REFUGE2 CHALLENGE: TREASURE FOR MULTI-DOMAIN LEARNING IN GLAUCOMA ASSESSMENT

A PREPRINT

Huihui Fang, Fei Li, Huazhu Fu, Xu Sun, Xingxing Cao, Jaemin Son, Shuang Yu, Menglu Zhang,
Chenglang Yuan, Cheng Bian, Baiying Lei, Benjian Zhao, Xinxing Xu, Shaohua Li, Francisco Fumero, Jose Sigut, Haidar Almubarak, Yakoub Bazi, Yuanhao Guo, Yating Zhou, Ujjwal Baid, Shubham Innani, Tianjiao Guo, Jie Yang, José Ignacio Orlando, Hrvoje Bogunović, Xiulan Zhang, Yanwu Xu*

ABSTRACT

Glaucoma is the second leading cause of blindness and is the leading cause of irreversible blindness disease in the world. Early screening for glaucoma in the population is significant. Color fundus photography is the most cost effective imaging modality to screen for ocular diseases. Deep learning network is often used in color fundus image analysis due to its powful feature extraction capability. However, the model training of deep learning method needs a large amount of data, and the distribution of data should be abundant for the robustness of model performance. To promote the research of deep learning in color fundus photography and help researchers further explore the clinical application signification of AI technology, we held a REFUGE2 challenge. This challenge released 2,000 color fundus images of four models, including Zeiss, Canon, Kowa and Topcon, which can validate the stabilization and generalization of algorithms on multi-domain. Moreover, three sub-tasks were designed in the challenge, including glaucoma classification, cup/optic disc segmentation, and macular fovea localization. These sub-tasks technically cover the three main problems of computer vision and clinicly cover the main researchs of glaucoma diagnosis. Over 1,300 international competitors joined the REFUGE2 challenge, 134 teams submitted more than 3,000 valid preliminary results, and 22 teams reached the final. This article summarizes the methods of some of the finalists and analyzes their results. In particular, we observed that the teams using domain adaptation strategies

*H. Fang and F. Li contributed equally to this work.
X. Zhang and Y. Xu are the corresponding authors (E-mail: zhangxl2@mail.sysu.edu.cn; xuyanwu@baidu.com).
H. Fang, F. Li, H. Fu, X. Sun, X. Cao, X. Zhang, and Y. Xu co-organized the ADAM challenge.
F. Li, and X. Zhang are with State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science,Guangzhou, China.
H. Fang, X. Sun, X. Cao, and Y. Xu are with Intelligent Healthcare Unit, Baidu Inc., Beijing, China.
H. Fu is with the Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore.
J.I. Orlando is with CONICET, Yatiris lab of Pladema Institute, Tandil, Argentina.
H. Bogunović is with Department of Ophthalmology and Optometry, Medical University of Vienna, Vienna, Austria.
J. Son is with VUNO Inc. Seoul, Republic of Korea.
S. Yu is with Tencent HealthCare, Tencent, Shenzhen, China.
M. Zhang is with Computer Vision Institute, College of Computer Science and Software Engineering of Shenzhen University, Shenzhen, China.
C. Yuan is with School of Biomedical Engineering, Health Science Center, Shenzhen University, China.
C. Bian is with Tencent Jarvis Lab.
B. Lei are with Shenzhen University, Shenzhen, China.
B. Zhao is with College of Computer Science and Software Engineering, Shenzhen University, China.
X. Xu and S. Li are with Institute of High Performance Computing, The Agency for Science, Technology and Research (A*STAR), Singapore.
F. Fumero and J. Sigut are with Department of Computer Science and Systems Engineering, Universidad de La Laguna.
H. Almubarak and Y. Bazi are with King Saud University.
Y. Guo and Y. Zhou are with Institute of Automation, Chinese Academy of Sciences.
U. Baid and S. Innani are with SGGS Institute of Engineering and technology, Nanded, India.
T. Guo is with Institute of Medical Robotics, Shanghai Jiao Tong University.
J. Yang is with Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University.

had high and robust performance on the dataset with multi-domain. This indicates that UDA and other multi-domain related researches will be the trend of deep learning field in the future, and our REFUGE2 datasets will play an important role in these researches.

