A Scanning Laser Ophthalmoscopy Image Database and Trustworthy Retinal Disease Detection Method

Yichen Hu[1], Chao Wang*[2], Weitao Song[2], Aleksei Tiulpin[3], and Qing Liu*[1,4]

[1] School of Computer Science, Central South University, China
[2] Xiangya Hospital of Central South University, China
[3] Research Unit of Health Sciences and Technology, University of Oulu, Finland
[4] Center for Machine Vision and Signal Analysis, University of Oulu, Finland
yichenhu@csu.edu.cn wangchao_csu@163.com
wtsong1980@126.com {aleksei.tiulpin, qing.liu}@oulu.fi

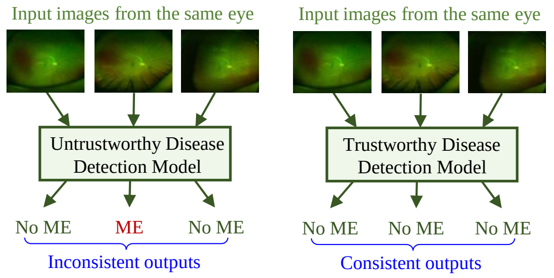
Abstract. Scanning laser ophthalmoscopy (SLO) images provide ophthalmologists with a non-invasive way to examine the retina for diagnostic and treatment purposes. Manual reading SLO images by ophthalmologists is a tedious task. Thus, developing trustworthy disease detection algorithms becomes urgent. However, up to now, there are no large-scale SLO image databases. In this paper, we collect and release a new SLO image dataset, named *Retina-SLO*, containing 7943 images of 4102 eyes from 2440 subjects with labels of three diseases, i.e., macular edema (ME), diabetic retinopathy (DR), and glaucoma. To our knowledge, *Retina-SLO* is the largest publicly available SLO image dataset for multiple retinal disease detection. While numerous deep learning-based methods for disease detection with medical images have been proposed, they ignore the model trust. Particularly, from a user's perspective, the detection model is highly untrustworthy if it makes inconsistent predictions on different SLO images of the same eye captured within relatively short time intervals. To solve this issue, we propose *TrustDetector*, a novel disease detection method, leveraging eye-wise consistency learning and rank-based contrastive learning to ensure consistent predictions and ordered representations aligned with disease severity levels on SLO images. Experimental results show that our *TrustDetector* achieves better detection performances and higher consistency than the state-of-the-arts. Dataset and code are available at https://drive.google.com/drive/TrustDetector/Retina-SLO.

Keywords: SLO images · eye-wise consistency · trustworthy detector

1 Introduction

Sight and vision help people to perceive their surroundings and provide vital information for human survival. Unfortunately, according to WHO, at least one

* Co-corresponding authors


Input images from the same eye
Untrustworthy Disease Detection Model
No ME
ME
No ME
Inconsistent outputs
Input images from the same eye
Trustworthy Disease Detection Model
No ME
No ME
No ME
Consistent outputs

(a) Untrustworthy model. (b) Trustworthy model.

Fig. 1: Illustration for trustworthy model and untrustworthy model defined in our paper.

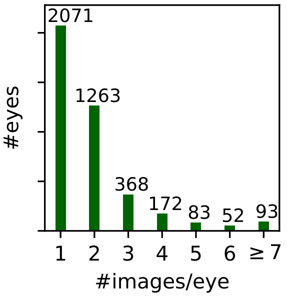


Fig. 2: Histogram of image number per eye of our dataset.

billion people around the world have a vision impairment that could have prevented it or is to be addressed [15] due to multiple factors such as insufficient eye care services and untimely interventions, etc. The causes include but are not limited to macular edema (ME), diabetic retinopathy (DR), glaucoma, etc. Clinically, scanning laser ophthalmoscopy (SLO) has been widely used as it provides ophthalmologists with a non-invasive way to assess the condition of symptomatic patients and screen for fundus diseases in a community setting. To relieve the workload of ophthalmologists and make eye care services available wider, automating the analysis of SLO images has an unmet clinical need.

