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

Rui Fan, Kamran Alipour, Christopher Bowd, Mark Christopher, Nicole Brye, James A. Proudfoot, Michael H. Goldbaum, Akram Belghith, Christopher A. Girkin, Massimo A. Fazio, Jeffrey M. Liebmann, Robert N. Weinreb, Michael Pazzani, David Kriegman, Linda M. Zangwill

PII: S2666-9145(22)00122-1
DOI: https://doi.org/10.1016/j.xops.2022.100233
Reference: XOPS 100233

To appear in: Ophthalmology Science

Received Date: 17 June 2022
Revised Date: 4 October 2022
Accepted Date: 12 October 2022


This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier Inc. on behalf of American Academy of Ophthalmology.
Detecting Glaucoma from Fundus Photographs Using Deep Learning without Convolutions: Transformer for Improved Generalization

Rui Fan¹,²,³, Kamran Alipour², Christopher Bowd¹, Mark Christopher¹, Nicole Brye¹, James A. Proudfoot¹, Michael H. Goldbaum¹, Akram Belghith¹, Christopher A. Girkin⁴, Massimo A. Fazio¹,⁴,⁵, Jeffrey M. Liebmann⁶, Robert N. Weinreb¹, Michael Pazzani², David Kriegman², and Linda M. Zangwill¹

1. Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, University of California San Diego, La Jolla, CA, United States.
2. Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, United States.
3. Department of Control Science and Engineering, Tongji University, Shanghai 201804, China P. R.
4. Department of Ophthalmology, School of Medicine, The University of Alabama at Birmingham, Birmingham, Alabama, United States.
5. Department of Biomedical Engineering, School of Engineering, The University of Alabama at Birmingham, Birmingham, Alabama, United States.
6. Bernard and Shirlee Brown Glaucoma Research Laboratory, Edward S. Harkness Eye Institute, Columbia University Medical Center, New York, United States.

Data sharing statement

Deidentified individual participant data that underlie the results reported in the article and data dictionaries, including the study protocol, will be made available upon request and approval of an analysis proposal and signed data access agreement to Dr. Mae Gordon (gordon.mae@wustl.edu), Director of the OHTS Data Coordinating Center. In the future, the OHTS data used in this report will be made available through the National Eye Institute Data Commons, which is currently under construction.
ABSTRACT

Purpose: To compare the diagnostic accuracy and explainability of a new 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 to identify in general the salient areas of the photographs most important for each model’s decision-making process.

Study Design: Evaluation of a diagnostic technology

Subjects, Participants, and/or Controls: 66,715 photographs from 1,636 OHTS participants and an additional five external datasets of 16137 photographs of healthy and glaucoma eyes.

Methods, Intervention, or Testing: DeiT models were trained to detect five ground truth OHTS POAG classifications: OHTS Endpoint Committee POAG determinations due to disc changes (Model 1), visual field changes (Model 2), or either disc or visual field changes (Model 3) and reading center determinations based on disc (Model 4) and visual fields (Model 5). The best-performing DeiT models were compared to ResNet-50 on OHTS and five 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 gradient-weighted class activation map generation strategies.

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

Conclusions: Vision transformer has the potential to improve the generalizability and explainability of deep learning models for the detection of eye disease and possibly other medical conditions that rely on imaging modalities for clinical diagnosis and management.
Introduction

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. As the population ages, the number of people with POAG will increase worldwide to an estimated 111.8 million by 2040. As 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.

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/or to different cameras, illuminations, and photographer skills/preferences. There is considerable evidence that the assessment of optic nerve head (ONH) photographs for glaucoma determination is highly variable, even among glaucoma experts. 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 one-dimensional (1-D) visual features. A fully connected layer is simultaneously fine-tuned to classify these visual features as either glaucoma or healthy. Such visual features are learned without encoding the connection between pixels, 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, and they have since become the SoTA method in many natural language processing (NLP) tasks. The explanation of their success in NLP tasks lies in the adopted self-attention mechanism, which differentially weighs the significance of each part of the sequential input data. Unlike recurrent neural networks (RNNs), 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 NLP). However, when using Transformers to evaluate images, applying self-attention between pixels is often challenging.

Vision Transformer (ViT) 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. Since 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-D structure — it learns relevant information from the training data and encodes structural information in the position embeddings. As opposed to convolution 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.

However, ViT models performed slightly worse than ResNet models of comparable sizes when the training set sample size was small. 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 is sufficient, ViT models can overcome the inductive biases and can outperform the SoTA CNNs significantly. In this regard, ViT is typically pre-trained with hundreds of millions of images, thereby limiting their adoption. Data-efficient image Transformer (DeiT) was recently proposed to overcome this limitation. DeiT 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 to...
ViT, DeiT achieves better performance in terms of both model accuracy and training efficiency without the large sample size requirement\textsuperscript{13}. The main objective of this study is to compare the accuracy of widely used ResNet-50 and new state-of-the-art 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\textsuperscript{14–16}. 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\textsuperscript{17}. 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 is 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 five external independent datasets. 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.