Keywords Glaucoma · Color fundus photography · multi-model dataset · Domain adaptation · REFUGE2 Challenge

1 Introduction

Glaucoma is a neurodegenerative illness defined by characteristic damage to the optic nerve and accompanying visual field deficits. Early diagnosis and treatment are critical to prevent irreversible vision loss and ultimate blindness. However, the early symptoms of glaucoma are relatively difficult to recognize and detect. Clinically, examination for glaucoma includes measurement of intraocular pressure using a tonometer, observation of the optic nerve head using an ophthalmoscope or optical coherence tomography, examination of drainage of intraocular fluid using an anterior chamber angioscope, and visual field measurement. However, for screening glaucoma in the population, these examinations are so complex. In contrast to the above examinations, color fundus photography (CFP) is the most cost-effective imaging method for screening glaucoma (Han et al. [2021]). The collection of CFP is convenient, and CFP can provide the information of fundus structures such as optic disc (OD) and optic cup (OC) where are greatly affected by glaucoma. Clinical evaluation of cup-disc ratio (CDR) is an indicator for the diagnosis of glaucoma. Specifically, a vertical CDR value larger than 0.7 is considered at risk for glaucoma, and the larger the value, the higher the risk (Aung and Crowston [2016]). In addition to CDR, other features which can be observed from the CFPs, such as serious retinal nerve fiber layer defect and narrowing of disc rim, can be used as the basis for screening glaucoma. Meanwhile, another important ocular structure, fovea, can also be observed from the CFPs. Fovea is the most sensitive part of human vision, and it is related to the position of optic disc (Niemeijer et al. [2009]). Detection of the optic cup, optic disc and fovea from CFPs can assist in the diagnosis of ocular diseases.

Recent years, due to the rapid development of deep learning technology(Li et al. [2021a], Shamshad et al. [2022]), many related methods have been developed for CFP automatic analysis. For example, various U-Net variants (Ronneberger et al. [2015], Sevastopolsky [2017], Yu et al. [2019], Wang et al. [2019a], Fu et al. [2018]) have been proposed to realize the optic disc and cup segmentation. There are also studies on fovea localization based on the convolutional neural network framework (Al-Bander et al. [2018], Hasan et al. [2021]). However, due to the small dataset with fovea location annotation, deep learning-based methods have not been widely studied. In addition of the fundus structure extraction, automatic glaucoma detection from the CFPs based on the deep learning methods have also been actively investigated. Li et al. (Li et al. [2018]) trained a Inception-v3 model to achieve the glaucoma classification, where the classification label during training were given by the ophthalmologist according to the signs of glaucomatous variation in the optic nerve and other parts. Due the optic disc is heavily affected by glaucoma, many researchers (Bajwa et al. [2019], Jiang et al. [2019], Fu et al. [2018]) used the local information of optic disc region to further identify glaucoma. Since deep learning based methods require a large amount of annotated data to train the model, while there are small annotated CFPs, Ruben et al. (Hemelings et al. [2020]) proposed to use active learning method to train the ResNet-50 model. Active learning based method (Felder and Brent [2009]) queries the most useful unlabeled samples through certain algorithms, and the samples are tagged by experts, and then are added to training dataset to train the classification model to improve the accuracy of the model.

