Supplementary Material

S1 Supplementary Tables

Table S1: Criteria used for labeling cases in the Stanford, India, and Nepal datasets as glaucoma and normal.

Labels	Glaucoma	Normal
Criteria	• Clinical Glaucomatous Disc changes (as per ISGEO classification [16]), *and* • OCT Glaucomatous defects on deviation maps and not all green on OCT RNFL and/or OCT GCIPL maps, *and* • 2 repeatable VF defects as per HAP criteria [17]. Reliably measured data were used, *i.e.* with a fixation loss $<$ 20%, false positive errors $<$ 15%, and false negative errors $<$ 33%, *or* total cupping of the optic nerve and unable to perform VF evaluation, *and* • On Treatment for Glaucoma or has undergone surgery/SLT-ALT.	• No disc changes for glaucoma (few cases have high cup disc ratio $>$ 0.6 but no other glaucomatous disc changes), *and* • No OCT glaucomatous defects on deviation maps and all green OCT RNFL and OCT GCIPL maps, *and* • No visual field defects, *and* • No treatment/review after a duration no lesser than a year as per chart review.

Table S2: Dataset distribution (in numbers) for all the data used in this study. **N** indicates normal cases and **G** indicates cases with glaucoma.

Dataset	Patients			Eyes			Scans		
	N	**G**	**Total**	**N**	**G**	**Total**	**N**	**G**	**Total**
Stanford (Training)	167	207	374	291	363	654	542	1022	1564
Stanford (Validation)	23	27	50	39	48	87	61	142	203
Stanford (Test)	66	89	155	113	157	270	241	453	694
Hong Kong	99	155	254	196	277	473	666	959	1625
India	73	101	174	121	171	292	184	461	645
Nepal	102	114	216	166	181	347	187	229	416

Table S3: Demographic background of the training set from Stanford. Significance tests for the Normal subset are performed relative to the Glaucoma subset of the Stanford training set. **Mean Deviation** (MD) is an overall value of the total amount of visual field loss. Note that for some patients, demographic data was incomplete and therefore, aggregate numbers do not necessarily add up to the dataset size.

	Glaucoma	**Normal**
Age (years)	69.41 ($\pm$14.70)	61.84 ($\pm$15.20, $p < 0.005$)
Asian (n)	163 (39.9%)	118 (49.0%)
Caucasian (n)	147 (36.0%)	77 (32.0%)
African American (n)	15 (3.6%)	10 (4.1%)
Hispanic (n)	32 (7.8%)	19 (7.9%)
Data of ethnicity unavailable (n)	50 (12.2%)	14 (5.8%)
Average MD	-9.75 ($\pm$7.50)	-0.79 ($\pm$1.20)
Mean Refractive Error	-3.57 ($\pm$3.37)	-2.20 ($\pm$4.62, $p < 0.001$)

Table S4: Demographic background of the validation set from Stanford. Significance tests of the normal and glaucoma subsets are performed relative to the normal and glaucoma subsets of the Stanford training set, respectively.

	Glaucoma	**Normal**
Age (years)	70.09 ($\pm$10.37, $p = 0.74$)	67.03 ($\pm$11.30, $p = 0.0715$)
Asian (n)	14 (25.0%)	13 (43.3%)
Caucasian (n)	16 (29.9%)	6 (20.0%)
African American (n)	2 (3.6%)	4 (13.3%)
Hispanic (n)	6 (11.0%)	2 (6.6%)
Data of ethnicity unavailable (n)	17 (31.0%)	6 (20%)
Average MD	-7.89 ($\pm$4.17, $p = 0.0724$)	-1.31 ($\pm$1.06, $p = 0.0241$)
Mean Refractive Error	-2.16 ($\pm$4.17, $p = 0.1949$)	-0.53 ($\pm$1.99, $p < 0.005$)

Table S5: Demographic background of the test set from Stanford. Significance tests of the normal and glaucoma subsets are performed relative to the normal and glaucoma subsets of the Stanford training set, respectively.

