam optimizer Kingma and Ba (2017) with cosine decay learning rate schedule Loshchilov and Hutter (2017) without restarts and minimum learning rate $1e-5$. The model is trained for 200 epochs, with base learning rate of 0.0001, α in triplet loss set to 5, and batch size of 32.

Data Used	**Expertise Accuracy**	**Glaucoma Accuracy**
(1) g_i^{region}	64%	**62%**
(1) g_i^{rg}	**75%**	53%
(2) g_i^{region}	51%	**62%**
(2) g_i^{rg}	49%	51%
(3) $g_i^{region\_cv}$	63%	56%
(4) $g_i^{region\_cv}$	73%	61%

Table 2: Baseline results on expertise and glaucoma classification tasks. Models achieving highest accuracy on each classification task are bolded.

3.3.4. Obtaining the pseudo-labels:

As shown in Figure 1, after training GazeFormerMD, the embeddings are gathered to create pseudo-labels. First, the embeddings are used to compute the cosine similarity matrix. We assume the embeddings with high cosine similarity came from the same class of images and from the same clinician, since this is the criteria for generating positive pairs in triplet loss. Therefore, images with similar gaze embeddings should also be similar. More formally, image x_i has gaze embeddings $E = \{e_j\}_{j=1}^{M(x_i)}$.[1] Each e_j has a sorted list of neighbors $E = [e_p, e_q, ...]$, which correspond to $\mathbb{I}_i = [I_{e_p}, I_{e_q}, ...]$. However, since it is not guaranteed that all images in $\mathbb{I}$ are unique (same image may be viewed by different clinicians), the top-k unique images are considered neighbors. The set of unique neighbors for each x_i is M_i.

3.4. Weakly-Supervised Contrastive Learning

By adapting supervised contrastive learning (SupCon), the pseudo-labels learned from gaze contrastive learning (GazeFormerMD) are used as weak labels in SupCon (WSupCon). After obtaining the set of neighbors $\{M_i\}_{i=1}^N$, we use OCT report images to train with $SupCon_{out}$ from Khosla et al. (2021):

$$\mathcal{L}_{W supcon} = - \sum_{i \in \mathrm{I}} \frac{1}{|P(i)|} \sum_{p \in P(i)} log \frac{exp((z_i \cdot z_p)/\tau)}{\sum_{a \in A(i)} exp((z_i \cdot z_a)/\tau)} \quad (3)$$

Let $i \in I = \{1, .., 2N\}$ be the index of augmented samples, and let $A(i) = I \setminus i$ be the set of indices minus the anchor. $P(i) = \{p \in A(i) : x_p \in M_i\}$ is the set of indices with the same label as the anchor within the augmented mini-batch. τ is the temperature parameter for softmax. $z_i = g(h_i)$ is the projected embedding similar to SimCLR. One important observation is that equation 3, does not enforce neighbors's neighbors to be positive pairs. For example, if z_1 is the current anchor and $P(i = 1) = \{z_2, z_3\}$, equation 3 does not necessarily maximize similarity with neighbors of z_2 in $P(i = 2)$. This is due to not having disjoint sets of pseudo-labels. Furthermore, since number of neighbors k is a hyper-parameter, similar embeddings might not be attracted together in the loss function if they are not within the k closest neighbors.

Architecture We use ResNet-50 He et al. (2015) as our encoder f. The representation h_i is the output of the final average pooling layer, with dimension 2048. The projection head g is a MLP with 2 layers, with 2048 and 1024 as the output dimensions, respectively.

Optimization Adam with cosine decay learning rate schedule without restart is also used with the same minimum learning rate. The model is trained for 200 epochs using equation 3, followed by training one linear layer attached to f with its weights frozen for 50 epochs using cross entropy. The base learning rate is 0.001, temperature τ is set to 0.1, and batch size is 16.

[1] $M(x_i)$ is the number of gaze points on x_i

Loss	Data Used	Accuracy
$\mathcal{L}_{CE}*$	g_i^{rg}	51.75% ±1.399
$\mathcal{L}_{CE}\dagger$	g_i^{rg}	84.65% ±1.351
$\mathcal{L}_{CE}\dagger$	$g_i^{rg}*$	**86.70 %** ±1.732

Table 3: GazeFormerMD training accuracy with Cross-Entropy Objective Only. We either train on expertise classification ($*$) using $\tilde{y}$ or glaucoma classification (†).

4. Results

4.1. Gaze Representation Learning

Below we show results on LabSet and validation of our method on a completely unseen set of 6941 glaucoma OCT reports from Topcon Maestro (Topcon Healthcare, Tokyo, Japan), hereafter referred to as the Topcon dataset (see Section 4.2.2 for more details). In addition, Appendix D shows results of corroborating our method on an external chest X-ray dataset with 1038 scans.

4.1.1. Baseline

Table 2 shows results from training baseline models (using configurations shown in Table 1) on either expertise or glaucoma classification. Five-fold cross-validation was performed on all models to obtain model performance. We used the best of these for downstream WSupCon training; results are shown in Table 5.

4.1.2. GazeFormerMD

To evaluate GazeFormerMD, we employed linear evaluation. Linear evaluation was performed on the pre-trained transformer model by freezing the weights and attaching a linear layer for predictions. Additionally, since novice's (resident's) gaze data are not as informative as that of experts (faculty) Brunyé et al. (2019), we also trained with gaze data from experts only. Table 3 shows the expertise and glaucoma classification accuracy on the training set using cross entropy loss applied to embeddings extracted from pathway 1.1 shown in Figure 1. Ground truth glaucoma vs. no glaucoma classification accuracy is obtained using each clinician's diagnoses $\tilde{y}_i$.