Due to the discrepancy of different devices for fundus image collection, a well-trained neural network is usually unsuitable for another new dataset. To solve this problem, more and more studies (Chen and Wang [2021], Lei et al. [2021], Kadambi et al. [2020], Liu et al. [2019], Wang et al. [2019b], Chen et al. [2021]) focus on the unsupervised domain adaptation (UDA). For example, Liu et al. (Liu et al. [2019]) provided Collaborative Feature Ensembling Adaptation (CFEA) to effectively overcome domain shift. Wang et al. (Wang et al. [2019b]) presented Boundary and Entropy-driven Adversarial Learning (BEAL), which utilized the adversarial learning to encourage the boundary prediction and mask probability entropy map of the target domain to be similar to the source ones, generating more accurate boundaries and suppressing the high uncertainty predictions of OD and OC segmentation. Chen et al. (Chen et al. [2021]) presented a novel denoised pseudo-labeling method for the source-free unsupervised domain adaptation problem. Among these studies for multi-domain training and validation, Liu et al. (Liu et al. [2019]) used REFUGE1 dataset, Wang et al. (Wang et al. [2019b]) used Drishti-GS and RIM-ONE_r3, and Chen et al. (Chen et al. [2021]) used REFUGE1, Drishti-GS and RIM-ONE_r3. These data satisfy the requirements of collecting by different devices, but belong to different datasets. Different data annotation standards will affect the training and evaluation of the models.

Tabel 1 summarizes all the databases used for training or testing processes of the papers described above. As can be seen from the first 15 rows (except REFUGE1 and REFUGE2) of the table, the existing databases have three deficiencies. First, each dataset contains a small number of samples, so researchers need to use different datasets for experiments, but this will introduce the problem of different labeling standards. Second, there is no database that provides glaucoma

Table 1: Summary of the machine model used and the labels annotated in different databases. GC-Glaucoma.

Database	Cameras	Num. of images			Ground truth labels		
		GC	Non GC	Total	GC classification	optic disc/cup	fovea
ARIA (Zheng et al. [2012])	Zeiss	0	143	143	×	✓/×	✓
DIARETDB0 (Kauppi et al.)	-	-	-	130	×	✓/×	✓
DIARETDB1 (Kauppi et al. [2007])	Zeiss	-	-	89	-	✓/×	✓
DRIONS-DB (Carmona et al. [2008])	-	-	-	110	×	✓/×	×
DRISHTI-GS (Sivaswamy et al. [2014])	-	70	31	101	✓	✓/✓	×
DRIVE (Staal et al. [2004])	Canon	-	-	40	-	-	-
HEI-MED (Giancardo et al. [2012])	Zeiss	-	-	169	-	-	-
HRF (Budai et al. [2013])	Canon	15	30	45	✓	×/×	×
IDRiD (Porwal et al. [2018])	Kowa	0	516	516	×	✓/×	✓
MESSIDOR (Decencière et al. [2014])	Topcon	-	-	1200	-	✓/×	×
ORIGA (Zhang et al. [2010])	Canon	168	482	650	✓	✓/✓	×
RIGA (Almazroa et al. [2018])	Topcon, Canon	-	-	750	×	✓/✓	×
RIM-ONE (Fumero et al. [2011])	Canon	74	85	169	✓	✓/×	×
SCES (Baskaran et al. [2015])	Canon	46	1630	1676	✓ (screening)	×/×	×
STARE (Goldbaum [2013])	Topcon	-	-	81	×	✓/×	✓
REFUGE1 (Orlando et al. [2020a])	Zeiss, Canon	120	1080	1200	✓	✓/✓	✓
REFUGE2	**Canon, Zeiss, Topcon, Kowa**	**280**	**1720**	**2000**	✓	✓/✓	✓

category, optic cup and disc segmentation mask and fovea localization labels at the same time. Some researchers made their own labels of the used datasets, such as GeethaRamani et al. (GeethaRamani and Balasubramanian [2018]) labeled with macular fovea location for DRIVE and HEI-Med. Third, the existing database acquisition machine models are relatively single which makes it difficult to verify the stability and the generalization of the algorithm on multi-domain data.