During the past two decades, several SLO image datasets have been collected and automatic retinal disease diagnosis methods were proposed accordingly. For example, Haleem et al. collected a dataset containing 65 SLO images from 65 patients and proposed a novel computer-aided method based on regional image features for glaucoma detection [2]. In [22], a dataset named IOSTAR containing 30 SLO images for vessel segmentation has been made publicly available. More recently, Tang et al. collected a large-scale dataset containing 9392 ultra-widefield SLO images of 1903 eyes from 1022 subjects with diabetes and proposed to identify the vision-threatening diabetic retinopathy and referable diabetic retinopathy with ResNet-50 [16]. Besides, numerous retinal disease detection/grading methods focus on images captured by standard fundus cameras [9] and formulate the tasks as either a classification problem [8][7] or an ordinal regression problem [19].

However, these SLO image databases were either small in scale of patients and labels for retinal diseases or unreleased for public. Thus, it is necessity to build a large-scale SLO image database with labels for multiple diseases to foster the development of automated retinal disease diagnosis. From the view of users, previous methods simply treated the disease detection as a classification task and ignored the ability to make consistent predictions on different SLO images from the same eye. As illustrated in Fig. 1(a), for different SLO images from the same eye captured within a relatively short time interval, as no change happens on the patient's eye, the disease detection model becomes untrustworthy if it outputs different predictions. This seriously reduces the model reliability by users. To evaluate the reliability of the model, including multiple SLO images from patients' same eye in the new database is highly desired.

To start the research of trustworthy retinal disease detection with SLO images, in this paper, we first build a new large-scale SLO image dataset, which is comprised of 7943 SLO images of 4102 eyes from 2440 subjects. Specifically, 2071 eyes have one SLO image while 2031 eyes have at least two SLO images as shown in Fig. 2. Then, we propose a novel trustworthy disease detection method, named *TrustDetector*. To enforce the eye-wise consistency across different SLO images, we introduce an eye-wise consistency learning module, which pulls together the features of images from the same eye. Besides, considering the disease severity levels are ordinal, we introduce the rank-based contrastive learning module to enforce the ordinal distance in feature space be well ordered, increasing the feature discrinativeness of diseases at different severity levels. The contributions of this work can be briefly summarized as follows:

- **An open-source database**: We build an ultra-widefield SLO image database for trustworthy multiple retinal disease detection. Here, the trustability is defined as the ability of the model to make consistent predictions on multiple SLO images from the same eye.
- **A trustworthy disease detection method**: We design TrustDetector for multiple disease detection, in which eye-wise consistency learning module is proposed to learn eye-wise consistent features and rank-based contrastive learning module is proposed to learn an ordered representation in line with the disease severity levels. Experimental results show that our TrustDetector outperforms the state-of-the-arts in terms of both detection metrics such as accuracy, F-score and Kappa coefficients and consistency related metrics.

2 The SLO Image Dataset: Retina-SLO

Overview Following the standard clinical acquisition protocols, the SLO images are collected from 2440 patients who ever visited Ophthalmic Outpatient Department, Xiangya Hospital of Central South University between January of 2019 and December of 2022. All SLO images were captured with Optos Panoramic200 scanning laser ophthalmoscope. The study was approved by the Medical Ethics Committee of Xiangya Hospital (reference number: 202311944) and data are protected without disclosure of any personal information. Informed consent of the patients was waived due to the retrospective nature of the study.

Image Collection and Labeling 7943 images of 4102 eyes were collected. Specifically, as illustrated in Fig. 2, 2031 eyes have at least two SLO images, which enables us to consider the eye-wise consistency of the disease detection model. Among all the images, 7091 are of the size of 3900 × 3072 and 852 are of the size of 3072 × 3072.

Table 1: Data split where eyes$_{\text{single}}$ and eyes$_{\text{multiple}}$ are the numbers of eyes with single image and multiple images, respectively.

	train	**validation**	**test**	**total**
number of eyes	3053	541	508	4102
– eyes$_{\text{single}}$	1471	309	291	2071
– eyes$_{\text{multiple}}$	1582	232	217	2031
number of images	6129	942	872	7943

For each image, the labels for three diseases, i.e., macular edema (ME), diabetic retinopathy (DR), and glaucoma were determined via indexing the electronic medical record system by experienced ophthalmologists. The labels of ME and DR were binary. Differently, glaucoma at an ultra-early phase does not always have obvious clinical manifestations and is extremely difficult even for experienced ophthalmologists to make a confident diagnosis decision. For cases in which ophthalmologists could not make diagnostic decisions confidently, they assigned a label “suspicious”. Thus, the labels for the glaucoma included “glaucoma”, “suspicious” and “non-glaucoma”. For samples whose diagnosis decisions were recorded in the system, their diagnosis decisions were directly used as the disease labels. For samples whose diagnosis decisions were not recorded but their detailed medical treatments were well recorded, the disease labels were determined by experienced ophthalmologists according to the medical records in the system. Otherwise, the disease labels were marked as “unclear”.