Loss	Data	% Linear Eval Acc$*$	% Linear Eval Acc†
$\mathcal{L}_{triplet}$	g_i^{rg}	52.66% ±1.284	52.66% ±1.281
$\mathcal{L}_{triplet}$	$g_i^{rg}*$	NA	47.80% ±4.489
$\mathcal{L}_{MTL}*$	g_i^{rg}	51.81% ±1.373	47.59% ±5.869
$\mathcal{L}_{MTL}\dagger$	g_i^{rg}	50.23% ±4.568	85.11% ±1.988
$\mathcal{L}_{MTL}\dagger$	$g_i^{rg}*$	NA	**88.02%** ±1.615

Table 4: GazeFormerMD training accuracy with Triplet and Multi-task Objectives. $*$ denotes only expert data was used or linear evaluation on expertise was performed; † denotes linear evaluation on glaucoma classification was performed. For each loss type, we trained a linear layer on top of the transformer, and we used either expertise or expert glaucoma diagnoses $\tilde{y}$ as targets.

4.1.3. Obtaining the Pseudo-Labels

After training our transformer model with contrastive triplet and multi-task loss as shown in pathway A of Figure 1, we obtained the accuracies shown in Table 4, which motivated our decision of which model to use to generate pseudo-labels for WSupCon.

Table 4 shows that the best pseudo-label performance is achieved by the model trained on multi-task loss with glaucoma classification cross-entropy using expert-only gaze data. Thus, pseudo-labels generated from this model are used for WSupCon.

4.2. Weakly-Supervised Contrastive Learning

Here we present results from using pseudo-labels generated using gaze from the previous section. We validated our framework using our LabSet and using the encoder trained with WSupCon to train a linear classifer on the Topcon dataset. We also compare our WSupCon approach (MTL and triplet) to our baselines (pseudo-labels generated using baselines in Table 2).

4.2.1. WSupCon on LabSet

Table 5 shows linear evaluation results with pseudo-labels $\hat{y}$ from baselines vs. those created via gaze contrastive learning (GazeFormerMD). For both baselines and GazeFormerMD, we use different amounts of data out of the training set to pre-train with WSup-

Pseudo-Label Type	**% Pre-training Data**	**% Linear Eval Data**	**Linear Eval Acc**
Baseline	100 %	70 %	79.63 % ±16.04
Triplet	100 %	70 %	72.22 % ±16.90
MTL ♢	100 %	70 %	**91.67 %** ±10.01
Baseline	75 %	30 %	72.22 % ±22.04
Triplet	75 %	30 %	76.85 % ±18.50
MTL	75 %	30 %	**87.03 %** ±11.56
Baseline	50 %	40 %	68.52 % ±15.30
Triplet	50 %	40 %	62.03 % ±15.29
MTL	50 %	40 %	**80.56 %** ±2.778

Table 5: Linear Evaluation after WSupCon pre-training on our LabSet dataset. We compare pseudo-labels generated from baseline, triplet, and MTL losses.

Con. Then, for each model, we train the linear layer with varying amounts of labeled data. We present results for best data fractions in Table 5 and for all data fractions in Appendix C. All models were trained with the same procedure described in 3.4. The overall training procedure is shown in Figure 3.

4.2.2. WSupCon on Topcon Dataset

We use a WSupCon pre-trained model with its weights frozen and trained a linear classifier on the Topcon dataset. We utilized the best model with MTL ♢ (row 3 in Table 5) and 100% data for pre-training, which achieved 91.67% on LabSet's test set, for Topcon dataset validation (containing a different data distribution than LabSet). The Topcon dataset contains 6941 OCT reports: 5008 acceptable (healthy) and 1933 unacceptable. Unacceptable includes glaucoma as well as other pathologies or poor scans. We used 20% of the dataset for testing and varied the rest of 80% for training. Table 6 shows results on the Topcon dataset. We compare our WSupCon pre-trained model with a model trained from scratch with Binary Cross Entropy loss for 200 epochs. Both baseline and WSupCon pre-trained models employ base learning rate of 0.001, batch size of 48, and Adam optimizer with cosine decay learning rate schedule without restart.

5. Discussion

5.1. Gaze Representation Learning

5.1.1. Baselines

Our baseline results indicate that although expertise classification reached 75% accuracy, glaucoma classification accuracy barely exceeded chance when embeddings were derived from SGT, PCA, MLP, or logistic regression approaches. However, the fact that these baselines crossed the chance threshold validates our encoding scheme using language-inspired region-based, region-based count vector, and region grid methods.

5.1.2. GazeFormerMD

Gaze contrastive learning was effective in achieving the best results compared to baselines and when only trained with cross entropy loss. The multi-task nature of our objective further encouraged the model to bring gaze embeddings closer together. Although the multi-task objective with glaucoma classification and expert data only (row 5, Table 4) outperformed that trained with a cross-entropy objective alone and only expert data (row 3, Table 3), the performance gap was not significant (88.02% vs. 86.70%). This may be due to the triplet loss being noisy, since triplets are selected at each training step from the current mini-batch. The triplet selection strategy could also affect performance as the batch-hard strategy was found to work the best. Since past work has shown that gaze data can be used to classify expertise with accuracy beyond 90% Akerman et al. (2023), this suggests our

