AMERICAN ACADEMY OF OPHTHALMOLOGY®

Detecting Glaucoma from Fundus Photographs Using Deep Learning without Convolutions: Transformer for Improved Generalization

Rui Fan, PhD,[1,2,3] *Kamran Alipour, PhD,*[2] *Christopher Bowd, PhD,*[1] *Mark Christopher, PhD,*[1] *Nicole Brye,*[1] *James A. Proudfoot, MS,*[1] *Michael H. Goldbaum, MD,*[1] *Akram Belghith, PhD,*[1] *Christopher A. Girkin, MD,*[4] *Massimo A. Fazio, PhD,*[1,4,5] *Jeffrey M. Liebmann, MD,*[6] *Robert N. Weinreb, MD,*[1] *Michael Pazzani, PhD,*[2] *David Kriegman, PhD,*[2] *Linda M. Zangwill, PhD*[1]

Purpose: To compare the diagnostic accuracy and explainability of a Vision Transformer deep learning technique, Data-efficient image Transformer (DeiT), and ResNet-50, trained on fundus photographs from the Ocular Hypertension Treatment Study (OHTS) to detect primary open-angle glaucoma (POAG) and identify the salient areas of the photographs most important for each model's decision-making process.

Design: Evaluation of a diagnostic technology.

Subjects, Participants, and Controls: Overall 66 715 photographs from 1636 OHTS participants and an additional 5 external datasets of 16 137 photographs of healthy and glaucoma eyes.

Methods: Data-efficient image Transformer models were trained to detect 5 ground-truth OHTS POAG classifications: OHTS end point committee POAG determinations because of disc changes (model 1), visual field (VF) changes (model 2), or either disc or VF changes (model 3) and Reading Center determinations based on disc (model 4) and VFs (model 5). The best-performing DeiT models were compared with ResNet-50 models on OHTS and 5 external datasets.

Main Outcome Measures: Diagnostic performance was compared using areas under the receiver operating characteristic curve (AUROC) and sensitivities at fixed specificities. The explainability of the DeiT and ResNet-50 models was compared by evaluating the attention maps derived directly from DeiT to 3 gradient-weighted class activation map strategies.

Results: Compared with our best-performing ResNet-50 models, the DeiT models demonstrated similar performance on the OHTS test sets for all 5 ground-truth POAG labels; AUROC ranged from 0.82 (model 5) to 0.91 (model 1). Data-efficient image Transformer AUROC was consistently higher than ResNet-50 on the 5 external datasets. For example, AUROC for the main OHTS end point (model 3) was between 0.08 and 0.20 higher in the DeiT than ResNet-50 models. The saliency maps from the DeiT highlight localized areas of the neuroretinal rim, suggesting important rim features for classification. The same maps in the ResNet-50 models show a more diffuse, generalized distribution around the optic disc.

Conclusions: Vision Transformers have the potential to improve generalizability and explainability in deep learning models, detecting eye disease and possibly other medical conditions that rely on imaging for clinical diagnosis and management. *Ophthalmology Science 2023;3:100233* © *2022 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org

Primary open-angle glaucoma (POAG) is a blinding but treatable disease in which damage to the optic nerve can result in progressive and irreversible vision loss. In 2010, there were an estimated 1.4 billion people worldwide with myopia, and the prevalence is rapidly rising to an estimated 4.75 billion by 2050.[1,2] As the population ages, the number of people with POAG will increase worldwide to an estimated 111.8 million by 2040.[3,4] Because its symptoms often only occur when the disease is severe, early detection and treatment play an important role in preventing visual impairment and blindness from the disease.

Digital fundus photography is an important modality for detecting glaucoma. Benefiting from the recent advances in artificial intelligence (AI), deep learning (DL) models using fundus photographs have achieved compelling results for glaucoma detection.[3,5–9]

© 2022 by the American Academy of Ophthalmology
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Published by Elsevier Inc.

https://doi.org/10.1016/j.xops.2022.100233
ISSN 2666-9145/22

Convolutional neural networks (CNNs), the most common and prevalent type of DL model, have dominated this research area. Although CNNs have achieved high accuracy in glaucoma detection, there is still a burning problem: *CNNs often do not generalize well on unseen fundus data*. This is likely due, in part, to differences in glaucoma ground truth and study populations across datasets and to different cameras, illuminations, and photographer skills/preferences. There is considerable evidence that the assessment of optic nerve head photographs for glaucoma determination is highly variable, even among glaucoma experts.[3] The development of glaucoma is also related to factors, such as race and age, which can vary significantly across datasets. Therefore, there is a need to develop techniques to enhance a DL model's ability to generalize across fundus datasets.

From the algorithm aspect, the state-of-the-art (SoTA) CNNs typically employ convolutional layers to represent fundus photographs with 1-dimensional visual features.[10] A fully connected layer is simultaneously fine-tuned to classify these visual features as either glaucoma or healthy.[11] Such visual features are learned without encoding the connection between pixels,[12] and therefore, CNNs can be easily misled when the visual feature of an unseen image is similar to one of the learned examples, even if they have significantly different spatial structures.

Transformers were initially introduced for machine translation,[13] and they have since become the SoTA method in many natural language processing tasks.[12] The explanation of their success in natural language processing tasks lies in the adopted self-attention mechanism,[13] which differentially weighs the significance of each part of the sequential input data. Unlike recurrent neural networks, Transformers do not necessarily process the data in order. In contrast, the self-attention mechanism provides context for any position in the input sequence, showing the capability to understand the connection between inputs (such as the words in a sentence when a Transformer is applied for natural language processing).[13] However, when using Transformers to evaluate images, applying self-attention between pixels is often challenging.

Vision Transformer (ViT)[12] was recently introduced to tackle this problem. It divides an image into a grid of square patches. Each patch is flattened into a single vector by concatenating the channels of all pixels in a patch and then linearly projecting it to the desired input dimension. Because ViT is agnostic to the structure of the input elements, learnable position embeddings were added to each patch to enable the model to learn about the structure of the images. A ViT does not know about the relative location of patches in the image or even that the image has a 2-dimensional structure—it learns relevant information from the training data and encodes structural information in the position embeddings. As opposed to convolutional layers whose receptive field is a small neighborhood grid, the self-attention layer's receptive field is always the full image. Therefore, ViT can learn many more global visual features.[12]

However, ViT models performed slightly worse than ResNet models of comparable sizes when the training set sample size was small.[12] This is likely due at least in part because ViT models lack some of the inductive biases inherent to CNNs (such as translation equivariance and locality) and do not generalize well when trained on insufficient amounts of data. In contrast, when the training data are sufficient, ViT models can overcome the inductive biases and can outperform the SoTA CNNs significantly. In this regard, ViT is typically pretrained with hundreds of millions of images, thereby limiting their adoption. Data-efficient image Transformer (DeiT)[14] was recently proposed to overcome this limitation. Data-efficient image Transformer introduces a teacher-student strategy specific to Transformers. It relies on a distillation token to ensure that the student model learns from the teacher model through attention. Compared with ViT, DeiT achieves better performance in terms of both model accuracy and training efficiency without the large sample size requirement.[14] The main objective of this study was to compare the accuracy of widely used ResNet-50 and new SoTA DeiT models for the detection of glaucoma from expert graded fundus photographs and one of the most extensive clinical trials for glaucoma prevention, the Ocular Hypertension Treatment Study (OHTS).