To address these issues, we hosted the REFUGE challenges at MICCAI. In response to the small number, single models and different labeling standards of the samples provided by the existing database, we released 1,200 fundus images from two equipment models (Canon and Zeiss) in REFUGE1 challenge in 2018. Then, in 2020, we released another 800 fundus images at the REFUGE2 Competition from two other equipment models (Kowa and Topcon). The purpose of the REFUGE challenge is to provide researchers with multi-model CFPs and verify the generalization ability of the algorithms. The REFUGE2 challenge considers the importance of the fundus states to the glaucoma diagnosis, and designs three sub-tasks, including glaucoma classification, cup/optic disc segmentation, and macular fovea localization (as shown in Fig. 1). At the same time, it provides an online evaluation platform for these three tasks. The REFUGE2 challenge attracted dozens of internationally renowned universities and institutions and over 1300 international participants. 134 teams submitted over 3000 valid preliminary results. In the end, 22 teams made it to the final round, of which 40% were enterprises. In the end, South Korean company VUNO won the title, and Shenzhen University and Tencent United won the second and third places. In this article, we summarize and analyze the methods of some of the participating teams, compare their results and discuss the clinical significance of these artificial intelligence (AI) methods. To encourage further developments and to ensure a proper and fair comparison of

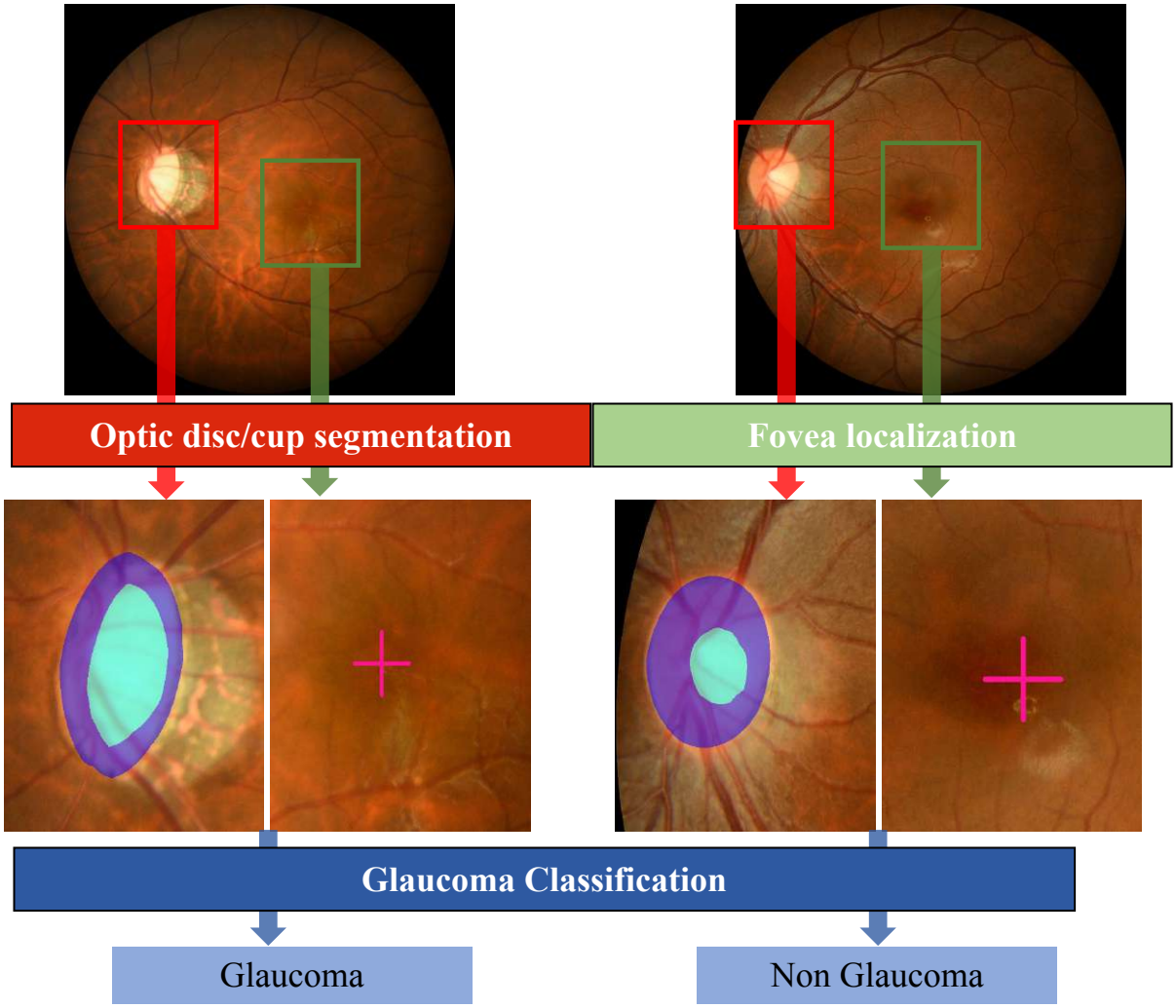


Figure 1: REFUGE2 challenge tasks: Classification of clinical glaucoma, segmentation of optic disc and cup, and localization of fovea from color fundus photographs.

new methods, the REFUGE2 data and the evaluation platform remain open through the grand challenge website at `https://refuge.grand-challenge.org/`. The main contributions of this work are as follows:

- Introduce the REFUGE2 challenge, and describe the details of the 2000 CFPs as well as the labeling process. To the best of our knowledge, REFUGE2 dataset was the first public dataset of CFPs collected from four equipment models, and simultaneously give the ground truth in regard to the ocular structures and glaucoma diagnosis.
- Summary the 10 methods of the finalists and compare their results on three subtasks, and find that domain shift strategy can improve the model performances.