Data Statistic Characteristics and Dataset Splits We split the dataset into three subsets, i.e., training, validation and test sets. To ensure that the disease distributions in three subsets are similar, the stratified sampling strategy is adopted. Particularly, the eyes with the same labels of three diseases are allocated into the same group. For each group, the eyes with the same number of images are then allocated into subgroups. For each subgroups with labels of “unclear”, 2/3 and 1/3 samples are randomly allocated to validation and test sets respectively. For subgroups without labels of “unclear”, we adopt random sampling without replacement strategy to ensure that around 80%, 10% and 10% eyes are allocated to train, validation and test sets respectively. Tab. 1 illustrates the numbers of eyes and SLO images in each subset and the class distribution for each set can be found in the supplementary.

Evaluation Metrics Similar to previous studies [10][19], metrics used to quantify the effectiveness of disease detection models on each disease are accuracy (Acc) and Cohen’s Kappa [13], and metrics for overall effectiveness evaluation are mean accuracy over tasks ($mAcc$) and mean Kappa over tasks ($mKappa$). As the class distribution of our Retina-SLO is extremely imbalanced, F1-score ($F1$) as a harmonic mean of specificity and selectivity, is used for the two-class classification of ME and DR, and macro-F1 [14] for the 3-class glaucoma grading. Similarly, the mean F1-score over tasks $mF1$ is used. Inspired by [17], two met-

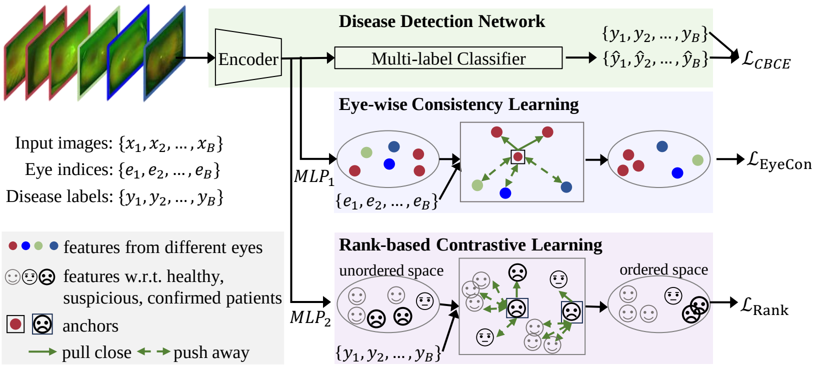

Disease Detection Network
Encoder
Multi-label Classifier
$\{y_1, y_2, \ldots, y_B\}$
$\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_B\}$
$\mathcal{L}_{CBCE}$
Eye-wise Consistency Learning
MLP_1
$\{e_1, e_2, \ldots, e_B\}$
$\mathcal{L}_{\mathrm{EyeCon}}$
Rank-based Contrastive Learning
unordered space
ordered space
MLP_2
$\{y_1, y_2, \ldots, y_B\}$
$\mathcal{L}_{\mathrm{Rank}}$
Input images: $\{x_1, x_2, \ldots, x_B\}$
Eye indices: $\{e_1, e_2, \ldots, e_B\}$
Disease labels: $\{y_1, y_2, \ldots, y_B\}$
● ● ● ● features from different eyes
☺ 😐 ☹ features w.r.t. healthy, suspicious, confirmed patients
▣ ☹ anchors
→ pull close ← → push away

Fig. 3: Overview of TrustDetector, in which the disease detection network predicts the risks for multi-diseases and Eye-wise Consistency Learning module enforces consistent prediction and Rank-based Contrastive Learning module enforces the detection network learn ordered representations.

rics to quantify the consistency of disease detection model over different images of the same eye are used and they are *consistency* (Con) and *correct-consistency* ($AccCon$) where $Con = \sum_{m=1}^{M_{multiple}} || \prod_{k \in S(m)} \hat{y}_k^{(t)} ||_1 / M_{multiple}$ and $AccCon = \sum_{m=1}^{M_{multiple}} || \prod_{k \in S(m)} \hat{y}_k^{(t)} \cdot y_k^{(t)} ||_1 / M_{multiple}$ and $M_{multiple}$ is the number of eyes with multiple images, and $S(m)$ is the image set from the eye m, and $|| \cdot ||_1$ is the L$_1$ norm. Their mean over tasks are denoted as $mCon$ and $mAccCon$.