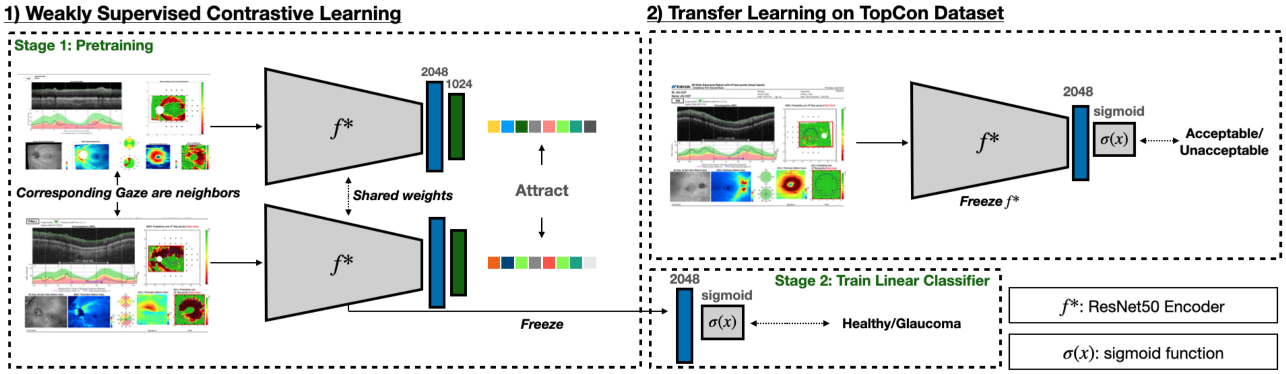


Figure 3: WSupCon and transfer learning training pipeline. 1) WSupCon training: OCT reports whose clinician gaze data are neighbors; these are an example of a positive pair. After WSupCon training, encoder f^* is frozen to train a linear classifier for healthy vs glaucoma. This same encoder is then used for training the Topcon dataset in 2).

	% Linear Eval Data	**Linear Eval Acc**	**Sensitivity**	**Specificity**
ResNet50-$\mathcal{L}_{CE}$	100 %	78.92% ±5.165	78.88% ±8.436	79.02% ±4.165
ResNet50-$\mathcal{L}_{CE}$	50 %	81.00% ±2.580	85.72% ±4.325	68.52% ±4.571
ResNet50-WSupCon	100 %	**81.41 %** ±1.220	86.66 % ±1.732	67.51 % ±1.783
ResNet50-WSupCon	50 %	79.97 % ±0.636	85.71 % ±1.359	65.61 % ±1.378
ResNet50-WSupCon	25 %	**81.31 %** ±0.957	86.69 % ±1.647	67.01 % ±2.209

Table 6: Validation results on Topcon dataset. This dataset has a different data distribution than that of our LabSet. We compared Topcon WSupCon results to a model trained-from-scratch with Cross-Entropy loss and only supervised labels (ResNet50-$\mathcal{L}_{CE}$).

current training method may be improved by learning more generative patterns in gaze data. We could extend our approach by training GazeFormerMD with additional tasks, such as masked language modeling and next-word prediction, effectively learning a generative model for gaze data. Our results also suggest the potential for improvement by combining gaze and image data into one model to better capture the correlation between image location and gaze while maintaining temporal information.

5.1.3. Obtaining the Pseudo-Labels

We generated pseudo-labels from gaze contrastive learning and multi-task loss (including triplet loss combined with glaucoma cross-entropy loss) and applied those for WSupCon from OCT report images. These pseudo-labels achieved up to 88.02% linear glaucoma classification accuracy on training data. Further experiments could use k-means clustering or other techniques to assign labels.

5.2. Weakly-Supervised Contrastive Learning

5.2.1. WSupCon on LabSet

WSupCon with GazeFormerMD-generated pseudo-labels outperformed baselines for each variation of the model shown in Table 5 (when 50%, 75%, or 100% of the training dataset was used for WSupCon pre-training and 30%, 40%, or 70% labeled data was used for linear evaluation). In addition, multi-task objectives (row 5, Table 4) performed better than triplet objectives (row 2, Table 4) alone across varying fractions of pre-training (50%, 75%, 100%), with multi-task accuracy exceeding 91% at 100% WSupCon pre-training and 70% labeled data during linear evaluation. There were occasional drops in testing accuracy (as shown in Appendix C), which could be partially explained by the fact that we used five distinct seeds to evaluate model performance and present average results. Since our LabSet is relatively small, each of these seeds could have resulted in a different data distribution, leading to a large variance in performance. We also include t-SNE plots of both gaze embeddings and image embeddings (in Appendix F) to visualize our LabSet data distribution.

The inclusion of triplet loss contributes to the model's ability to construct robust positive and negative relations between gaze patterns, generating a well-defined clustering between similar embeddings and thus more accurate pseudo-labels. Simultaneously, the integration of cross-entropy loss serves a crucial role in guiding the model to distinguish between glaucomatous and healthy OCT data. We conducted Mann-Whitney U tests of the average WSupCon accuracy results between a model using only triplet loss and and a model using our combined multi-task loss. We observed a significant improvement in glaucoma detection accuracy for the model trained with multi-task loss compared to the model trained with triplet loss only with a p-value of 0.00018 at 75% WSupCon pre-training. Therefore, by combining these loss functions, GazeFormerMD achieves a holistic learning strategy, exhibiting enhanced performance through use of contrastive similarity and classification objectives.

5.2.2. WSupCon on Topcon Dataset

There are two findings from training the Topcon dataset on our WSupCon pre-trained model. First, our model is able to achieve test accuracy that is comparable to a model trained from scratch with fully-labeled data, which takes considerably more resources to train. This implies that pre-training with gaze information (from our LabSet alone) helped our model to generalize to unseen data. Second, we tried to vary the amount of data used to train the linear classifier. With only 25% of the data, we were able to achieve comparable results to 100% of the data. This could be due to the pre-trained encoder providing useful representations, reducing the complexity of the classification task.