In addition to high accuracy, it is also essential that AI algorithms provide mechanisms to explain their decisions. In vision-based AI, explanation techniques, such as attention maps and saliency maps, have become very popular in recent years.[15–17] Saliency maps can highlight parts of the input that are most influential on AI's outcome; however, these techniques have shown limitations in the interpretability of AI, especially in medical applications.[18] In contrast, attention maps in AI algorithms are usually generated during the inference and can more directly identify the parts of input considered as important. This paper also objectively compares saliency maps and attention maps from the DeiT algorithm to the saliency maps from the ResNet-50 classifier. The goal of this analysis was to provide information to developers and validators on the regions considered by the deep learner to ensure that it is not focusing on irrelevant areas that will lead to poor performance.

Methods

Description of Study Populations and Datasets

This study compares the accuracy, generalizability, and explainability of the SoTA Transformer DeiT and CNN ResNet-50 models for detecting glaucoma from OHTS fundus photographs and 5 external independent datasets. University of California, San Diego–based Diagnostic Innovations in Glaucoma Study/the multicenter African Descent and Glaucoma Evaluation Study and OHTS studies adhered to the tenets of the Declaration of Helsinki and were approved by the institutional review boards of UC San Diego and all other study sites. Other data were assembled from publicly available datasets.

The OHTS,[19,20] a large randomized clinical trial of 1636 subjects with ocular hypertension, was designed to determine the safety and efficacy of topical ocular hypotensive medication in delaying or preventing the onset of POAG in ocular hypertensive eyes. The OHTS was unique in that the primary POAG end point, of a reproducible clinically significant POAG optic disc changes or a reproducible glaucomatous visual field (VF) defect, was decided

Table 1. Characteristics of the Ocular Hypertension Treatment Study Training, Validation, and Test Sets (n [%] or mean [95% CI])[11]

Measurement	Train/Validation	Test	P Value
Age (yrs)	56.8 (56.3, 57.4)	57.2 (56.1, 58.2)	0.924
Number of participants (eyes)	1314 (2628)	322 (644)	
Number of eye visits	29 644	7088	
Self-reported race			
European descent	991 (75.4%)	238 (73.9%)	0.566
African descent	323 (24.6%)	84 (26.1%)	
Sex			
Female	758 (57.7%)	173 (53.7%)	0.209
Male	556 (42.3%)	149 (46.3%)	
Baseline visual field MD (dB)	−0.03 (−0.11, 0.05)	−0.12 (−0.27, 0.04)	0.239
Baseline photograph based vertical cup-to-disc ratio	0.39 (0.38, 0.40)	0.39 (0.37, 0.41)	0.735
No glaucoma			
Number of participants (eyes*)	1093 (2344)	250 (240)	
Number of eye visits	27 966	6675	
Visual field MD (dB)	−0.18 (−0.23, −0.12)	−0.16 (−0.28, −0.05)	0.871
Developed a POAG end point by visual field or photograph			
Number of participants (eyes)	221 (284)	72 (96)	
Number of eye visits	1678	96	

CI = confidence interval; dB = decibels; MD = mean deviation; POAG = primary open-angle glaucoma.
*Eyes without glaucoma are included from patients with and without glaucoma.

by a 3-member masked end point committee of glaucoma experts who reviewed clinical information and both the photographs and VFs to determine whether observed changes were because of POAG or another disease. The end point committee reviewed cases only after changes in the optic disc and VF from baseline were determined by masked readers at the independent Optic Disc Reading Center and VF Reading Center. We trained DL models on the OHTS fundus photographs to detect the following 5 outcomes from the end point committee and Optic Disc Reading Center and VF Reading Center POAG determinations:

- End point committee determination:

Model 1: optic disc changes attributable to POAG.
Model 2: VF changes attributable to POAG.
Model 3: optic disc or VF changes attributable to POAG.

- Reading Center determination:

Model 4: optic disc changes attributable to POAG by Optic Disc Reading Center.
Model 5: VF changes attributable to POAG by VF Reading Center.

The DL models were then evaluated on the following 5 independent external test sets: (1) the Diagnostic Innovations in Glaucoma Study, United States dataset and the African Descent and Glaucoma Evaluation Study, San Diego, CA, Birmingham, AL, and New York City, NY, United States dataset[21]; (2) the public fundus dataset funded by the Ministerio de Economía y Competitividad of Spain (ACRIMA) (Spain)[22] dataset; (3) the Large-Scale Attention-Based Glaucoma (LAG, China)[23] dataset; (4) the Retinal IMage database for Optic Nerve Evaluation (Spain)[24] dataset; and (5) The Online Retinal Fundus Image Dataset for Glaucoma Analysis and Research (Singapore)[25] dataset.

Dataset Preparation and Augmentation

We performed the same preprocessing and data augmentation strategies to prepare the fundus photograph dataset for DL model training as we used in our previous work.[11] In brief, a region centered on the optic nerve head was first extracted from each raw fundus photograph using a semantic segmentation network. A square region surrounding the extracted optic nerve head was then automatically cropped from each image and resized to 224 × 224 pixels for input in the DL model. During the DL model training, several data augmentation strategies, such as random rotation, translation, and horizontal flipping, were applied to increase the amount and type of variation of the training set.

DL Model

In our experiments, a DeiT,[14] pretrained on the ImageNet database,[26] was trained to detect the 3 end point committee and the 2 Reading Center committee POAG determinations. We modified the last layer of the pretrained DeiT to produce 2 scalars, indicating the probability distribution of healthy and POAG classes, respectively.

Data-efficient image Transformer[14] is a convolution-free Transformer. A fundus image is split into a collection of 16 × 16 pixel patches, which are then embedded linearly (with position embeddings included). A distillation token interacts with the class and the patch tokens through the self-attention layers. This distillation token is used similarly to the class token, which minimizes a cross-entropy loss L_{CE}, except that, on the output of the network, its objective is to reproduce the hard label predicted by the teacher (by minimizing another cross-entropy loss $L_{teacher}$) instead of a true label. Both the class and distillation tokens input to the transformers is learned by backpropagation (Fig S1, available at www.ophthalmologyscience.org).