The OHTS\textsuperscript{18, 19}, a large randomized clinical trial of 1,636 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 endpoint, of a reproducible clinically significant POAG optic disc changes or a reproducible glaucomatous visual field (VF) defect, was decided by a three-member masked Endpoint Committee of glaucoma experts who reviewed clinical information and both the photographs and VFs to determine whether observed changes were due to POAG or another disease. The Endpoint Committee reviewed cases only after changes in the optic disc and/or VF from baseline were determined by masked readers at the independent Optic Disc Reading Center (ODRC) and VF Reading Center (VFRC). We trained DL models on the OHTS fundus photographs to detect the following five outcomes from the Endpoint Committee and ODRC and VFRC POAG determinations:

- **Endpoint 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 ODRC.
  - Model 5: VF changes attributable to POAG by VFRC.

The DL models were then evaluated on the following five independent external test sets: (1) the University of California, San Diego (UCSD)-based Diagnostic Innovations in Glaucoma Study (DIGS, United States) dataset and the multi-center African Descent and Glaucoma Evaluation Study (ADAGES, San Diego, CA, Birmingham, AL and New York City, NY, United States) dataset\textsuperscript{20}, (2) the ACRIMA (Spain)\textsuperscript{21} dataset, (3) the Large-Scale Attention-based Glaucoma (LAG, China)\textsuperscript{22} dataset, (4) the Retinal IMage database for Optic Nerve Evaluation (RIM-ONE, Spain)\textsuperscript{23} dataset, (5) The Online Retinal Fundus Image Dataset for Glaucoma Analysis and Research (ORIGA, Singapore)\textsuperscript{24} 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\cite{10}. In brief, a region centered on the ONH was first extracted from each raw fundus photograph using a semantic segmentation network. A square region surrounding the extracted ONH 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.

**Deep Learning Model**

In our experiments, a DeiT\cite{13}, pre-trained on the ImageNet database\cite{25}, was trained to detect the three Endpoint committee and the two Reading Center committee POAG determinations. We modified the last layer of the pre-trained DeiT to produce two scalars indicating the probability distribution of healthy and POAG classes, respectively.

DeiT\cite{13} 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 \(L_{\text{CE}}\), 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 \(L_{\text{teacher}}\)), instead of a true label. Both the class and distillation tokens input to the transformers is learned by back-propagation (Supplemental Figure 1, available at https://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\cite{10}, including (a) the same OHTS training, validation, and test sets randomly chosen by the patient, (b) five POAG determinations: 3 Endpoint Committee Determinations (Models 1-3), and 2 Reading Center Determinations (Models 4-5), (c) the same strategy to reduce the class imbalance problem, (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 GPU, which has an 11 GB GDDR6 memory and 4352 CUDA cores.

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

**Performance Evaluation and Statistical Analysis**

The model performance of the five different POAG ground truths was evaluated based on the area under the receiver operating curves (AUROC) within 95% confidence intervals (CI) 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 MD -6 dB) on the OHTS test set. AUROC 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 five 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\textsuperscript{16}, GradCAM++\textsuperscript{26}, and ScoreCAM\textsuperscript{27}. 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, while 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 (a) classifying healthy versus all glaucoma eyes and (b) classifying healthy versus mild glaucoma eyes. Compared to our best-performing ResNet-50 models\textsuperscript{10} (Table 2), the DeiT models demonstrated similar performance on the OHTS test sets for all five types of glaucoma determinations (see Figure 1a). The DeiT and ResNet-50 AUROC of Model 1 and Model 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 Model 2, Model 4, and Model 5. It is worth noting that the number of eyes incorrectly classified as POAG (false positives) was much higher than the number of endpoints 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 five POAG endpoints on five external independent fundus photograph test sets (Figure 1b–f). 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 five POAG endpoints by DeiT ranged from 0.08 (Model 2) to 0.16 (Model 4), higher than that of ResNet-50. DeiT 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 to 0.74 (0.73, 0.76) by ResNet-50. Additionally, ResNet-50 performs best w.r.t. Model 1 but its AUROC (95% CI): 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 ORIGA 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 Endpoint Committee’s POAG attribution by optic disc photographs on the RIM-ONE test set is the only case in which ResNet-50 performs slightly better than DeiT (0.87 and 0.83, respectively, Figure 1e). In addition, confusion matrices show that the number of false positive classifications is higher than false negative ones in the external datasets (Supplemental Figure 2, available at https://www.ophthalmologyscience.org/).