	Glaucoma	**Normal**
Age (years)	69.82 ($\pm$16.15, $p = 0.7886$)	63.00 ($\pm$16.93, $p = 0.4838$)
Asian (n)	60 (38.2%)	57 (50.0%)
Caucasian (n)	65 (41.4%)	42 (36.8%)
African American (n)	8 (5.0%)	6 (5.2%)
Hispanic (n)	13 (8.2%)	4 (3.5%)
Data of ethnicity unavailable (n)	11 (7.0%)	5 (4.3%)
Average MD	-9.01 ($\pm$7.52, $p = 0.2709$)	-0.79 ($\pm$0.98, $p = 1.000$)
Mean Refractive Error	-2.64 ($\pm$2.86, $p = 0.0011$)	-1.92 ($\pm$2.03, $p = 0.1552$)

Table S6: Demographic background of the Hong Kong test set, such as gender and ethnicity distribution, and mean values (standard deviations) for visual field parameter mean deviation (MD) and Mean Refractive error. Significance tests of the normal and glaucoma subsets are performed relative to the normal and glaucoma subsets of the Stanford training set, respectively.

	Glaucoma	**Normal**
Age (years)	65.90 ($\pm$9.30, $p < 0.005$)	61.05 ($\pm$8.50, $p = 0.5139$)
Asian (n)	277 (100%)	196 (100%)
Average MD	-8.50 ($\pm$6.81, $p = 0.035$)	-0.90 ($\pm$1.30, $p = 0.3526$)
Mean Refractive Error	-0.85 ($\pm$2.57, $p < 0.005$)	-0.51 ($\pm$2.15, $p < 0.005$)

Table S7: Demographic background of the India test set, such as gender and ethnicity distribution, and mean values (standard deviations) for visual field parameter mean deviation (MD) and Mean Refractive error. Significance tests of the normal and glaucoma subsets are performed relative to the normal and glaucoma subsets of the Stanford training set, respectively.

	Glaucoma	**Normal**
Age (years)	63.84 ($\pm$11.72, $p < 0.005$)	54.76 ($\pm$14.95, $p < 0.005$)
Asian (n)	173 (100%)	130 (100%)
Average MD	-12.74 ($\pm$9.22, $p < 0.005$)	-2.10 ($\pm$1.30, $p < 0.0001$)
Mean Refractive Error	-0.48 ($\pm$2.25, $p < 0.005$)	-0.44 ($\pm$2.19, $p < 0.005$)

Table S8: Demographic background of the Nepal test set, such as gender and ethnicity distribution, and mean values (standard deviations) for visual field parameter mean deviation (MD) and Mean Refractive error. Significance tests of the normal and glaucoma subsets are performed relative to the normal and glaucoma subsets of the Stanford training set, respectively.

	Glaucoma	**Normal**
Age (years)	45.34 ($\pm$17.08, $p < 0.005$)	39.17 ($\pm$12.28, $p < 0.005$)
Asian (n)	184 (100%)	173 (100%)
Average MD	-8.30 ($\pm$7.04, $p = 0.037$)	-2.32 ($\pm$1.47, $p < 0.005$)
Mean Refractive Error	-1.38 ($\pm$2.38, $p < 0.005$)	-1.17 ($\pm$1.36, $p < 0.005$)

Table S9: Distribution of cases in terms of glaucoma severity. Classification based on Mean Deviation (Severe: MD ≤ -12, Moderate: $-12 <$ MD ≤ -6, Mild: $-6 <$ MD).

	Stanford	**Hong Kong**	**India**	**Nepal**
Severe Glaucoma	28.40%	24.00%	44.80%	21.10%
Moderate Glaucoma	18.93%	26.10%	17.20%	22.76%
Mild Glaucoma	52.66%	49.70%	37.90%	56.10%

Table S10: Comparison of myopia severity (in terms of spherical equivalent) between the Stanford, Hong Kong, India, and Nepal test sets. **G** stands for glaucoma and **N** stands for normal. Chi-squared test was used for severe myopia distribution analysis (Myopia severity distribution: Severe: $D \leq -6$, Moderate: $-6 < D \leq -3$, Mild: $-3 < D$, where D is diopter). Emmetropia is defined as spherical equivalent of $-0.25D$ to $+0.25D$.