Model Training and Selection

The demographic and clinical characteristics of the OHTS participants in training/validation and test sets are outlined in Table 1. We conducted our experiments with the identical training setup as our previous work,[11] including (a) the same OHTS training, validation, and test sets randomly chosen by the patient; (b) 5 POAG determinations, which are 3 end point committee determinations (models 1−3) and 2 Reading Center determinations (models 4 and 5); (c) the same strategy to reduce the class imbalance problem;

Table 2. Diagnostic Accuracy of DeiT and ResNet-50 Performance[11] in Identifying POAG by the OHTS End Point Committee and Optic Disc and VF Reading Centers

Ground Truth Determined by	POAG Detection Modality	POAG (n) Subjects (Eyes)/Visits	AUROC (95% CI)					
			DeiT			ResNet-50		
			All Eyes	Early Glaucoma (VF MD ≥ −6 dB)	Severe Glaucoma (VF MD < −6 dB)	All Eyes	Early Glaucoma (VF MD ≥ −6 dB)	Severe Glaucoma (VF MD < −6 dB)
End point committee	Optic disc photograph and/or VF	52 (71)/352	0.88 (0.82, 0.92)	0.87 (0.80, 0.91)	0.82 (0.65, 0.94)	0.88 (0.82, 0.92)	0.86 (0.79, 0.91)	0.83 (0.63, 0.95)
	Optic disc photograph	41 (56)/262	0.91 (0.87, 0.93)	0.90 (0.87, 0.93)	0.77 (0.51, 0.93)	0.91 (0.88, 0.94)	0.90 (0.87, 0.94)	0.81 (0.60, 0.95)
	VF	35 (41)/195	0.84 (0.75, 0.90)	0.82 (0.73, 0.89)	0.74 (0.54, 0.89)	0.86 (0.76, 0.93)	0.83 (0.70, 0.91)	0.81 (0.56, 0.93)
Reading centers	Optic disc photograph	60 (77)/318	0.86 (0.83, 0.89)	0.86 (0.82, 0.89)	0.70 (0.49, 0.89)	0.89 (0.85, 0.92)	0.87 (0.83, 0.91)	0.81 (0.57, 0.96)
	VF	61 (78)/242	0.82 (0.76, 0.87)	0.80 (0.73, 0.86)	0.66 (0.43, 0.82)	0.83 (0.76, 0.88)	0.80 (0.72, 0.86)	0.68 (0.49, 0.87)

AUROC = areas under the receiver operating characteristic curve; dB = decibels; CI = confidence interval; DeiT = Data-efficient image Transformer; MD = mean deviation; OHTS = Ocular Hypertension Treatment Study; POAG = primary open-angle glaucoma; VF = visual field.

and (d) the same metric (F-score) to select the best-performing models from the validation set. The models were trained on an NVIDIA GeForce RTX 2080Ti graphics processing unit, which has an 11 GB GDDR6 memory and 4352 CUDA cores.

The minibatch sizes for DeiT and ResNet-50 training are set to 40 and 110, respectively, to optimize graphics processing unit memory use. The maximum epoch is set to 200. We also implemented an early stopping mechanism to reduce over-fitting in which the training of DeiT or ResNet is terminated if the F-score has not increased for 10 epochs. We used the stochastic gradient descent optimization algorithm to minimize the training loss. The initial learning rate is set to 0.001, which decays gradually after the 100th epoch. Hyperparameters were chosen based on commonly used values and empirical testing on the training data. Our source code is available in PyTorch on request at https://github.com/visres-ucsd/vision-transformer.

Performance Evaluation and Statistical Analysis

The model performance of the 5 different POAG ground truths was evaluated based on the area under the receiver operating curves (AUROC) within 95% confidence intervals (CIs) calculated using clustered bootstrapping techniques. Sensitivity (recall) was calculated at 80%, 85%, 90%, and 95% fixed specificity values. We also evaluated the performance in eyes with early glaucoma (eyes with a VF mean deviation −6 decibels [dB]) on the OHTS test set. Areas under the receiver operating characteristic curve scores for classifying healthy eyes versus all glaucoma eyes and early glaucoma eyes were computed for each model. The best-performing DeiT models were evaluated on the OHTS test set and 5 external datasets.

To evaluate the explainability of models, we compared the differences in the explanatory feature maps from DeiT and ResNet-50. This comparison was made based on the attention maps derived directly from the DeiT and different gradient-weighted class activation maps from ResNet-50, including GradCAM,[17] GradCAM++,[27] and ScoreCAM.[28] In gradient-based saliency maps, we highlight the pixels with the highest impact on the decision-making process based on activation maps in the last layer of the transformer. For DeiT, we can visualize both the attention from the transformer and the saliency maps for the last layer of the transformer, whereas for the ResNet-50 model, only the saliency maps are available. Additional details of these explainability approaches are provided in the Supplemental Methods (available at https://www.ophthalmologyscience.org/).

Role of Funding Source

The funding sources had no involvement in the study design, collection, analysis, or interpretation of data or in the writing of this manuscript.

Results

Table 2 and Figure 1 show the performance of our trained DeiT models on the OHTS test set in terms of the following: (a) classifying healthy versus all glaucoma eyes and (b) classifying healthy versus mild glaucoma eyes. Compared with our best-performing ResNet-50 models 10 (Table 2), the DeiT models demonstrated similar performance on the OHTS test sets for all 5 types of glaucoma determinations (see Fig 1A). The DeiT and ResNet-50 AUROC of models 1 and 3 were both 0.91 and 0.88, respectively, and differences in AUROC between the 2 DL strategies ranged from 0.01 to 0.03 for models 2, 4,