Model 1 attention and saliency maps for selected healthy and glaucoma examples are shown in Figure 2, while a comparison of the average attention and saliency maps are shown in Figure 3. In comparing attention and saliency maps between the two AI algorithms utilized in this study, ResNet-50 shows a general sensitivity to the central part of the image while 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 Supplemental Figures 3 and 4 (available at https://www.ophthalmologyscience.org/).

**Discussion**

We found that the convolution-free Vision Transformer 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 five
ground-truths POAG determinations were used. Furthermore, DeiT outperforms ResNet-50 w.r.t. different POAG 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 as 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\textsuperscript{10}, DeiT 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 whole image, may therefore provide more generalizable results.

In general, diagnostic accuracy using artificial intelligence for the detection of glaucoma from fundus photographs and optical coherence tomography images is worse in external datasets than in test sets from the original data source\textsuperscript{4}. Others have reported better diagnostic accuracy on some of the external fundus photograph test sets used in the current study\textsuperscript{6,7,31–34}. 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., visual field vs. expert photo review, cup-to-disc ratio, etc.). It should be noted that the OHTS was designed to maximize specificity in its endpoint committee determinations of POAG\textsuperscript{37}. For this reason, it is not surprising that there were more false positive than false negative errors in the external datasets that had ground truth determinations that were more likely to classify an eye as glaucoma (more sensitive) than the OHTS (Supplemental Figure 2). We evaluated the alternative approaches on the ability to identify the most important regions of an image. This type of analysis explains in general where the deep learning 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 deep learning algorithms based on their saliency and attention maps provides a new perspective on explainable glaucoma detection compared to similar recent work that mostly focuses on post-hoc occlusion\textsuperscript{28}, cropping\textsuperscript{29}, or adversarial examples\textsuperscript{30} to explain the inference process. While 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 Vision Transformer 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 work, 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 investigated the general role of transformers in glaucoma detection and reported results comparable to ResNet\textsuperscript{35, 36}. We are 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 Vision Transformers typically require a large amount of data to train. However, we achieved good results with a training set of 1636 patients (3272 eyes). Secondly, 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 is located in the peripheral retina.

**Conclusion**

In summary, this study comprehensively explored the generalizability and explainability of a cutting-edge AI technique: Vision Transformer, to detect glaucoma using fundus photographs. The extensive experimental results suggested that Vision Transformer 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, Vision Transformer 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 deep learning solutions for the detection of not only eye disease but possibly also other conditions that require various imaging modalities for clinical diagnosis and management.
References


Figure Captions

Figure 1. Comparison of accuracy between ResNet-50 and DeiT for Model 1, Model 2, Model 3, Model 4, and Model 5 on five additional test sets: DIGS/ADAGES, ACRIMA, LAG, ORIGA, and RIM-ONE.

Figure 2. Comparisons between 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, b, c, d) compared to 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, while ResNet-50 is generally limited to learning the correlation between higher-level features due to multiple layers of convolution.

Figure 3. The averages of attention/saliency maps are compared between DeiT and ResNet-50 for Model 1 in OHTS and LAG datasets. In each subfigure, we categorize results based on model outcome (TP, TN, FP, FN, borderline). The borderline cases are those where 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 Materials for similar figures on other labels (Model 4, Model 2, and Model 5).
Table 1. Characteristics of the Ocular Hypertension Treatment Study (OHTS) training, validation, and test sets. (n (%) or mean (95% CI))

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Train / Validation</th>
<th>Test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>56.8 (56.3, 57.4)</td>
<td>57.2 (56.1, 58.2)</td>
<td>0.924</td>
</tr>
<tr>
<td>Number of participants (eyes)</td>
<td>1314 (2628)</td>
<td>322 (644)</td>
<td></td>
</tr>
<tr>
<td>Number of eye visits</td>
<td>29644</td>
<td>7088</td>
<td></td>
</tr>
<tr>
<td>Self-reported Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>European Descent</td>
<td>991 (75.4%)</td>
<td>238 (73.9%)</td>
<td>0.566</td>
</tr>
<tr>
<td>African Descent</td>
<td>323 (24.6%)</td>
<td>84 (26.1%)</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Female)</td>
<td>758 (57.7%)</td>
<td>173 (53.7%)</td>
<td>0.209</td>
</tr>
<tr>
<td>(Male)</td>
<td>556 (42.3%)</td>
<td>149 (46.3%)</td>
<td></td>
</tr>
<tr>
<td>Baseline Visual Field Mean Deviation (MD) (dB)</td>
<td>-0.03 (-0.11, 0.05)</td>
<td>-0.12 (-0.27, 0.04)</td>
<td>0.239</td>
</tr>
<tr>
<td>Baseline Photograph Based Vertical Cup Disc Ratio</td>
<td>0.39 (0.38, 0.40)</td>
<td>0.39 (0.37, 0.41)</td>
<td>0.735</td>
</tr>
<tr>
<td><strong>No glaucoma</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of participants (eyes*)</td>
<td>1093 (2344)</td>
<td>250 (240)</td>
<td></td>
</tr>
<tr>
<td>Number of eye visits</td>
<td>27966</td>
<td>6675</td>
<td></td>
</tr>
<tr>
<td>Visual field MD (dB)</td>
<td>-0.18 (-0.23, -0.12)</td>
<td>-0.16 (-0.28, -0.05)</td>
<td>0.871</td>
</tr>
<tr>
<td><strong>Developed a POAG endpoint by visual field or photograph</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of participants (eyes)</td>
<td>221 (284)</td>
<td>72 (96)</td>
<td></td>
</tr>
<tr>
<td>Number of eye visits</td>
<td>1678</td>
<td>96</td>
<td></td>
</tr>
</tbody>
</table>