Subset	**Severe Myopia**	**Moderate Myopia**	**Mild Myopia**	**Emmetropia**	**Hypermetropia**
Stanford (G)	8.88% ($p = 0.70$)	8.10%	42.20%	11.11%	20.00%
Stanford (N)	4.20% ($p = 0.98$)	10.08%	31.09%	5.88%	47.89%
Hong Kong (G)	4.70% ($p = 0.12$)	12.50%	37.50%	5.90%	39.20%
Hong Kong (N)	0.0% ($p < 0.001$)	21.01%	15.70%	10.50%	47.30%
India (G)	0.0% ($p < 0.001$)	3.94%	43.20%	22.30%	38.10%
India (N)	0.0% ($p < 0.001$)	16.60%	30.30%	15.15%	37.87%
Nepal (G)	2.50% ($p < 0.001$)	14.28%	43.80%	10.70%	25.00%
Nepal (N)	0.0% ($p < 0.001$)	6.38%	53.00%	0.0%	40.40%

Table S11: Comparison of additional clinical data between the primary set and four external evaluation datasets. The statistical analysis was performed with the MedCalc Software (Version 19.4). Results are expressed as mean (± standard deviation) and independent 2 sample t-test was used to evaluate the level of significance. A p-value of 0.005 or less was considered significant. Chi-squared test was used for comparisons of categorical demographic data for proportions. **G** stands for glaucoma and **N** stands for normal. n indicates the number of eyes in each set. p-values for Stanford Training (N) have been computed against Stanford Training (G). For all other datasets, (N) subsets have been compared against Stanford Training (N), and (G) subsets have been compared against Stanford Training (G), respectively. **Visual Field Index** (VFI) is a global metric that assigns a number between 1-100 percent based on aggregate percentage of visual function with 100% being perfect age-adjusted visual field. **Pattern Standard Deviation** (PSD) depicts focal defects on visual fields by comparing the differences between the adjacent points on the visual field.

Subset	Cup-Disc Ratio	IOP	Gender Distribution (F:M)	PSD	VFI
Stanford Training (G) ($n = 363$)	0.80 (±0.12)	20.07 (±4.75)	55:45	7.71 (±6.66)	74.40%
Stanford Training (N) ($n = 291$)	0.46 (±0.16) $p < 0.005$	15.67 (±2.72)		1.83 (±0.53)	98.46% ($p < 0.005$)
Stanford Validation (G) ($n = 48$)	0.77 (±0.13) $p = 0.0856$	19.60 (±4.80) $p = 0.492$	49:51 ($p = 0.3240$)	7.73 (±4.30) $p = 0.982$	77.70% ($p = 0.5971$)
Stanford Validation (N) ($n = 39$)	0.53 (±0.20) $p = 0.0422$	14.20 (±3.89) $p = 0.01$		1.72 (±2.46) $p = 0.54$	98.30% ($p = 0.9468$)
Stanford Test (G) ($n = 157$)	0.79 (±0.19) $p = 0.5262$	19.56 (±5.47) ($p = 0.244$)	49:51 ($p = 0.2194$)	6.37 (±4.46) ($p = 0.0205$)	77.01% ($p = 0.6191$)
Stanford Test (N) ($n = 113$)	0.45 (±0.16) $p = 0.4764$	16.00 (±2.72) ($p = 0.153$)		1.12 (±1.07) ($p < 0.005$)	98.06% ($p = 0.7678$)
Hong Kong (G) ($n = 277$)	No Data	16.19 (±4.17) ($p < 0.005$)	67:33 ($p < 0.005$)	6.44 (±4.21) ($p < 0.005$)	79.83% ($p = 0.5239$)
Hong Kong (N) ($n = 196$)	No Data	13.44 (±2.72) ($p < 0.005$)		1.46 (±0.30) ($p < 0.005$)	99.61% ($p = 0.2346$)
India (G) ($n = 171$)	No Data	No Data	40:60 ($p < 0.005$)	7.68 (±3.81) ($p = 0.951$)	65.38% ($p = 0.0331$)
India (N) ($n = 121$)	No Data	No Data		2.54 (±1.39) ($p < 0.005$)	93.17% ($p = 0.006$)
Nepal (G) ($n = 166$)	No Data	16.56 (±4.74) ($p < 0.005$)	40:60 ($p < 0.005$)	5.37 (±3.30) ($p < 0.005$)	77.00% ($p = 0.5791$)
Nepal (N) ($n = 181$)	No Data	15.68 (±2.90) ($p = 0.972$)		1.99 (±1.08) ($p = 0.051$)	97.58% ($p = 0.5362$)