6. Conclusions and Future Directions

Our work showcases that medical expert gaze data (specifically eye movements of ophthalmologists as they view optical coherence tomography reports for glaucoma detection) has the potential to enhance disease detection accuracy especially in settings when access to labeled data is lacking. Pseudo-labels generated purely from region-based encodings of gaze data on OCT reports enabled downstream glaucoma classification with suboptimal accuracy; however, pseudo-labels derived from supervised gaze contrastive learning (GazeFormerMD) using cross-entropy, triplet, and multi-task loss achieved up to 88.02% training accuracy. WSupCon using these pseudo-labels was effective at achieving accuracy beyond 91% with only 70% labeled data and 100% pseudo-label pre-training. The process of generating pseudo-labels introduces noise when learning with WSupCon; thus, our relatively high pseudo-label accuracy using gaze data (88.02%) did not always transfer to downstream glaucoma classification model performance using OCT image data. Nonetheless, when a model pre-trained on our LabSet with GazeFormerMD-generated pseudo-labels was used for linear evaluation on an unseen OCT report dataset, it achieved comparable performance to a fully-supervised, trained-from-scratch model while using only 25% of the labeled data. Our method showcases the power of gaze as a form of contrastive pre-training for potential development of foundation models for medical applications.

Future work will explore other ways to preserve performance from gaze to the images being observed. Future directions also include integrating other modalities of data to provide complementary information for better model comprehension of glaucoma patterns, such as the subject's corresponding visual field images alongisde OCT reports Tian et al. (2023), or clinicians' textual comments on each OCT report observation Radford et al. (2021). Another potential direction is employing alternate embedding approaches by quantizing eye tracking into features based on characteristics beyond region alone or pre-training on large datasets of eye tracking data (even those of non-experts) prior to fine tuning on clinician data. Lastly, leveraging text responses of clinicians in addition to their eye movements would enable the use of large language model inspired methods to predict the next fixation in a sequence, much like predicting the next word in a sentence.

Acknowledgments

We would like to thank Richard Zemel for his valuable comments and feedback, as well as Mary Durbin from Topcon Healthcare for collaboration through use of the Topcon dataset.

References

M. Akerman, S. Choudhary, J.M. Liebmann, G.A. Cioffi, R.W.S. Chen, and K.A. Thakoor. Extracting decision-making features from the unstructured eye movements of clinicians on glaucoma oct reports and developing ai models to classify expertise. *Frontiers in Medicine*, 10(1251183): 2579–2605, 2023. URL doi:10.3389/fmed.2023.1251183.

Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E. O'Connor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2331–2338, 2017. doi: 10.1109/ICCVW.2017.275.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022.

Tad T. Brunyé, Trafton Drew, Donald L. Weaver, and Joann G. Elmore. A review of eye tracking for understanding and improving diagnostic interpretation. *Cognitive Research: Principles and Implications*, 4(1):7, Feb 2019. ISSN 2365-7464. doi: 10.1186/s41235-019-0159-2. URL https://doi.org/10.1186/s41235-019-0159-2.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12546–12558. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/949686ecef4ee20a62d16b4a2d7ccca3-Paper.pdf.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning, 2020.

E.S. Dalmaijer. Pygaze: Open-source toolbox for eye tracking in python, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Ary L. Goldberger, Luís A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger Mark, ..., and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification, 2017.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, Dec 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

Antonis Karargyris, Sujata Kashyap, Ioanna Lourentzou, Jie Wu, Matthew Tong, Akshay Sharma, Syed Abedin, David Beymer, Vishal Mukherjee, Elizabeth Krupinski, and Mehdi Moradi. Eye gaze data for chest x-rays. PhysioNet, 2020. URL https://doi.org/10.13026/qfdz-zr67. Version 1.0.0.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Rayan Krishnan, Pranav Rajpurkar, and Eric J. Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12): 1346–1352, Dec 2022. ISSN 2157-846X. doi: 10.1038/s41551-022-00914-1. URL https://doi.org/10.1038/s41551-022-00914-1.

Peng Kuang, Emma Söderberg, Diederick C Niehorster, and Martin Höst. Applying machine learning to gaze data in software development: a mapping study. In *Eleventh International Workshop on Eye Movements in Programming, EMIP 2023*, 2023.

Matthias Kümmerer, Matthias Bethge, and Thomas S A Wallis. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *J. Vis.*, 22 (5):7, April 2022.

Jason Li, Nicholas Watters, Yingting, Wang, Hansem Sohn, and Mehrdad Jazayeri. Modeling human eye movements with neural networks in a maze-solving task, 2022.

L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu. Attention based glaucoma detection: A large-scale database and cnn model. In *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, Apr 2019. doi: 10.48550/arXiv.1903.10831. URL https://doi.org/10.48550/arXiv.1903.10831.

Jia Zheng Lim, James Mountstephens, and Jason Teo. Eye-tracking feature extraction for biometric machine learning. *Frontiers in neurorobotics*, 15:796895, 2022.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Ildar Rakhmatulin and Andrew T Duchowski. Deep neural networks for low-cost eye tracking. *Procedia Computer Science*, 176:685–694, 2020.

Chitta Ranjan, Samaneh Ebrahimi, and Kamran Paynabar. Sequence graph transform (sgt): A feature embedding function for sequence data mining, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

Khaled Saab, Sarah M Hooper, Nimit S Sohoni, Jupinder Parmar, Brian Pogatchnik, Sen Wu, Jared A Dunnmon, Hongyang R Zhang, Daniel Rubin, and Christopher Ré. Observational supervision for medical image classification using gaze data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–614. Springer, 2021.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015. doi: 10.1109/cvpr.2015.7298682. URL https://doi.org/10.1109%2Fcvpr.2015.7298682.

Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/file/d3b1fb02964aa64e257f9f26a31f72cf-Paper.pdf.

J.N. Stember, H. Celik, E. Krupinski, P.D. Chang, S. Mutasa, B.J. Wood, A. Lignelli, G. Moonis, L.H. Schwartz, S. Jambawalikar, and U. Bagci. Eye tracking for deep learning segmentation using convolutional neural networks. *Journal of digital imaging*, 32(4):597–604, 2019. URL https://pubmed.ncbi.nlm.nih.gov/31044392/.

Ye Tian, Mingyang Zang, Anurag Sharma, Sophie Z Gu, Ari Leshno, and Kaveri A Thakoor. Glaucoma progression detection and humphrey visual field prediction using discriminative and generative vision transformers. In *International Workshop on Ophthalmic Medical Image Analysis*, pages 62–71. Springer, 2023.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. URL https://proceedings.neurips.cc/paper_files/paper/2005/file/a7f592cef8b130a6967a90617db5681b-Paper.pdf.

Chen Xia, Junwei Han, Fei Qi, and Guangming Shi. Predicting human saccadic scanpaths based on iterative representation learning. *IEEE Transactions on Image Processing*, 28(7):3502–3515, 2019. doi: 10.1109/TIP.2019.2897966.

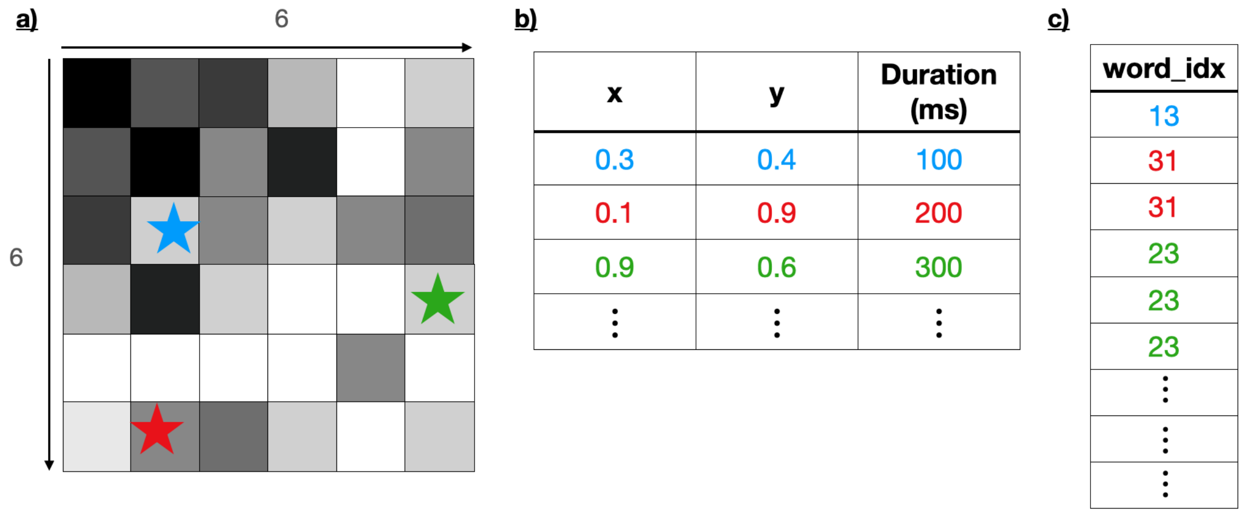
Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association : JAMIA*, oct 2018. URL https://doi.org/10.1093/jamia/ocy068.

Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning, 2020.

Niklas Zdarsky, Stefan Treue, and Moein Esghaei. A deep learning-based approach to video-based eye tracking for human psychophysics. *Frontiers in human neuroscience*, 15:685830, 2021.

Appendix A. Processing Gaze Data Details

A.1. Preserving Spatial-temporal Relationships



x	y	Duration (ms)
0.3	0.4	100
0.1	0.9	200
0.9	0.6	300
⋮	⋮	⋮

word_idx
13
31
31
23
23
23
⋮
⋮
⋮

Figure 4: Data Preprocessing. a) shows an image being split into 6×6 grids, and locations of the fixations are shown in b) with their corresponding color. b) shows the recorded fixation data and duration of each fixation. c) shows fixations after processing by assigning a word index and repeating to denote duration.

Figure 4 shows the overall process of preprocessing data. Each row in c) denotes fixation for a chosen duration, $100ms$ in our case. If there are fixations lasting $150ms$ followed by another $150ms$ fixation, we linearly interpolate the fixation. The $word_{idx}$ denotes which cell the fixation falls in (as shown in a), with upper left corner as 0. Since each unique $word_{idx}$ always corresponds to the same location, and all the reports shown in Fig. 2 have the same scan of different patients at the same location, location information is preserved. By maintaining the order and the duration (by repeating the same $word_{idx}$ to indicate longer duration), temporal information is preserved.

Appendix B. Additional Details on Strategies to Prevent Overfitting

Data Augmentation used for our model differs from the ones used in Chen et al. (2020) Khosla et al. (2021). The following is our data augmentation process. It is important to note that color distortion was removed since it interferes with colors used to show intensity in OCT reports. A pseudo code for our augmentation is below using PyTorch.

```
transforms = transforms.Compose([
    transforms.RandomResizedCrop(
        size=(512, 512), scale=(0.4, 1.0), antialias=True),
    transforms.RandomHorizontalFlip(p=0.5),
    RandomCutOut(size=(512, 512), min_cutout=0.1, max_cutout=0.7),
    transforms.RandomRotation(degrees=(0, 360)),
])
```

Appendix C. Additional Experimental Results on LabSet

Figure 5 shows glaucoma detection accuracy via WSupCon with varying % pre-training data and varying % linear evaluation data using pseudo-labels generated from baseline models and GazeFormerMD.