- The significance of using different datasets, prior knowledge and other strategies in AI-based methods to the clinical glaucoma diagnosis was discussed.

(A) Zeiss (B) Canon (C) KOWA (D) TOPCON

Figure 2: Samples collected from the four machine models. First row: glaucoma samples, second row: non glaucoma samples.

Table 2: Summary of the main characteristics of each subset of the REFUGE2 dataset

Characteristics	Subset			
	Training		Online	Onsite
Acquisition device	Zeiss Visucam 500	Canon CR-2	KOWA	TOPCON TRC-NW400
Resolution	2124×2056	1634×1634	1940×1940	1848×1848
Num. images	400	800	400	400
Glaucoma/Non glaucoma	40/360	80/720	80/320	80/320
Public labels	✓	✓	×	×

2 The REFUGE2 challenge

2.1 REFUGE2 Database

The REFUGE2 database consists of 2000 retinal CFPs provided by Zhongshan Ophthalmic Center, stored in jpg format with 8 bits per color channel. The examinations were performed in a standardized darkroom, and the CFPs were centered on the following situations: on the optic disc region, on the macular area, and on the midpoint between optic disc and macula with both optic disc and macula visible, which greatly conforms to the clinical shooting situations. Both left and right eyes of each patient were included if the images were eligible. Where, 1200 CFPs of the REFUGE2 database is from REFUGE1 database acquired by a Zeiss Visucam 500 camera with a resolution of 2124×2056 pixels (400 images) and a Canon CR-2 device with a resolution of 1634×1634 pixels (800 images). These 1200 CFPs are released as training set in REFUGE2 challenge (as shown in Table 2). 400 of the remaining 800 CFPs acquired by a KOWA device with a resolution of 1940×1940 pixels are released as online set, and the other 400 CFPs acquired by a Topcon TRC-NW400 device with a resolution of 1848×1848 are released as onsite set. Fig. 2 shows the glaucoma and non glaucoma samples collected from different machine models. Fig. 3 shows the data distribution of our dataset, which was visualized by t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton [2008]). The 800 Canon training images are distributed differently from the KOWA and Topcon data in the online and onsite sets. The last three rows of Table 3 show the mean and standard deviation of the pixel values of RGB channels in the Zeiss, Canon, KOWA and Topcon datasets. It can also be seen that color deviation were in the data collected by different equiments.