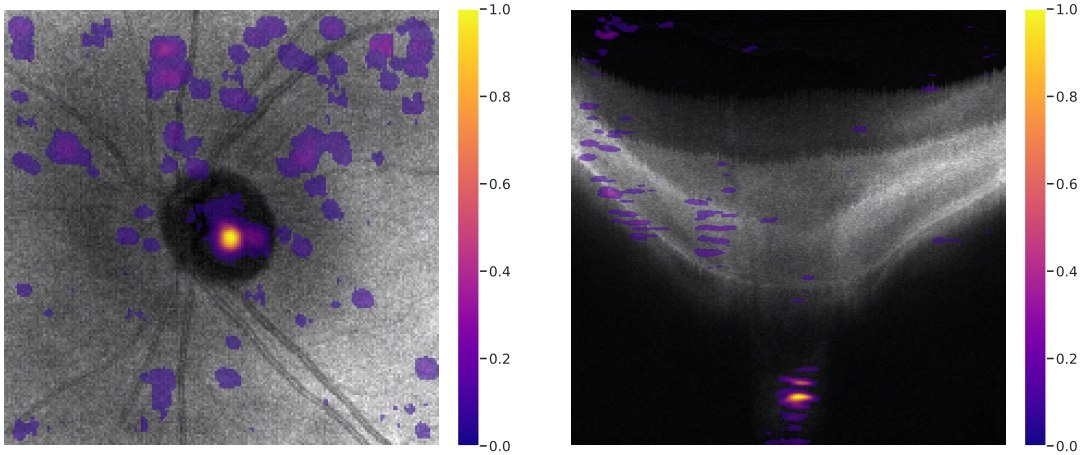
Table S12: Results of the proposed model on the Stanford and external test sets. 95% confidence intervals are computed over 5 independent runs of the model.

Dataset	AUC, 95% CI	Sensitivity at 90% Specificity, 95% CI	Sensitivity at 95% Specificity, 95% CI
Stanford	0.9098 (0.9027-0.9168)	77.48% (73.68-81.29)	73.02% (69.23-76.81)
Hong Kong	0.8023 (0.7838-0.8208)	61.69% (59.40-63.98)	56.16% (51.89-60.44)
India	0.9444 (0.9319-0.9569)	85.00% (81.10-88.90)	77.48% (71.06-83.89)
Nepal	0.8738 (0.8515-0.8961)	68.02% (63.24-72.80)	59.46% (53.85-65.07)

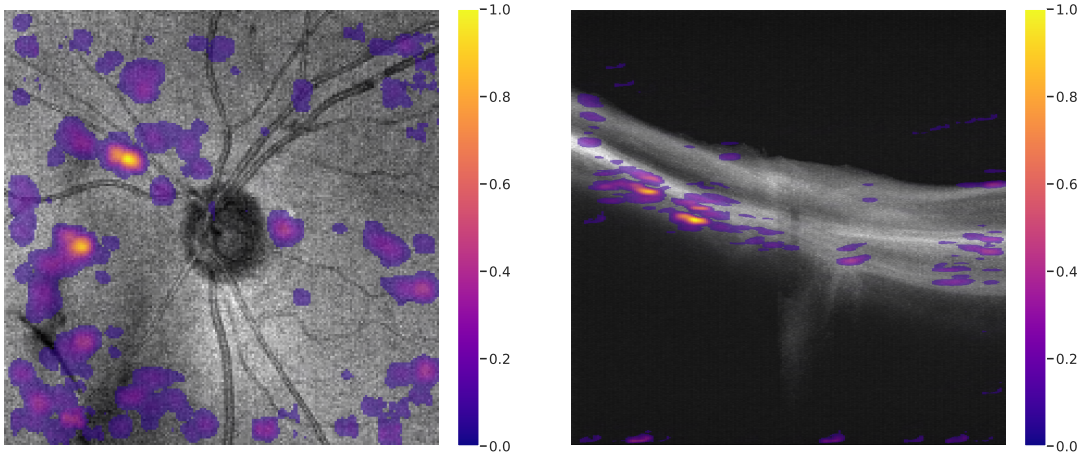
Table S13: Results of the proposed model on the Stanford and external test sets, divided by each eye. A p-value of 0.005 or less was considered significant. 95% confidence intervals are computed over 5 independent runs of the model. Z test has been used to measure the p-value of the difference in metric values between the different eyes from each test set.