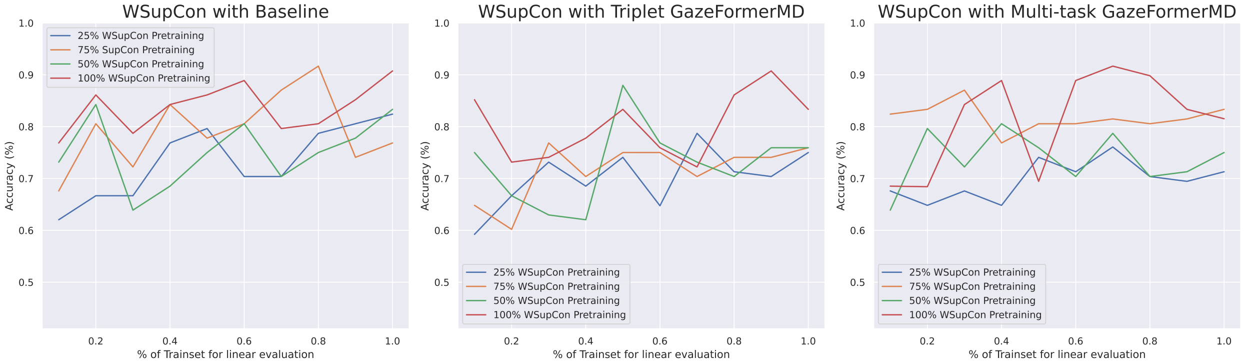


Figure 5: Glaucoma detection accuracy via WSupCon with varying % pre-training data and varying % linear evaluation data using pseudo-labels generated from baseline models and GazeFormerMD (triplet model, row 2 in Table 4 and multi-task model, row 5 in Table 4).

Appendix D. Additional Experimental Results Showcasing Generalization of our Method on Different Gaze and Image Data

D.1. Eye Gaze Data for Chest X-ray Dataset

The "Eye Gaze Data for Chest X-ray Dataset (EGD-CXR)" is a public dataset containing radiologist gaze information on the chest X-ray dataset, MIMIC-CXR Karargyris et al. (2020) Johnson et al. (2019). It contains X-ray scans from 1083 patients that are either diagnosed as normal, with congestive heart failure (CHF), or with Pneumonia. To remain consistent with our dataset, we perform binary classification between healthy and unhealthy (normal vs. CHF or Pneumonia). We applied our GazeFormerMD + WSupCon pipeline to this dataset to show that our method can generalize across other eye tracking and imaging datasets.

Since EDG-CXR contains gaze data only from one radiologist, images belonging to the same class serve as positives for triplet loss. The same training method described in the main paper is used here. Unlike our glaucoma dataset (Fig. 2) in which there are 6 sub-images per OCT report, this dataset contains just one scan per image. Therefore, g^{rg} splits the scan into grids and applies the same method as described in the main paper.

Table 7 shows the model accuracy of our GazeFormerMD at classifying healthy vs not-healthy chest X-rays. Triplet loss yields the highest accuracy with linear evaluation while multi-task loss yields the highest kNN accuracy. Table 8 shows our model's performance using pseudo-labels as well as other methods.

Appendix E. Augmenting Supervised Learning with Self-Supervision and Expert Gaze Data

This section describes a sub-study in which we performed self-supervised pre-training via SimCLR followed by supervised fine-tuning with gaze-overlaid OCT reports. We observed significant improvement in accuracy when we utilized a self-supervised pre-training strategy to detect glaucoma from OCT reports superimposed

Loss	**Data**	**% Pre-training Data**	**% Linear Eval Data**	**Linear Eval Acc**	**kNN**	**MCC**
$\mathcal{L}_{CE}$	g_i^{rg}	100%	100%	80.80%	80.49%	0.5534
$\mathcal{L}_{triplet}$	g_i^{rg}	100%	100%	**81.57%**	80.34%	0.5711
$\mathcal{L}_{MTL}$	g_i^{rg}	100%	100%	81.26%	**82.03%**	**0.6002**
$\mathcal{L}_{CE}$	g_i^{rg}	50%	50%	82.03%	**81.87%**	**0.5840**
$\mathcal{L}_{triplet}$	g_i^{rg}	50%	50%	80.95%	80.03%	0.5446
$\mathcal{L}_{MTL}$	g_i^{rg}	50%	50%	**82.33%**	81.41%	0.5801

Table 7: **Gaze**: Test accuracy with cross entropy, triplet and multi-task losses, respectively, at classifying healthy vs not-healthy chest X-rays using gaze data. The images are split into (28x28) grids for gaze processing. $\mathcal{L}_{MTL}$ achieves highest linear evaluation accuracy.

Loss	**% Pre-training Data**	**% Linear Eval Data**	**Linear Eval Data Acc**
$\mathcal{L}_{CE}$	100%	100%	75.57%
$\mathcal{L}_{SimCLR}$	100%	100%	77.08%
$\mathcal{L}_{SupCon}$	100%	100%	74.65%
$\mathcal{L}_{WSupCon}$ (ours)	100%	100%	**71.43%**
$\mathcal{L}_{CE}$	50%	50%	68.66%
$\mathcal{L}_{SimCLR}$	50%	50%	65.90%
$\mathcal{L}_{SupCon}$	50%	50%	70.05%
$\mathcal{L}_{WSupCon}$ (ours)	50%	50%	**67.28%**

Table 8: **Image**: Downstream image classification accuracy with cross-entropy, SimCLR, SupCon, and our method, respectively, at classifying healthy vs not-healthy chest X-rays. Our methods maintain high test accuracy even with 50% labeled data. The pseudo-labels are trained with 50% labeled data and the multi-task loss shown in Table 7.

with fixation heatmaps vs. clean OCT reports (without overlaid gaze data) followed by supervised fine-tuning with partially-labeled data.