3 Methodology

Overview of TrustDetector Given the training dataset $\{x_n, e_n, y_n\}_{n=1}^{N}$ with N data from M eyes for T diseases, where x_n is the n-th image sample, and $e_n \in [1, M]$ is the index of eye, and $y_n = \{y_n^{(t)}\}_{t=1}^{T}$ and $y_i^{(t)} \in \mathbb{R}^{C^{(t)}}$ is the one-hot encoding for disease t with $C^{(t)}$ states, our TrustDetector aims to learn to make correct and consistent disease predictions for different images with the same eye index. Fig. 3 illustrates the overview. It consists of three components: the disease detection network with SLO images $\{x_1, \cdots, x_B\}$ as input and the predictions for multiple diseases denoted by $\{\hat{y}_1, \cdots, \hat{y}_B\}$ as output for each batch of size B, the eye-wise consistency learning module pulling together the features of images with the same eye index, the rank-based contrastive learning module enforcing the features in embedding space be ordered in line with the disease severity levels.

Disease Detection Network The disease detection network consisting of an encoder Enc and a multi-label classifier Cls is supposed to predict the risks of diseases and grade the disease severity level:

$$\hat{y}_i = Cls(Enc(x_i)) \; . \tag{1}$$

Particularly, we employ the modernized CNN backbone ConvNeXt [12] as the encoder and a linear projector as the classifier. To reduce the mis-classification on minority classes, inspired by [20], we adopt the class-balanced cross-entropy loss, which for each batch can be expressed as

$$\mathcal{L}_{CBCE} = \sum_{i=1}^{B} \sum_{t=1}^{T} \sum_{c=1}^{C^{(t)}} -\frac{1-\omega_c^{(t)}}{C^{(t)}-1} \cdot y_{i,c}^{(t)} \log(\hat{y}_{i,c}^{(t)}) \ , \tag{2}$$

where $\omega_c^{(t)}$ is the sample ratio of class c of disease t.

Eye-wise Consistency Learning To enforce the detection model make consistent predictions and increase the reliability, inspired by the supervised contrastive learning [5], eye-wise consistency learning module is proposed to pull the features of images with same eye index while push away features from different eyes. In detail, given the features in one batch $\{Enc(x_i)\}_{i=1}^{B}$ produced by the backbone where B is the batch size, we project them into a dimension reduced feature space size of 128 via a multiple layer perception (MLP) with one hidden layer, followed by a normalization. The output feature vectors are denoted as $\{u_i\}_{i=1}^{B}$. Then, the module enforces eye-wise consistency by minimizing:

$$\mathcal{L}_{EyeCon} = \sum_{i=1}^{B} \sum_{p \in Eye(i)} -\frac{1}{|Eye(i)|} \log \frac{\exp\left(u_i \cdot u_p/\tau\right)}{\sum_{k \in \bar{A}(i)} \exp\left(u_i \cdot u_k/\tau\right)} \ , \tag{3}$$

where i is the anchor sample from the eye e_i, $\bar{A}(i) = \{1, \cdots, B\} \setminus i$ is the set of indices excluding the anchor sample, and $Eye(i) = \{p \in \bar{A}(i) : e_p = e_i\}$ is the set of image indices with the same eye index with the anchor sample, and $\tau = 0.07$ is the temperature hyper-parameter.

Rank-based Contrastive Learning As the disease severity levels are ordinal, learning an ordered representation where distances of features are ordered in line with distances in the label space is desired. Taking glaucoma grading as an example, the increasing order of severity levels is non-glaucoma, suspicious and glaucoma. What we desire is that, in the ordered embedding space, the feature distances between samples of non-glaucoma and glaucoma are larger than that of non-glaucoma and suspicious. To this end, we draw inspiration from Rank-N-Contrast [21] and propose the rank-based contrastive learning module. For each batch, we first employ a MLP with one hidden layer to project the features produced by backbone network to dimension of 128 followed by a normalization. The normalized features are denoted as $\{v_i\}_{i=1}^{B}$. Then, taking sample i as the anchor sample, for any other sample j in the batch, we contrast them against each other, enforcing the feature distance between i and j to be less than that of other samples if their differences of severity levels are greater than that of i and j. Enumerating over all samples in one batch as anchors, all features are enforced to be ordered in line with their orders in label space via minimizing:

$$\mathcal{L}_{Rank} = \sum_{i=1}^{B}\sum_{t=1}^{T}\sum_{j\in\bar{A}(i)} -\frac{1}{B(B-1)} \log \frac{\exp(v_i{\cdot}v_j/\eta)}{\sum_{k\in Rank_{i,j}^{(t)}} \exp(v_i{\cdot}v_k/\eta)} , \tag{4}$$

where $Rank_{i,j}^{(t)} = \{k \in \bar{A}(i) : \Delta_{i,k}^{(t)} \geq \Delta_{i,j}^{(t)}\}$ is the set of samples satisfying $\Delta_{i,k}^{(t)} \geq \Delta_{i,j}^{(t)}$. $\Delta_{i,j}^{(t)}$ represents the severity level difference between sample i and j regarding disease t and $\eta = 0.07$ is the temperature hyper-parameter.

Total Loss The total loss of our TrustDetector is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{CBCE} + \alpha \cdot \mathcal{L}_{EyeCon} + \beta \cdot \mathcal{L}_{Rank} , \tag{5}$$

where α and β are hyper-parameters.

4 Experiments

Experimental Setup We implement our TrustDetector on MMPreTrain platform [1]. We use the pre-trained model on ImageNet-21K to initialize the parameters in the backbone of ConvNeXt V1 [12] and Gaussian distribution with zeros mean and standard deviation of 0.01 to initialize the parameters associated with the multi-label classifier in Fig. 3, and uniform distribution to initialize parameters in MLP_1 and MLP_2. We employ an AdamW [6] optimizer and train the model 150 epochs using a linear warm-up of 20 epochs and a cosine decay learning rate scheduler afterward. The loss weights α and β are set to 0.2 and 0.1 respectively. Other hyper-parameters include: initial learning rate of 5×10^{-5}, weight decay of 0.05, and a batch size of 32.

Pre-processing and Data Augmentation We scale images such that the short side is 512. For the data augmentation, we keep the long size be 640 via random crop/zero padding for images with long size greater/less than 640. Besides, random rotation ranging from $-30°$ to $30°$, random flipping (horizontal and vertical) and brightness enhancement (0-0.9) are also used.

Comparisons with State-of-the-arts We compare TrustDetector with two prevalent CNN classification methods, i.e., ResNet-50 [3] and SENet [4], and one transformer-based classification method Swin [11] and two latest CNN-based methods, i.e., ConvNeXt V1 [12] and ConvNeXt V2 [18]. The backbone types of Swin [11], ConvNeXt V1 [12] and ConvNeXt V2 [18] are the tiny one. Performances on the test set are reported in Tab. 2. Our TrustDetector achieves $mF1$ of 58.96% which surpasses the second best Swin [11] by 1.41%. In terms of $mKappa$ and $mAcc$, our TrustDetector achieves 53.67% and 93.42% which surpass the second best ConvNeXt V1 [12] by 2.78% and 0.24% respectively. In terms of the consistency evaluation metrics, our TrustDetector achieves the second best in $mCon$ and the best in $mAccCon$. The performances for each task can be found in the supplementary.

Table 2: Multi-disease detection performance comparisons on *test* set. The averages and standard deviations from 5-trails are reported.

Methods	$mF1$	$mKappa$	$mAcc$	$mCon$	$mAccCon$
ResNet-50 [3]	$56.14_{\pm 1.26}$	$49.49_{\pm 2.02}$	$92.62_{\pm 0.49}$	$93.77_{\pm 1.00}$	$88.11_{\pm 0.96}$
SENet [4]	$56.96_{\pm 2.61}$	$50.48_{\pm 2.67}$	$92.85_{\pm 0.50}$	$94.13_{\pm 1.02}$	$89.05_{\pm 1.33}$
Swin [11]	$\underline{57.55}_{\pm 1.29}$	$49.47_{\pm 1.54}$	$91.69_{\pm 0.38}$	$94.12_{\pm 0.46}$	$87.76_{\pm 0.45}$
ConvNeXt V1 [12]	$56.47_{\pm 1.26}$	$\underline{50.89}_{\pm 1.96}$	$\underline{93.18}_{\pm 0.41}$	$94.35_{\pm 0.57}$	$\underline{89.69}_{\pm 0.81}$
ConvNeXt V2 [18]	$52.04_{\pm 1.99}$	$47.11_{\pm 2.41}$	$93.02_{\pm 0.13}$	$\mathbf{95.17}_{\pm 0.90}$	$89.67_{\pm 0.58}$
TrustDetector (ours)	$\mathbf{58.96}_{\pm 0.89}$	$\mathbf{53.67}_{\pm 0.88}$	$\mathbf{93.42}_{\pm 0.30}$	$\underline{95.07}_{\pm 0.47}$	$\mathbf{90.58}_{\pm 0.47}$