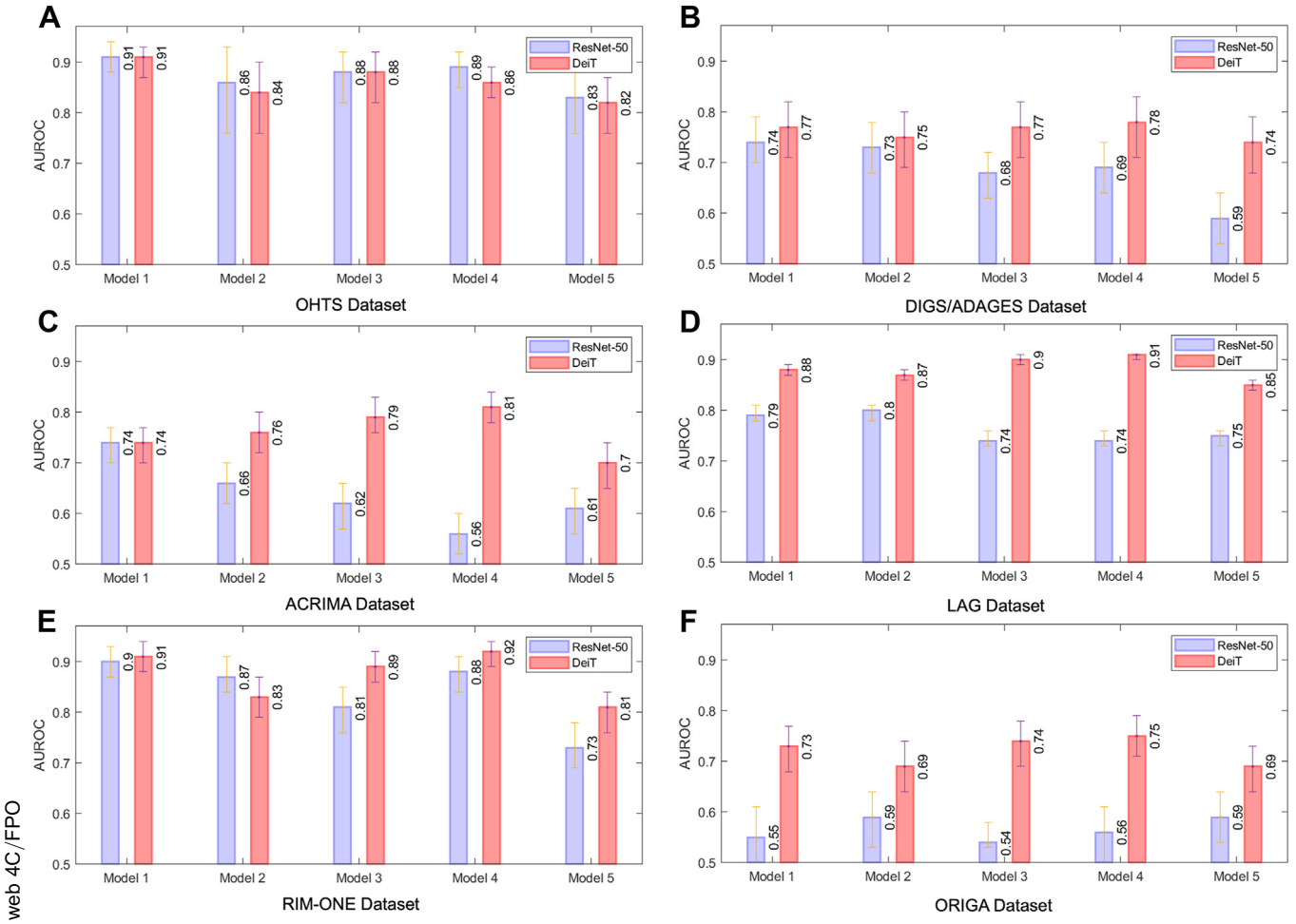


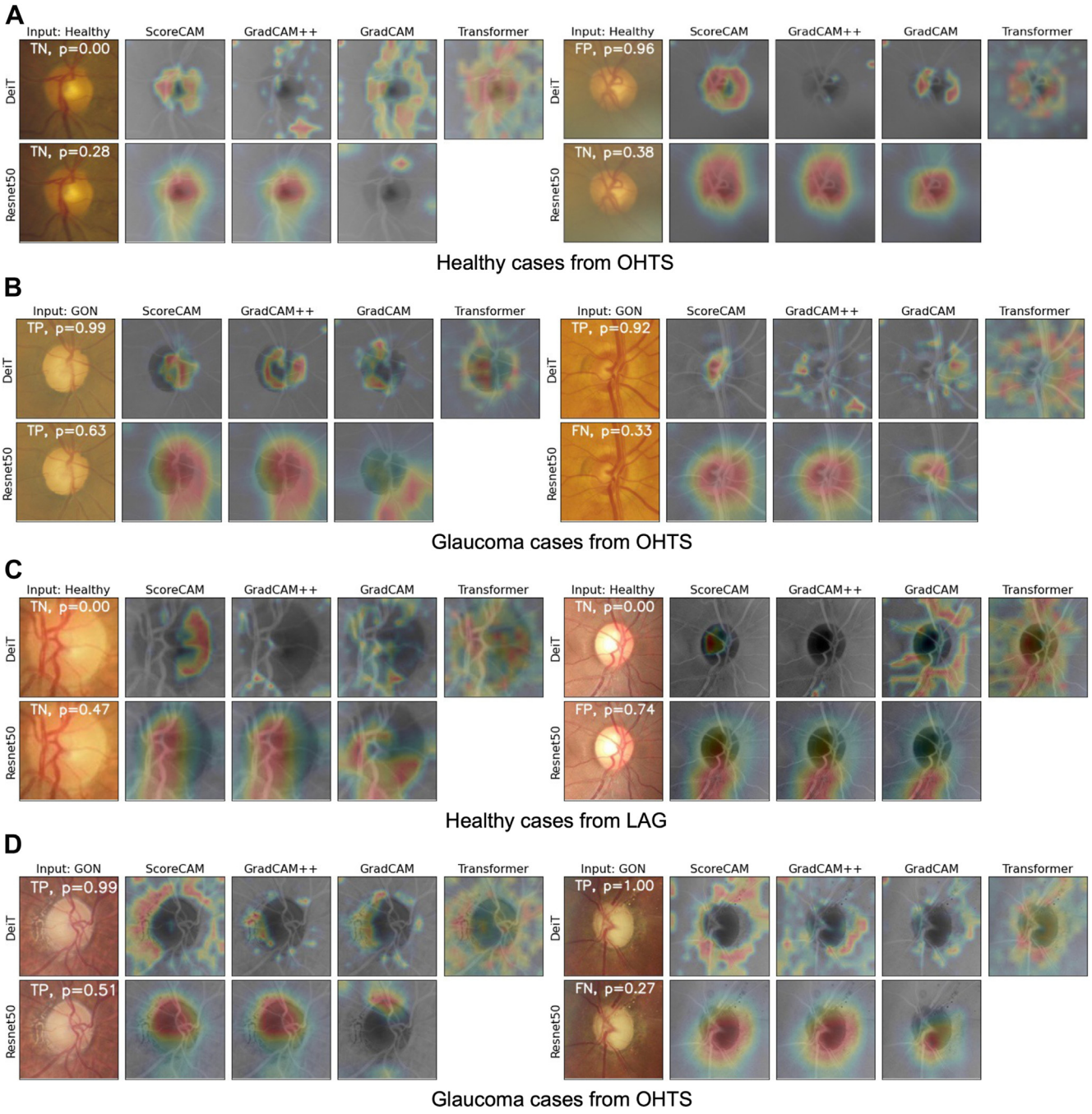
Figure 1. Comparison of accuracy between ResNet-50 and Data-efficient image Transformer (DeiT) for model 1, model 2, model 3, model 4, and model 5 on (**A**) OHTS 5 additional test sets: (**B**) DIGS/ADAGES, (**C**) ACRIMA, (**D**) LAG, (**E**) ORIGA, and (**F**) RIM-ONE. ACRIMA = public fundus dataset funded by the Ministerio de Economía y Competitividad of Spain; ADAGES = The African Descent and Glaucoma Evaluation Study; AUROC = areas under the receiver operating characteristic curve; DIGS = Diagnostic Innovations in Glaucoma Study; LAG = Large-Scale Attention-Based Glaucoma; OHTS = Ocular Hypertension Treatment Study; ORIGA = Online Retinal Fundus Image Dataset for Glaucoma Analysis and Research; RIM-ONE = Retinal IMage database for Optic Nerve Evaluation.