By leveraging Self-Supervised Learning (SSL) as a pre-training method, our model can learn intricate patterns from the data, capitalizing on various learned (pre-text) tasks, such as predicting relationships within medical reports or reconstructing masked portions. Integrating ophthalmologist gaze data enhances this process, enabling the model to understand spatial cues from clinicians' gaze patterns. Subsequently, fine-tuning the pre-trained model with Supervised Learning (SL) on the available labeled data further refines its features for the specific clinical task.

This approach offers several advantages. First, the model is equipped with an understanding of the intrinsic data structure through SSL. Second, the inclusion of gaze data provides a nuanced perspective, enhancing the model's temporal and spatial comprehension. Third, the model's ability to generalize is significantly improved, due to the combined power of SSL, SL, and expert gaze insights.

In our study, we delved into the intricate relationship between unlabeled data and model performance when pre-training with SSL. We randomly sampled 25%, 50%, 75%, and 100% of our training data to treat as unlabeled data in our pre-training task. Then, we systematically sampled the dataset into different proportions of labeled data, ranging from 10% to 90% for our SL fine-tuning task. We used a ResNet-50 backbone and SimCLR loss for SSL pre-training (200 epochs), followed by SL fine-tuning (50 epochs) with cross-entropy loss for obtaining final glaucoma vs. healthy classification.

We augmented our models in two ways: one set was trained exclusively on clean OCT reports, while the other incorporated gaze fixation data (fixation information was overlaid on the image via PyGaze heatmaps Dalmaijer (2021)). This innovative augmentation strategy aimed to provide the models with an additional layer of information, particularly in situations where complete labeled data was lacking.

To ensure the reliability of our results, we implemented a robust testing procedure. The entire dataset was randomly reordered three times, and each experiment for each unlabeled data percentage was run three times and then tested. By averaging the outcomes, we accounted for any potential variability in the data and training process, ensuring the integrity of our findings.

E.1. Results: Augmenting Supervised Learning with Self-Supervision and Expert Gaze Data

To understand the statistical significance of these improvements, especially in comparison to baseline models or other methods, we carefully selected statistical tests based on the unique characteristics of our data. Kolmogorov-Smirnov tests were chosen for assessing normality, considering the non-normally distributed nature of the accuracy data. Mann-Whitney U tests were employed for comparing accuracy between different conditions due to their suitability for non-parametric analysis.

In our analysis, Kolmogorov-Smirnov tests confirmed non-normality in all models' accuracy distributions (p=0), emphasizing the unique nature of the data. Subsequent Mann-Whitney U tests of the average accuracy results between with and without gaze fixations yielded p-values of 0.0003873, 0.0003792, and 0.01706 for 25%, 50%, and 100% SSL pre-training, respectively (75% is excluded since its p-value was insignificant). This indicates a significant difference in accuracy between models with and without gaze fixation data. These results emphasize the impact of integrating gaze fixation data on glaucoma classification accuracy from OCT reports, highlighting its potential to enhance model performance compared to a model trained solely on clean OCT reports.

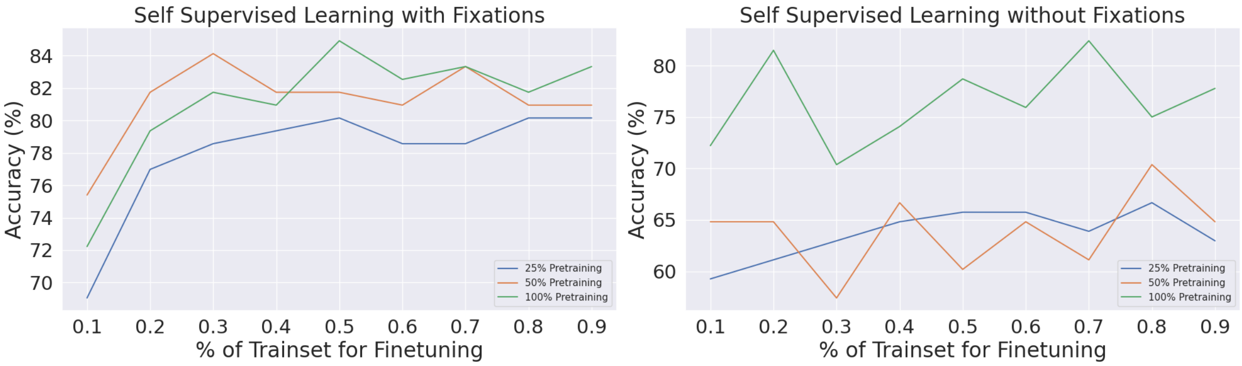


Figure 6: Comparison of average glaucoma detection accuracy of models trained using varying amounts of OCT report data for SSL pre-training, followed by supervised fine-tuning with 10% to 90% labeled data, with gaze data superimposed (left) or without gaze data superimposed (right).