Table 3: Influences of the proposed learning modules. The averages and standard deviations over 5-trials on the *test* set of Retina-SLO are reported.

	$mF1$	$mKappa$	$mAcc$	$mCon$	$mAccCon$
TrustDetector	$\mathbf{58.96}_{\pm 0.89}$	$\mathbf{53.67}_{\pm 0.88}$	$\underline{93.42}_{\pm 0.30}$	$\underline{95.07}_{\pm 0.47}$	$\mathbf{90.58}_{\pm 0.47}$
w/o $\mathcal{L}_{EyeCon}$	$56.34_{\pm 1.04}$	$51.22_{\pm 1.61}$	$93.31_{\pm 0.24}$	$94.78_{\pm 0.65}$	$89.83_{\pm 0.48}$
w/o $\mathcal{L}_{Rank}$	$\underline{57.78}_{\pm 1.17}$	$\underline{51.99}_{\pm 1.23}$	$\mathbf{93.61}_{\pm 0.19}$	$\mathbf{95.32}_{\pm 0.54}$	$\underline{90.45}_{\pm 0.35}$

Table 4: Influences of different settings of α and β in Eq.(5) on *validation* set. 5-trails are conducted and the averages and standard deviations are reported.

α	β	$mF1$	$mKappa$	$mAcc$	$mCon$	$mAccCon$
0.2	0.05	$49.31_{\pm 1.37}$	$42.39_{\pm 1.10}$	$93.25_{\pm 0.37}$	$94.38_{\pm 0.60}$	$89.86_{\pm 0.57}$
0.2	**0.1**	$\mathbf{51.77}_{\pm 1.11}$	$\mathbf{44.85}_{\pm 1.61}$	$\mathbf{93.72}_{\pm 0.19}$	$\mathbf{95.12}_{\pm 0.26}$	$\mathbf{90.75}_{\pm 0.27}$
0.2	0.2	$50.53_{\pm 1.61}$	$43.71_{\pm 1.76}$	$93.52_{\pm 0.26}$	$94.85_{\pm 0.75}$	$90.50_{\pm 0.63}$
0.1	0.1	$49.77_{\pm 1.25}$	$44.10_{\pm 1.32}$	$93.60_{\pm 0.15}$	$\mathbf{95.16}_{\pm 0.52}$	$90.52_{\pm 0.27}$
0.2	**0.1**	$\mathbf{51.77}_{\pm 1.11}$	$\mathbf{44.85}_{\pm 1.61}$	$\mathbf{93.72}_{\pm 0.19}$	$95.12_{\pm 0.26}$	$\mathbf{90.75}_{\pm 0.27}$
0.3	0.1	$49.69_{\pm 1.08}$	$42.30_{\pm 1.10}$	$93.39_{\pm 0.23}$	$94.74_{\pm 0.57}$	$90.14_{\pm 0.45}$

How $\mathcal{L}_{EyeCon}$ and $\mathcal{L}_{Rank}$ Contribute? Here we investigate how the proposed eye-wise consistency learning and rank-based contrastive learning contribute to trustworthy disease detection. As shown in Tab. 3, without $\mathcal{L}_{EyeCon}$, $mCon$ decreases to 94.78% from 95.07% and $mAccCon$ decreases to 89.3% from 90.58%, which indicate that the eye-wise consistency learning module contributes to consistent disease prediction over multiple SLO images from the same eye. Without $\mathcal{L}_{Rank}$, the class balanced metrics of $mF1$ and $mKappa$ decrease to 57.78% and 51.99% respectively while the majority class biased metrics of $mAcc$ and $mCon$ increase to 93.61% and 95.32% from 93.42% and 95.07%, respectively. These results indicate that the rank-based contrastive learning module can boost the detection performances on minority classes and reduce the miss identification of samples with diseases.