and 5. It is worth noting that the number of eyes incorrectly classified as POAG (false-positives) was much higher than the number of end points missed (false-negative). Specifically, the ratio of false-positives to false-negatives at 90% specificity ranges from 7.0 for model 1 to 3.75 for model 4.

We also compared the generalizability of DeiT and ResNet-50 models for detecting the 5 POAG end points on 5 external independent fundus photograph test sets (Fig 1B–F, Table 3). These results suggest that DeiT significantly outperforms ResNet-50 in almost all cases. Specifically, when evaluated on the LAG test set, the AUROC (95% CI) of the 5 POAG end points by DeiT ranged from 0.08 (model 2) to 0.16 (model 4), higher than that of ResNet-50. Data-efficient image Transformer achieves the best generalizability on the LAG test set (AUROC [95% CI], 0.91 [0.90, 0.91]) for the Reading Center's determination of change based on optic disc photographs, compared with 0.74 (0.73, 0.76) by ResNet-50. In addition, ResNet-50 performs best with respect to model 1, but its AUROC (95% CI) of 0.79 (0.78, 0.81) is significantly lower than that achieved by DeiT (AUROC [95% CI], 0.88 [0.87, 0.89]). Moreover, both DeiT and ResNet-50 do not generalize well on the Online Retinal Fundus Image Dataset for Glaucoma Analysis and Research test set. However, DeiT still significantly outperforms ResNet-50 in terms of the diagnostic AUROC (95% CI) of the POAG attribution in all cases. The end point committee's POAG attribution by optic disc photographs on the Retinal IMage database for Optic Nerve Evaluation test set is the only case in which ResNet-50 performs slightly better than DeiT (0.87 and 0.83, respectively, Fig 1E). In addition, confusion matrices show that the number of false-positive classifications is higher than false-negative ones in the external datasets (Fig S2, available at www.ophthalmologyscience.org).

Model 1 attention and saliency maps that are for selected healthy and glaucoma examples are shown in Figure 2, whereas a comparison of the average attention and saliency maps are shown in Figure 3. In comparing

web 4C/FPO



web 4C/FPO

Figure 2. Comparisons between Data-efficient image Transformer (DeiT) and ResNet-50 in predicting model 1 for OHTS and LAG datasets. These randomly selected samples demonstrate a tendency for a more localized feature attendance to the neuroretinal rim in DeiT (**A**–**D**) compared with the saliencies of the center of the image/whole optic disc from ResNet-50 (**A**, **B**, **D**). This greater attention to the details is enforced by the local image patches defined in the transformer layers of DeiT, whereas ResNet-50 is generally limited to learning the correlation between higher-level features because of multiple layers of convolution. LAG = Large-Scale Attention-Based Glaucoma; OHTS = Ocular Hypertension Treatment Study.

attention and saliency maps between the 2 AI algorithms used in this study, ResNet-50 shows a general sensitivity to the central part of the image, whereas DeiT is more focused on localized features around the borders of the optic disc and neuroretinal rim. Average attention and saliency maps for other models are shown in Figures S3 and S4 (available at www.ophthalmologyscience.org).

Discussion

We found that DeiT trained and validated on the OHTS dataset performs similarly to ResNet-50 in detecting POAG on the OHTS test set, regardless of which of the 5 ground-truths POAG determinations were used. Furthermore, DeiT outperforms ResNet-50 with respect to different POAG