Dataset	No. scans (eyes)	AUC, 95% CI	Sensitivity, 95% CI	Specificity, 95% CI	F1 Score, 95% CI
Stanford (Right)	349 (138)	0.91 (0.90-0.92)	84.00% (77.14-90.86)	79.66% (72.09-87.24)	0.86 (0.83-0.89)
Stanford (Left)	345 (132)	0.92 (0.90-0.93)	87.89% (82.91-92.87)	75.90% (62.31-89.49)	0.87 (0.87-0.88)
Stanford (p-value)		0.0818	0.2026	0.5020	0.2967
Hong Kong (Right)	756 (228)	0.79 (0.76-0.82)	70.30% (64.34-76.27)	74.65% (64.24-85.06)	0.74 (0.72-0.76)
Hong Kong (Left)	869 (245)	0.81 (0.80-0.82)	75.15% (68.60-81.70)	71.51% (56.72-86.30)	0.78 (0.77-0.78)
Hong Kong (p-value)		0.0319	0.1288	0.6302	0.0000
India (Right)	303 (142)	0.94 (0.92-0.96)	93.21% (87.76-98.67)	72.94% (51.47-94.41)	0.92 (0.90-0.93)
India (Left)	317 (141)	0.95 (0.94-0.96)	93.03% (87.47-98.58)	69.29% (49.87-88.71)	0.90 (0.89-0.91)
India (p-value)		0.3812	0.9478	0.7264	0.0189
Nepal (Right)	208 (177)	0.88 (0.85-0.90)	81.25% (69.76-92.74)	79.17% (65.48-92.85)	0.81 (0.79-0.84)
Nepal (Left)	199 (163)	0.87 (0.84-0.90)	77.27% (66.16-88.39)	78.20% (64.48-91.93)	0.79 (0.76-0.82)
Nepal (p-value)		0.6657	0.4898	0.8901	0.1109