Influences of Different Settings of α and β Tab. 4 shows the results of different settings for the loss weights α and β in Eq. (5). Overall, the best option for α and β are 0.2 and 0.1 respectively.

5 Conclusions and Future Work

This paper tackles the issue of trustworthy disease detection with SLO images from the perspective of detection system users. First, a large-scale SLO image database is collected and contains 7943 images of 4102 eyes from 2440 subjects and almost half of eyes have at least two SLO images, which enables the research of trustworthy disease detection. Then, *TrustDetector* is proposed for trustworthy disease detection, in which eye-wise consistency module enforces the encoder learn eye-wise consistent features. Finally, extensive experiments are conducted and results clearly demonstrate the superiority of our *TrustDetector*. In summery, our work opens up new possibilities for investigating the trustworthy detection of multiple retinal diseases with SLO images. In future, other attack types such as adversarial attack will be considered to increase the trustability of disease detection models.

Acknowledgments. This work was supported in part by Academy Research Fellow project under Grant 355095, the National Natural Science Foundation of China under Grant NSFC No. 62006249, Hunan Provincial Natural Science Foundation of China under Grant 2021JJ40788. As well, the authors wish to acknowledge the High Performance Computing Centre of Central South University, China and CSC–IT Center for Science, Finland, for computational resources.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Contributors, M.: Openmmlab's pre-training toolbox and benchmark. https://github.com/open-mmlab/mmpretrain (2023)
2. Haleem, M.S., Han, L., Hemert, J.v., Fleming, A., Pasquale, L.R., Silva, P.S., Song, B.J., Aiello, L.P.: Regional image features model for automatic classification between normal and glaucoma in fundus and scanning laser ophthalmoscopy (slo) images. Journal of medical systems **40**, 1–19 (2016)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
4. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
5. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
7. Li, L., Xu, M., Liu, H., Li, Y., Wang, X., Jiang, L., Wang, Z., Fan, X., Wang, N.: A large-scale database and a cnn model for attention-based glaucoma detection. IEEE Transactions on Medical Imaging **39**(2), 413–424 (2020)

8. 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. pp. 10571–10580 (2019)
9. Li, T., Bo, W., Hu, C., Kang, H., Liu, H., Wang, K., Fu, H.: Applications of deep learning in fundus images: A review. Medical Image Analysis **69**, 101971 (2021)
10. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences **501**, 511–522 (2019)
11. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
12. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
13. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia medica **22**(3), 276–282 (2012)
14. Opitz, J., Burst, S.: Macro f1 and macro f1. arXiv preprint arXiv:1911.03347 (2019)
15. Organization, W.H., et al.: World report on vision (2019)
16. Tang, F., Luenam, P., Ran, A.R., Quadeer, A.A., Raman, R., Sen, P., Khan, R., Giridhar, A., Haridas, S., Iglicki, M., et al.: Detection of diabetic retinopathy from ultra-widefield scanning laser ophthalmoscope images: a multicenter deep learning analysis. Ophthalmology Retina **5**(11), 1097–1106 (2021)
17. Wang, L., Ghosh, D., Gonzalez Diaz, M., Farahat, A., Alam, M., Gupta, C., Chen, J., Marathe, M.: Wisdom of the ensemble: Improving consistency of deep learning models. Advances in Neural Information Processing Systems **33**, 19750–19761 (2020)
18. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16133–16142 (2023)
19. Wu, J., Fang, H., Li, F., Fu, H., Lin, F., Li, J., Huang, Y., Yu, Q., Song, S., Xu, X., et al.: Gamma challenge: glaucoma grading from multi-modality images. Medical Image Analysis **90**, 102938 (2023)
20. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE international conference on computer vision. pp. 1395–1403 (2015)
21. Zha, K., Cao, P., Son, J., Yang, Y., Katabi, D.: Rank-n-contrast: Learning continuous representations for regression. Advances in Neural Information Processing Systems **36** (2024)
22. Zhang, J., Dashtbozorg, B., Bekkers, E., Pluim, J.P.W., Duits, R., ter Haar Romeny, B.M.: Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores. IEEE Transactions on Medical Imaging **35**(12), 2631–2644 (2016)