Appendix F. Ophthalmologist Gaze Dataset Summary

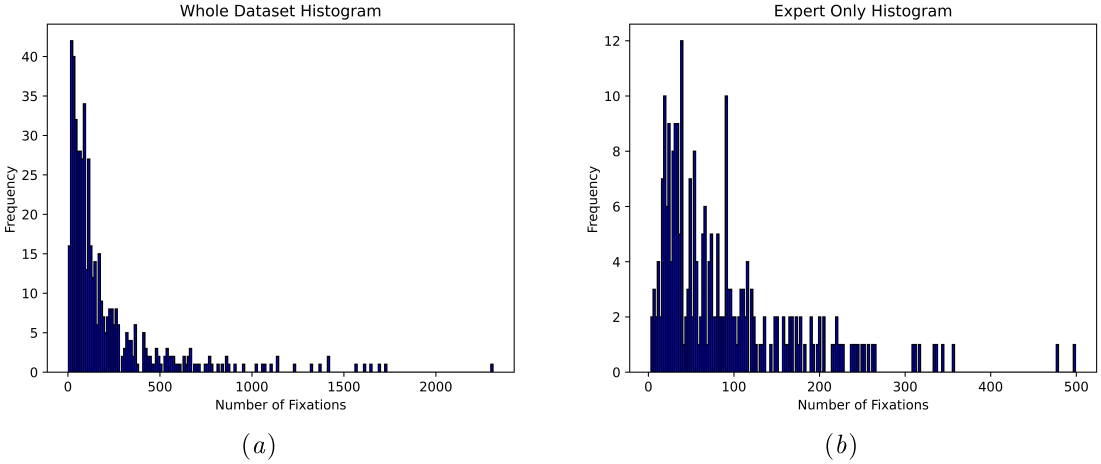
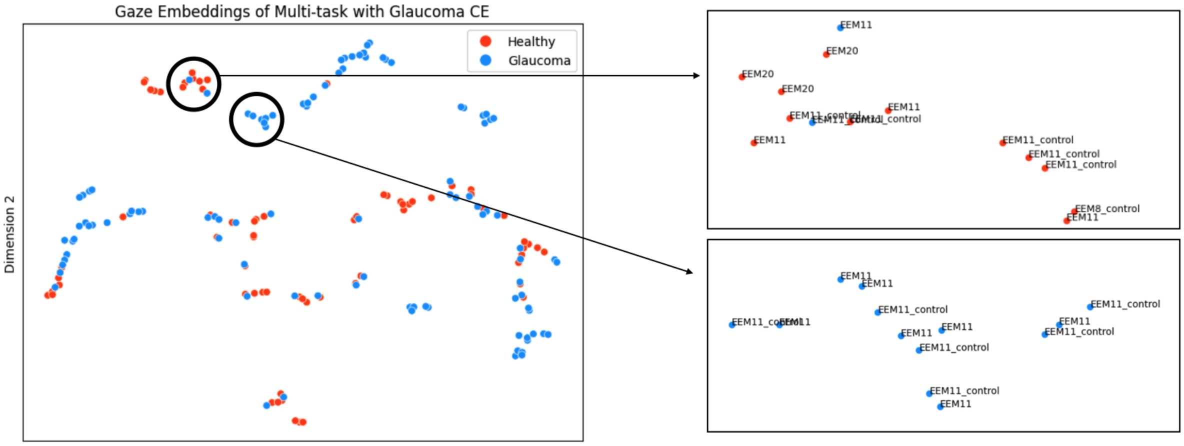


Figure 7: Histogram of fixation counts on OCT reports for all clinicians (left) and for experts only (right). Expert gaze data have fewer outliers with a high number of fixations.



EEM11 and EEM11_control are from the same clinician

Figure 8: t-SNE plot of Gaze embeddings. Note that the small clusters are successfully grouped by glaucoma vs healthy and which clinician the gaze data came from, which was the criteria for positive pairs in GazeFormerMD's triplet loss.

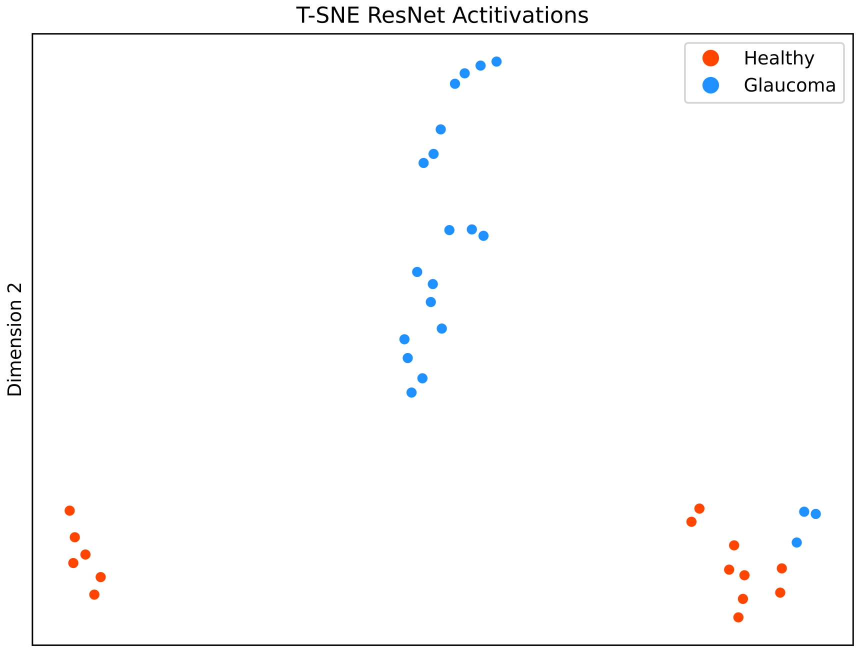


Figure 9: t-SNE plot of ResNet's activations after training with WSupCon on the test dataset. Glaucoma vs. healthy activations are relatively well-separated.