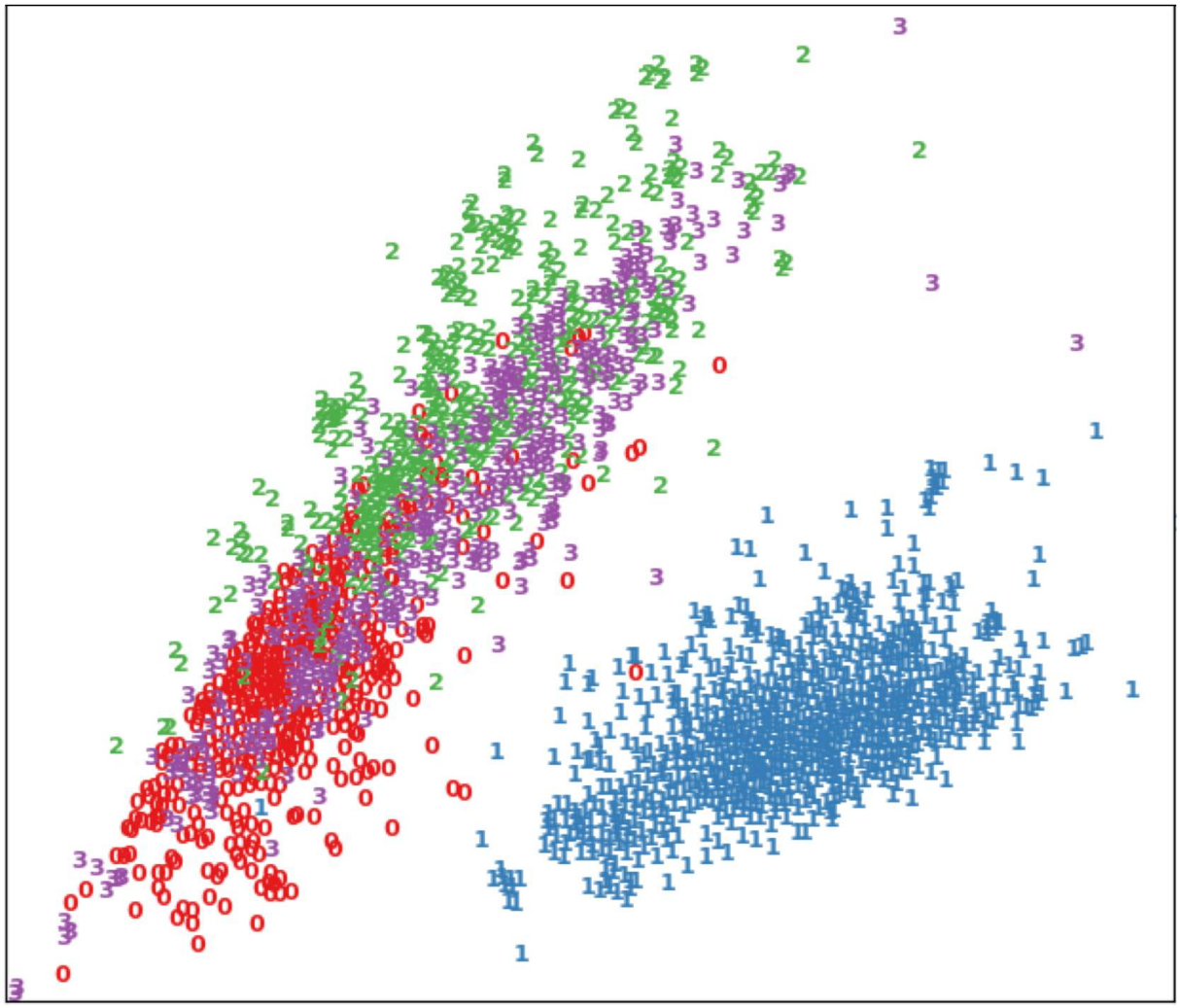


Figure 3: Data distribution of the four machine models collected by t-SNE dimension reduction. 0: Zeiss, 1:Canon, 2:KOWA, 3:Topcon