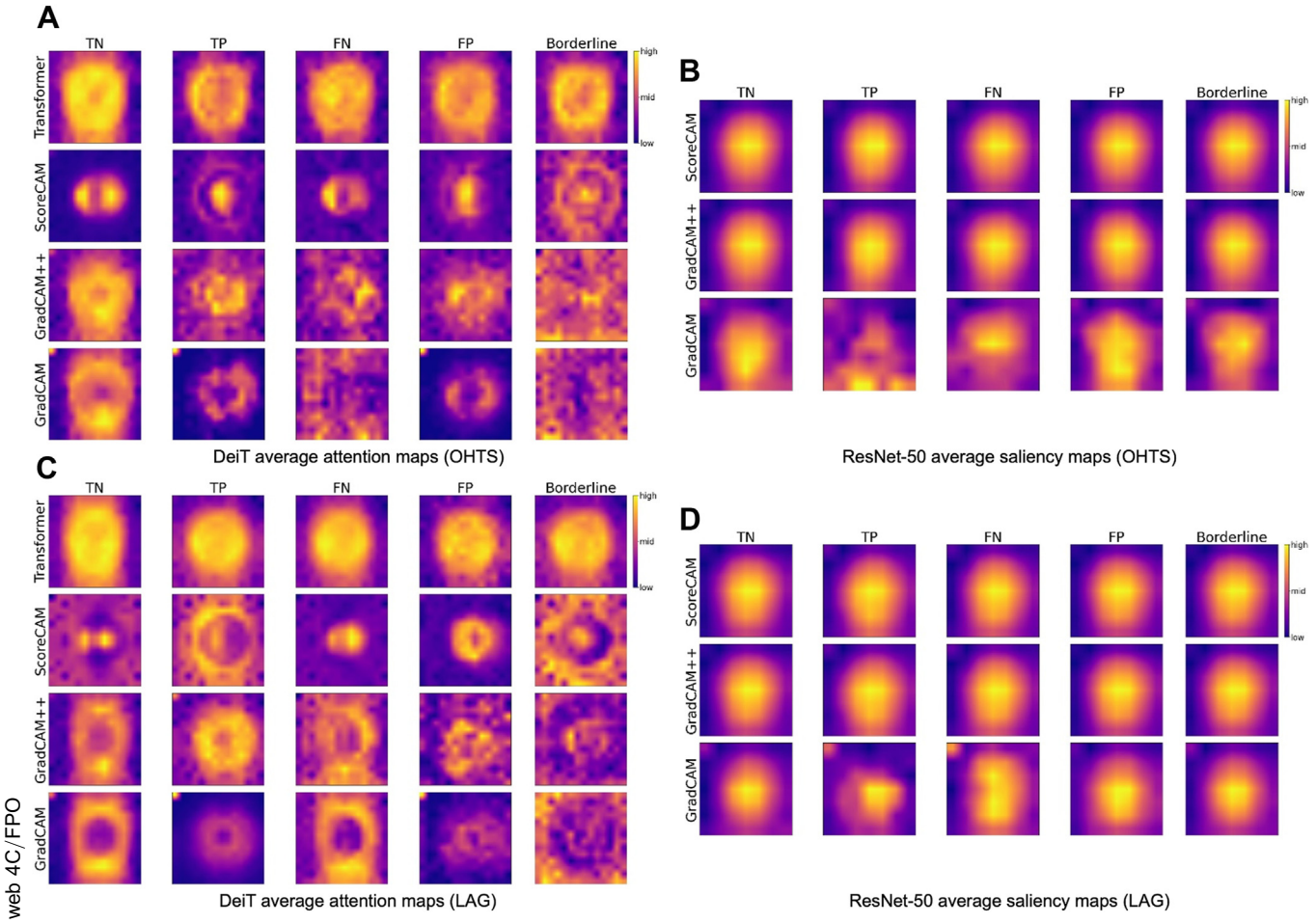


Figure 3. The averages of attention/saliency maps are compared between Data-efficient image Transformer (DeiT) (**A** and **C**) and ResNet-50 (**B** and **D**) for model 1 in OHTS and LAG datasets. In each subfigure, we categorize results based on model outcome (borderline). The borderline cases are those in which the model result is between 0.3 and 0.7. The saliency maps in the ResNet-50 model show a normal distribution around the center of the image. In contrast, the same maps from the DeiT indicate high sensitivity to the borders of the disc, suggesting a better understanding of important features. See Supplemental Methods (available at www.ophthalmologyscience.org) for similar figures on other labels (model 4, model 2, and model 5). LAG = Large-Scale Attention-Based Glaucoma; OHTS = Ocular Hypertension Treatment Study.

determination strategies on almost all the external independent test sets included in this study. Specifically, DeiT generalizes better than ResNet-50 in the diverse external test sets of fundus photographs, representing individuals of Chinese, Japanese, Spanish, African, and European descent, each with its own criteria for ground-truth determination of glaucoma. These results suggest that DeiT may be particularly useful to improve the generalizability of AI models in clinical applications because it outperforms the SoTA CNN ResNet-50 developed based on convolutions. Most importantly, biases in training sets (ground truth, study population, types of photographs, cameras, etc.) have been shown to lead to inaccurate detection results,[11] Data-efficient image Transformer with improved generalizability may serve as a critical tool to limit the effect of possible biases in training sets used for the detection of eye disease.

It is unclear why DeiT outperforms ResNet-50 when applied to external datasets. It is possible that CNNs such as ResNet-50, which rely on the relationship between pixels, may be sensitive to pixel-level noise in specific datasets, which reduces diagnostic accuracy when the model is applied to external datasets. Vision Transformers, which rely more on the entire image, may therefore provide more generalizable results.

In general, diagnostic accuracy using AI for the detection of glaucoma from fundus photographs and OCT images is worse in external datasets than in test sets from the original data source.[5] Others have reported better diagnostic accuracy on some of the external fundus photograph test sets used in the current study.[7,8,29–32] However, these reports trained and tested the datasets, so the diagnostic accuracy is expected to be higher than when the fundus photographs are independent external test sets as in the current study. Different glaucoma definitions between the external datasets may also explain the differences in performance (e.g., VF vs. expert photograph review, cup-to-disc ratio, etc.). It should be noted that the OHTS was designed to maximize specificity in its end point committee determinations of POAG.[33] For this reason, it is not surprising that there were more false-positive than false-negative errors in the external datasets

Table 3. Diagnostic Accuracy of DeiT and ResNet-50 Performance in Test Datasets for Model 1

	Test Set Size Photos		DeiT					ResNet-50				
				Sensitivity at Specificity of					Sensitivity at Specificity of			
	Healthy	Glaucoma	AUROC (95% CI)	80%	85%	90%	95%	AUROC (95% CI)	80%	85%	90%	95%
OHTS	6675	413	0.91 (0.87, 0.93)	0.83	0.79	0.70	0.56	0.91 (0.88, 0.94)	0.86	0.81	0.73	0.56
DIGS	5184	4289	0.77 (0.71, 0.82)	0.62	0.57	0.48	0.34	0.74 (0.69, 0.79)	0.59	0.52	0.43	0.30
ACRIMA	309	396	0.74 (0.70, 0.77)	0.57	0.46	0.38	0.31	0.74 (0.70, 0.77)	0.55	0.46	0.38	0.29
LAG	3143	1711	0.88 (0.87, 0.89)	0.81	0.77	0.70	0.59	0.79 (0.78, 0.81)	0.66	0.59	0.53	0.42
RIM-ONE	255	200	0.91 (0.88, 0.94)	0.85	0.83	0.81	0.73	0.90 (0.87, 0.93)	0.86	0.83	0.78	0.68
ORIGA	482	168	0.73 (0.68, 0.77)	0.48	0.40	0.35	0.21	0.55 (0.48, 0.61)	0.35	0.33	0.26	0.18

ACRIMA = public fundus dataset funded by the Ministerio de Economía y Competitividad of Spain; AUROC = areas under the receiver operating characteristic curve; CI = confidence interval; DeiT = Data-efficient image Transformer; DIGS = Diagnostic Innovations in Glaucoma Study; LAG = Large-Scale Attention-Based Glaucoma; OHTS = Ocular Hypertension Treatment Study; ORIGA = Online Retinal Fundus Image Dataset for Glaucoma Analysis and Research; RIM-ONE = Retinal IMage database for Optic Nerve Evaluation.