*Eyes w/o glaucoma are included from patients w/ and w/o glaucoma.
<table>
<thead>
<tr>
<th>Ground Truth determined by</th>
<th>POAG Detection Modality</th>
<th>POAG (n) Subjects (Eyes)/Visits</th>
<th>DeiT AUROC (95%CI)</th>
<th>ResNet50 AUROC (95%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All eyes</td>
<td>Early glaucoma (VF MD ≥ -6 dB)</td>
<td>Severe glaucoma (VF MD &lt; -6 dB)</td>
</tr>
<tr>
<td><strong>Endpoint Committee</strong></td>
<td>Optic Disc Photograph</td>
<td>52 (71)/352</td>
<td>0.88 (0.82, 0.92)</td>
<td>0.87 (0.80, 0.91)</td>
</tr>
<tr>
<td></td>
<td>Visual Field</td>
<td>41(56)/262</td>
<td>0.91 (0.87, 0.93)</td>
<td>0.90 (0.87, 0.93)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35(41)/195</td>
<td>0.84 (0.75, 0.90)</td>
<td>0.82 (0.73, 0.89)</td>
</tr>
<tr>
<td><strong>Reading Centers</strong></td>
<td>Optic Disc Photograph</td>
<td>60(77)/318</td>
<td>0.86 (0.83, 0.89)</td>
<td>0.86 (0.82, 0.89)</td>
</tr>
<tr>
<td></td>
<td>Visual Field</td>
<td>61(78)/242</td>
<td>0.82 (0.76, 0.87)</td>
<td>0.80 (0.73, 0.86)</td>
</tr>
</tbody>
</table>
Table 3. Diagnostic accuracy of DeiT and ResNet-50 performance in test datasets.

<table>
<thead>
<tr>
<th>Test set size photos</th>
<th>AUROC (95% CI)</th>
<th>Sensitivity at Specificity of:</th>
<th>AUROC (95% CI)</th>
<th>Sensitivity at Specificity of:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Healthy</td>
<td>Glaucoma</td>
<td>80%</td>
<td>85%</td>
</tr>
<tr>
<td>OHTS</td>
<td>6675</td>
<td>413</td>
<td>0.91 (0.87, 0.93)</td>
<td>0.83</td>
</tr>
<tr>
<td>DIGS</td>
<td>5184</td>
<td>4289</td>
<td>0.77 (0.71, 0.82)</td>
<td>0.62</td>
</tr>
<tr>
<td>ACRIMA</td>
<td>309</td>
<td>396</td>
<td>0.74 (0.70, 0.77)</td>
<td>0.57</td>
</tr>
<tr>
<td>LAG</td>
<td>3143</td>
<td>1711</td>
<td>0.88 (0.87, 0.89)</td>
<td>0.81</td>
</tr>
<tr>
<td>RIM-ONE</td>
<td>255</td>
<td>200</td>
<td>0.91 (0.88, 0.94)</td>
<td>0.85</td>
</tr>
<tr>
<td>ORIGA</td>
<td>482</td>
<td>168</td>
<td>0.73 (0.68, 0.77)</td>
<td>0.48</td>
</tr>
</tbody>
</table>
(a) Healthy cases from OHTS

(b) Glaucoma cases from OHTS

(c) Healthy cases from LAG

(d) Glaucoma cases from OHTS
Precis

In most cases, non-convolutional Vision Transformer models trained on Ocular Hypertension Treatment Study fundus photographs improved performance compared to convolutional ResNet-50 models for detecting glaucoma in 5 external fundus photograph datasets.