S2 Supplementary Figures

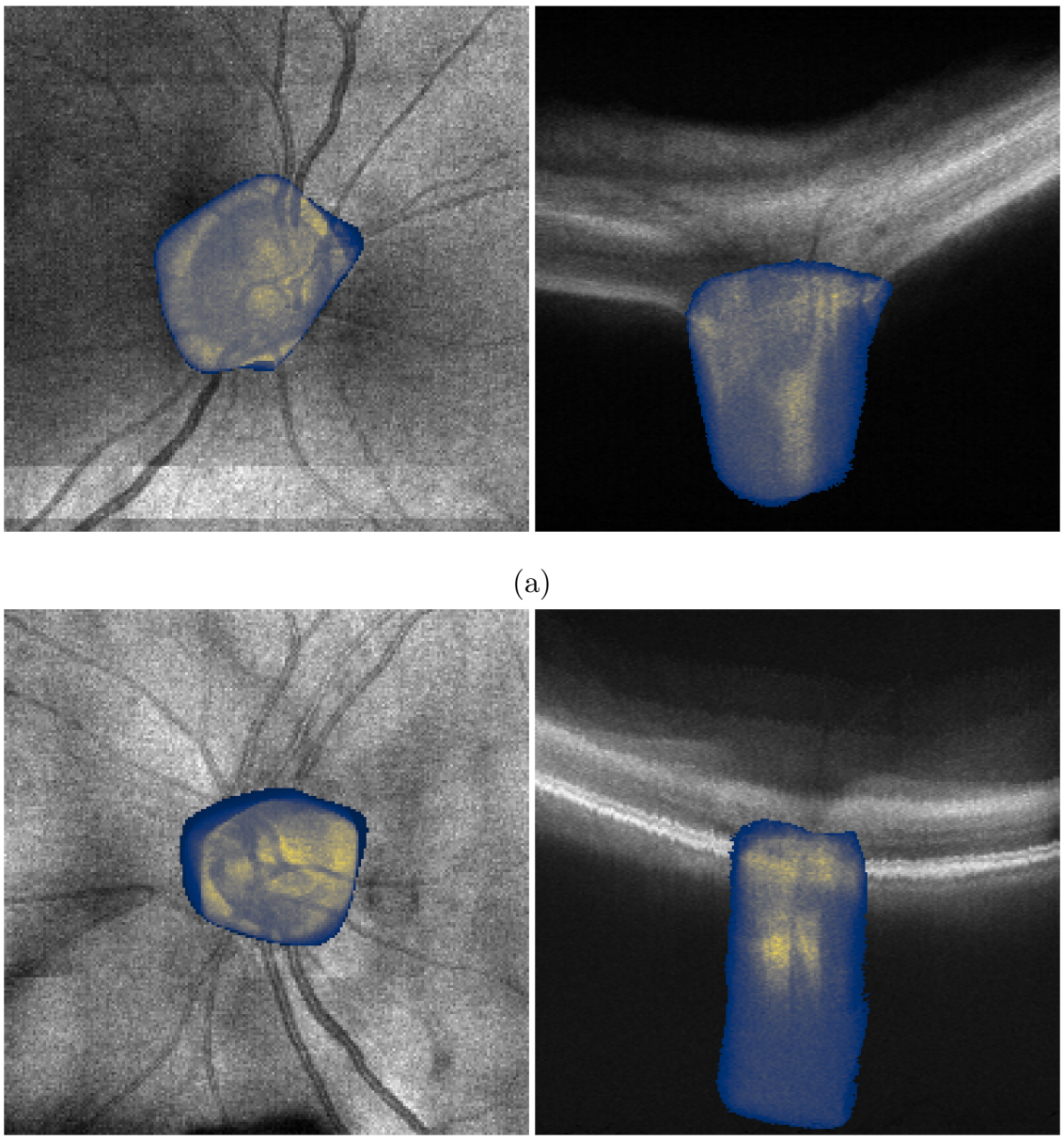


(a) (b)



(c) (d)

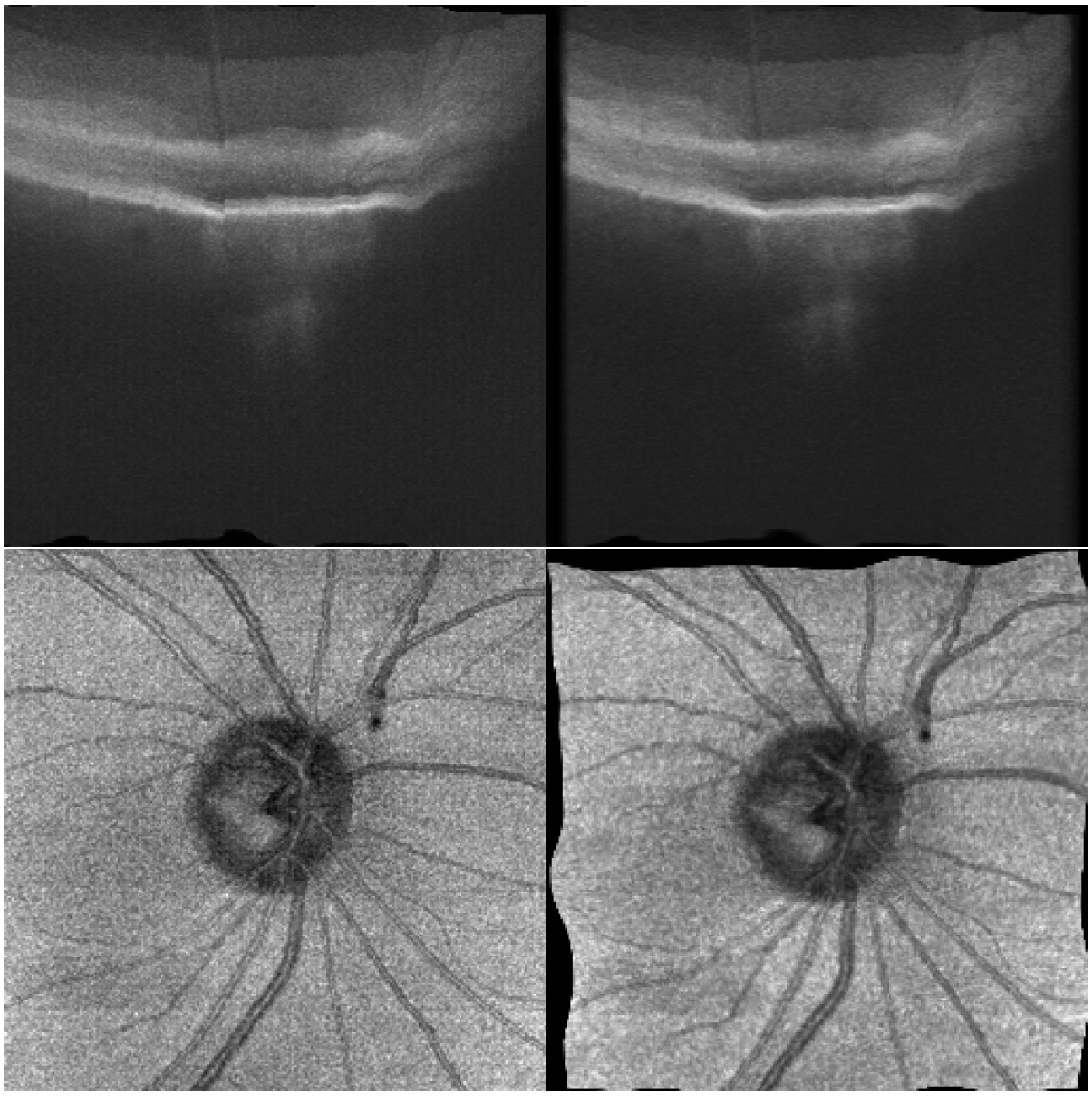
Figure S1: Saliency visualizations for two cases from the Stanford test set with wrong predictions. **(a)** Top, and **(b)** Side side view of saliency visualizations of a false positive case, where Lamina Cribrosa is highlighted, even though the case has normal ground truth label. **(c)** Top, and **(d)** Side view of saliency visualizations of a false negative case, where the retina is highlighted despite the case having glaucoma ground truth. Saliency visualization have been obtained with respect to the predicted class. Regions with higher value are more salient for the model in making the final prediction.