that had ground-truth determinations that were more likely to classify an eye as glaucoma (more sensitive) than the OHTS (Fig S2). We evaluated the alternative approaches on the ability to identify the most important regions of an image. This type of analysis explains in general in which the DL algorithm is focusing its attention. It is designed to give confidence to developers and validators that the system is not finding spurious correlations in less relevant parts of the image that will not generalize well to novel examples. Analyzing DL algorithms based on their saliency and attention maps provides a new perspective on explainable glaucoma detection compared with similar recent work that mostly focuses on post hoc occlusion,[34] cropping,[35] or adversarial examples[36] to explain the inference process. Although these post hoc editing methods can help expose the causal roots of a prediction, they are prone to noise imposed by the editing artifacts. In contrast, attention map explanations attempt to reflect the same information while preserving the original input and avoiding any unnatural artifacts in the inference process. Having been derived directly from the ViT modeling process, these attention maps may provide more direct information on their decision-making process. Regardless of the specific method, it is difficult to evaluate these explainability techniques with respect to specific cases in the absence of detailed, expert annotation. In this paper, we focused on the performance of these methods in general using attention and saliency maps averaged over many cases. In future work, we plan to incorporate detailed expert annotations of specific cases to further explore these techniques. A limited number of recent studies have investigated the general role of transformers in glaucoma detection and reported results comparable to ResNet.[37,38] To our knowledge, we are one of the first to use a data-efficient version of transformers, DeiT, to detect and explain glaucoma in fundus images. With DeiT, we showed better generalizability than ResNet and better explainability than saliency maps.

There are also several possible limitations to this study. First, the SoTA ViTs typically require a large amount of data to train. However, we achieved good results with a training set of 1636 patients (3272 eyes). Second, there were fewer eyes with POAG than without it, resulting in an imbalance in the dataset. Therefore, we implemented additional class weights into the model to address this imbalance problem. Third, we cropped all photographs, which may have reduced model performance if informative information was located in the peripheral retina.

Conclusion

In summary, this study comprehensively explored the generalizability and explainability of a cutting-edge AI technique, ViT, to detect glaucoma using fundus photographs. The extensive experimental results suggested that ViT generalizes well to the eyes of individuals of Chinese, Japanese, Spanish, African, and European descent represented in the external test sets of fundus photographs included in this study. Furthermore, ViT focused on localized features of the neuroretinal rim, which are often used in the clinical management of glaucoma. Vision Transformer has the potential to improve the scalability of DL solutions for the detection of not only eye disease but possibly also other conditions that require various imaging modalities for clinical diagnosis and management.

Footnotes and Disclosures

Originally received: June 17, 2022.
Final revision: October 4, 2022.
Accepted: October 12, 2022.
Available online: October 19, 2022.Manuscript no. XOPS-D-22-00131R1.

[1] Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, University of California San Diego, La Jolla, California.

[2] Department of Computer Science and Engineering, University of California San Diego, La Jolla, California.

[3] Department of Control Science and Engineering, Tongji University, Shanghai 201804, China.

[4] Department of Ophthalmology, School of Medicine, The University of Alabama at Birmingham, Birmingham, Alabama.

[5] Department of Biomedical Engineering, School of Engineering, The University of Alabama at Birmingham, Birmingham, Alabama.

[6] Bernard and Shirlee Brown Glaucoma Research Laboratory, Edward S. Harkness Eye Institute, Columbia University Medical Center, New York, New York.

Disclosures:

All authors have completed and submitted the ICMJE disclosures form.

The authors made the following disclosures: C.A.G: Grants or Contracts − National Eye Institute, EyeSight Foundation of Alabama, Research to Prevent Blindness, Heidelberg Engineering; Consultant − Topcon.

J.M.L.: Grant or Contracts − Bausch & Lomb, Carl Zeiss Meditec, Heidelberg Engineering, National Eye Institute, Novartis, Optovue, Reichert Technologies, Research to Prevent Blindness; Consultant − Alcon, Allergan, Bausch & Lomb, Carl Zeiss Meditec, Heidelberg Engineering, Reichert Technologies, Valeant Pharmaceuticals.

L.M.Z.: Grants or Contracts − National Eye Institute, Carl Zeiss Meditec, Heidelberg Engineering, Optovue, Topcon, The Glaucoma Foundation; Consultant − Abbvie, Digital Diagnostics; Patents − Carl Zeiss Meditec.

M.A.F.: Grants or Contracts − National Eye Institute, EyeSight Foundation of Alabama, Research to Prevent Blindness, Heidelberg Engineering; Consultant − Topcon.

M.C.: Support − National Eye Institute; Grants or Contracts − The Glaucoma Foundation.

R.N.W.: Grants or Contracts − Heidelberg Engineering, Carl Zeiss Meditec, Konan Medical, Optovue, Centervue, Bausch & Lomb, Topcon; Consultant − Abbvie, Aerie Pharmaceuticals, Allergan, Equinox, Nicox, Topcon; Patents − Toromedes, Carl Zeiss Meditec.

HUMAN SUBJECTS: Human subjects were included in this study. DIGS/ADAGES and OHTS studies adhered to the tenets of the Declaration of Helsinki and were approved by the IRBs of UC San Diego and all other study sites. Other data were assembled from publicly available datasets.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Fan, Alipour, Bowd, Christopher, Brye, Proudfoot, Goldbaum, Belghith, Girkin, Fazio, Liebmann, Weinreb, Pazzani, Kriegman, Zangwill.

Data collection: Fan, Alipour, Bowd, Christopher, Brye, Proudfoot, Goldbaum, Belghith, Girkin, Fazio, Liebmann, Weinreb, Pazzani, Kriegman, Zangwill.

Analysis and interpretation: Fan, Alipour, Bowd, Christopher, Brye, Proudfoot, Goldbaum, Belghith, Girkin, Fazio, Liebmann, Weinreb, Pazzani, Kriegman, Zangwill.

Obtained funding: Zangwill, Fazio, Girkin, Weinreb

Overall responsibility: Fan, Alipour, Bowd, Christopher, Brye, Proudfoot, Goldbaum, Belghith, Girkin, Fazio, Liebmann, Weinreb, Pazzani, Kriegman, Zangwill.

Abbreviations and Acronyms:
AI = artificial intelligence; **AUROC** = areas under the receiver operating characteristic curve; **CI** = confidence interval; **CNN** = convolutional neural network; **DeiT** = Data-efficient image Transformer; **DL** = deep learning; **LAG** = Large-Scale Attention-Based Glaucoma; **OHTS** = Ocular Hypertension Treatment Study; **POAG** = primary open-angle glaucoma; **SoTA** = state-of-the-art; **VF** = visual field; **ViT** = Vision Transformer.