(a)

(b)

Figure S2: Visualization of the cropped scans, overlaid on the un-cropped scans. Top row **(a)** shows a normal scan and the bottom row **(b)** shows a glaucoma scan.



(a) (b)

Figure S3: **(a)** Original OCT scans. **(b)** Elastic Deformation applied to the OCT scans. Darker regions are tissues in the eye that are less transparent against the light beamed to the eye.

S3 Development of the Deep Learning Algorithm

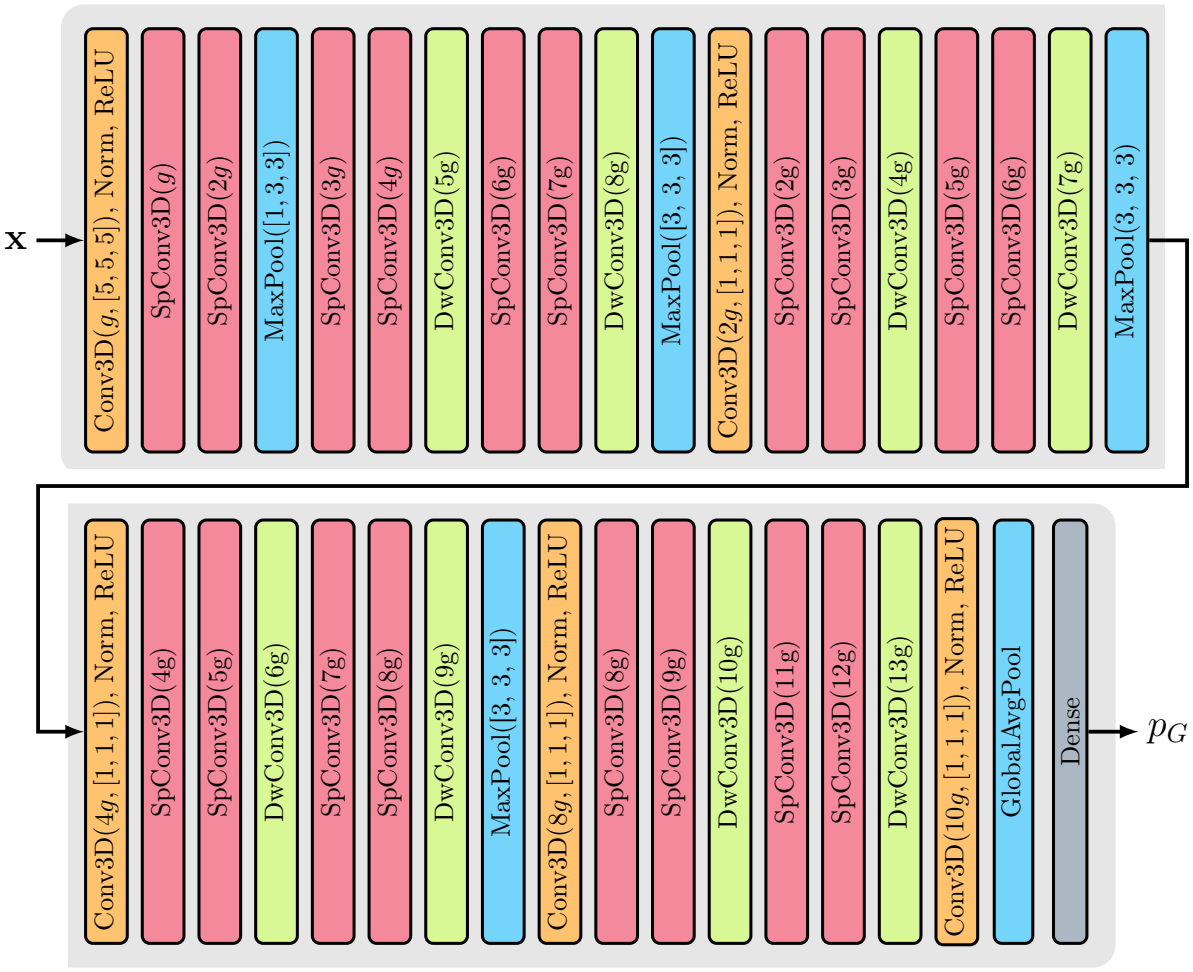


Figure S4: Architecture of the neural network used in the present study. **Norm** indicates Group Normalization, and g is a hyperparameter which was set to 16. **SpConv3D** and **DwConv3D** are defined in Figure 1a and Figure 1b, respectively.

S3.1 Finding Areas of Interest

Our hypothesis was that if a glaucoma detection network can achieve a performance better than a random classifier, given only the cropped scan, then it would show that the ONH area indeed includes informative signals for glaucoma detection. Remember the cohort includes real world scans even with lower signal strength.

Since manually cropping 3D OCT volumes is laborious and time-consuming, we only annotated 100 OCT scans with equal number of normal and glaucoma cases. Therefore, we were neither able to train our network on cropped data nor searching over hyperparameters was possible. To solve the first issue, we applied extra cropping data augmentation to make the model more robust against partial data. In this data augmentation, we randomly selected a smaller volume, and set the values outside the volume to zero. To mitigate the latter issue, we trained the best performing model from random initialization with the additional data augmentation and used it to get numbers on the cropped scan test set.

S4 Labeling Details

S4.1 Training, Validation, and Test sets from Stanford, and Datasets from India and Nepal

The inclusion criteria were (1) age equal to or older than 18 years old; (2) reliable visual field (VF) tests; and (3) availability of SD-OCT Optic Disc scans. A reliable visual field report is defined as (a) fixation losses less than 33%; (b) false positive rate less than 25%; (c) false negative rate less than 25%; and (d) no appearance of lid or lens rim artifacts, and no appearance of cloverleaf patterns. SD-OCT scans with signal strength less than 3 or any artifact obscuring imaging of the ONH, or any artifacts or missing data areas that prevented measuring the thickness of the RNFL at 3.4 mm diameter were excluded from the study. Artifacts included blink, motion, registration, and mirror artifacts. Eyes with optic nerve head pathologies, such as non-glaucomatous optic neuropathy, optic nerve head hypoplasia, or optic nerve pit, and other retinal pathologies such as retinal detachment, age-related macular degeneration, myopic macular degeneration, macular hole, diabetic retinopathy, and arterial and venous obstruction were carefully excluded.

S4.2 Hong Kong Dataset

For SD-OCT data from the Hong Kong test set, two trained medical students and a postgraduate ophthalmology trainee (with more than 3 years' of experience in Glaucoma) did the initial quality control and then graded the SD-OCT scans into gradable or non-gradable SD-OCT scans, according to the aforementioned criteria. Two glaucoma specialists then worked separately to label all the eyes with gradable SD-OCT scans into Normal/Glaucoma combined with VF results. Most of the images were labeled as normal/glaucoma when the two graders arrived at the same categorization separately, but a few disagreeable cases were reviewed by a senior Glaucoma specialist to make the final decision. Ungradable SD OCT scan was defined as when: signal strength < 5, or any artifacts affected the measurement circle or $> 25\%$ of peripheral area. Artifacts included: off-centeration, out of registration, missing OCT signal, motion, mirror artifacts, and blurriness. An SD-OCT volumetric scan was labeled as gradable when: signal strength was ≥ 5 without any of the aforementioned artifacts; or when the artifacts influenced $< 25\%$ of peripheral area, excluding the measurement center.