Keywords:
Deep learning, Fundus photographs, Glaucoma detection, Vision Transformers.

Correspondence:
Linda M. Zangwill, 9500 Gilman Dr., #0946, La Jolla, California 92093-0946. E-mail: lzangwill@ucsd.edu.

References

1. Holden BA, Fricke TR, Wilson DA, et al. Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050. *Ophthalmology*. 2016;123:1036−1042.
2. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: a review. *JAMA*. 2014;311:1901−1911.
3. Thompson AC, Jammal AA, Medeiros FA. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. *Transl Vis Sci Technol*. 2020;9:42.
4. Tham YC, Li X, Wong TY, et al. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*. 2014;121:2081−2090.
5. Wu JH, Nishida T, Weinreb RN, Lin JW. Performances of machine learning in detecting glaucoma using fundus and retinal optical coherence tomography images: a meta-analysis. *Am J Ophthalmol*. 2022;237:1−12.
6. Christopher M, Belghith A, Bowd C, et al. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Sci Rep*. 2018;8:16685.
7. Liao W, Zou B, Zhao R, Chen Y, He Z, Zhou M. Clinical interpretable deep learning model for glaucoma diagnosis. *IEEE J Biomed Health Inform*. 2020;24:1405−1412.
8. Yu S, Xiao D, Frost S, Kanagasingam Y. Robust optic disc and cup segmentation with deep learning for glaucoma detection. *Comput Med Imaging Graph*. 2019;74:61−71.
9. Christopher M, Nakahara K, Bowd C, et al. Effects of study population, labeling and training on glaucoma detection using deep learning algorithms. *Transl* Vis Sci Technol. 2020;9:27, 27.
10. Fan R, Bowd C, Brye N, et al. [Preprint] One-vote veto: semi-supervised learning for low-shot glaucoma diagnosis. https://arxiv.org/abs/2012.04841. Accessed November 7, 2022.
11. Fan R, Bowd C, Christopher M, et al. [Preprint] Detecting glaucoma in the Ocular Hypertension Treatment Study using deep learning: implications for clinical trial endpoints. doi: 10.36227/techrxiv.14959947.v2. https://www.techrxiv.org/articles/preprint/Detecting_Glaucoma_in_the_Ocular_Hypertension_Treatment_Study_Using_Deep_Learning_Implications_for_clinical_trial_endpoints/14959947/1 (Accessed 7 November 2022). *TechRxiv*. 2022. https://doi.org/10.36227/techrxiv.14959947.v2
12. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations*. 2020.
13. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Adv Neural Inf Process Syst*. 2017:5998−6008.
14. Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention. In: *International Conference on Machine Learning*. 2021:10347−10357.
15. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: *European Conference on Computer Vision*. Springer; 2014:818−833.

16. Hendricks LA, Akata Z, Rohrbach M, et al. Generating visual explanations. In: *European Conference on Computer Vision*. Springer; 2016:3−19.
17. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017:618−626.
18. Saporta A, Gui X, Agrawal A, et al. [Preprint]. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *medRxiv; 2021.*
19. Gordon MO, Kass MA. Ocular Hypertension Treatment Study Group. The Ocular Hypertension Treatment Study: design and baseline description of the participants. *Arch Ophthalmol*. 1999;117:573−583.
20. Kass MA, Heuer DK, Higginbotham EJ, et al. The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Arch Ophthalmol*. 2002;120:701−713.
21. Sample PA, Girkin CA, Zangwill LM, et al. The African Descent and Glaucoma Evaluation Study (ADAGES): design and baseline data. *Arch Ophthalmol*. 2009;127:1136−1145.
22. Diaz-Pinto A, Morales S, Naranjo V, et al. CNNs for automatic glaucoma assessment using fundus images: an extensive validation. *Biomed Eng OnLine*. 2019;18:29.
23. Li L, Xu M, Wang X, Jiang L, Liu H. Attention based glaucoma detection: a large-scale database and CNN model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019:10571−10580.
24. Fumero F, Alayo'n S, Sanchez JL, et al. Rim-one: an open retinal image database for Optic Nerve Evaluation. In: *2011 24th International Symposium on Computer-based Medical Systems*. *IEEE*. 2011:1−6.
25. Zhang Z, Yin FS, Liu J, et al. ORIGA-light: an online retinal fundus image database for glaucoma analysis and research. In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. *IEEE*. 2010:3065−3068.
26. Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE*. 2009:248−255.
27. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks. In: *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision*. IEEE; 2018:839−847.
28. Wang H, Wang Z, Du M, et al. Score-cam: score-weighted visual explanations for convolutional neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020:24−25.
29. Fu H, Cheng J, Xu Y, Liu J. Glaucoma detection based on deep learning network in fundus image. In: *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*. Springer; 2019:119−137.
30. Nawaz M, Nazir T, Javed A, et al. An efficient deep learning approach to automatic glaucoma detection using optic disc and optic cup localization. *Sensors*. 2022;22:434.
31. Islam MT, Mashfu ST, Faisal A. Siam, deep learning-based glaucoma detection with cropped optic cup and disc and blood vessel segmentation. *IEEE Access*. 2021;10: 2828−2841.
32. Xu X, Guan Y, Li J, et al. Automatic glaucoma detection based on transfer induced attention network. Biomed *Eng OnLine*. 2021;20:1−19.
33. Gordon MO, Higginbotham EJ, Heuer DK, et al. Assessment of the impact of an end point committee in the Ocular Hypertension Treatment Study. *Am J Ophthalmol*. 2019;199: 193−199.
34. Christopher M, Bowd C, Belghith A, et al. Deep learning approaches predict glaucomatous visual field damage from oct optic nerve head en face images and retinal nerve fiber layer thickness maps. *Ophthalmology*. 2020;127:346−356.
35. Hemelings R, Elen B, Barbosa-Breda J, et al. Deep learning on fundus images detects glaucoma beyond the optic disc. *Sci Rep*. 2021;11:20313.
36. Chang J, Lee J, Ha A, et al. Explaining the rationale of deep learning glaucoma decisions with adversarial examples. *Ophthalmology*. 2021;128:78−88.
37. Yu S, Ma K, Bi Q, et al. Mil-vt: multiple instance learning enhanced Vision Transformer for fundus image classification. In: *International Conference on Medical Image Computing and Computer- Assisted Intervention*. Springer; 2021:45−54.
38. Song D, Fu B, Li F, et al. Deep relation transformer for diagnosing glaucoma with optical coherence tomography and visual field function. *IEEE Trans Med Imaging*. 2021;40: 2392−2402.