Each CFP in the REFUGE2 dataset has three reference annotations, including glaucoma/non-glaucoma label, optic disc/cup segmentation mask, and macular fovea localization coordidate. The reference standard for glaucoma presence obtained from the health records, which is not based on fundus image only, but also take OCT, visual field, and other facts into consideration. There are 280 samples correspond to glaucomatous subjects in the REFUGE2 dataset, including 120 samples in the training set, 80 samples in the online set and 80 samples in the onsite set(as shown in Table 2). Manual pixel-wise annotations of the optic disc and cup edges were first created by 7 independent glaucoma specialists from Zhongshan Ophthalmic Center, Sun Yat-Sen University, and then were merged into single annotation by another senior glaucoma specialist. It is stored as a png image with the same size as the corresponding fundus image with the following labels: 128-optic disc exclude the OC region, 0-optic cup, 255-background.Manual pixel-wise annotations of the fovea were also obtained by 7 independent glaucoma specialists first. And the final reference standard of the localization coordinate was created by averaging the selected annotations from the 7 annotations. The selected annotations were determined by the senior glaucoma specialist. Table 3 represents the characteristics of OD, OC and fovea areas in each of the machine models. The fovea area is a square area with the fovea as the center and 2 times optic disc diameter as the side length. From the first two rows of the table, we can observe that no matter which machine model the sample was collected with, the proportion of optic cup and disc in the whole image is very small, which determined by the physiological structure. Meanwhile, the size of optic cup varies greatly among different samples, mainly because glaucoma will lead to the expand of optic cup. In terms of the OD and OC proportion of different machine models, the optic disc proportions of Zeiss and Canon were about 0.1% larger than those of KOWA and

Table 3: REFUGE2 dataset characteristics in each of the machine models (Zeiss, Canon, KOWA, and Topcon), including the OD and OC areas as percentages of the field-of-view area, and the pixel values of OD, OC and fovea areas.

		Zeiss	Canon	KOWA	Topcon
OD area as % of FOV		1.696±0.304	1.624±0.322	1.520±0.314	1.547±0.370
OC area as % of FOV		0.448±0.203	0.426±0.239	0.363±0.218	0.430±0.240
OD area pixel value	channel R	196.399±18.455	237.785±11.863	239.403±14.010	209.085±36.238
	channel G	105.447±17.481	173.798±19.720	144.824±21.768	122.608±32.951
	channel B	52.956±18.033	138.434±21.169	82.476±24.894	75.414±26.628
OC area pixel value	channel R	227.382±15.268	249.730±7.477	250.344±8.157	223.688±34.543
	channel G	142.929±28.423	205.175±24.006	172.916±26.253	142.831±38.802
	channel B	74.032±28.632	172.737±29.555	108.917±30.834	93.407±35.034
Fovea area pixel value	channel R	99.295±23.831	154.640±24.102	166.924±30.833	118.867±44.693
	channel G	53.174±13.375	100.196±19.248	83.465±19.891	56.780±22.625
	channel B	22.842±9.073	73.765±15.291	22.010±15.271	30.701±11.104
image pixel value	channel R	71.858±48.228	103.683±72.821	106.428±69.721	94.921±62.264
	channel G	42.796±28.058	68.977±51.654	53.048±38.663	48.590±35.111
	channel B	21.376±15.695	55.696±41.228	14.920±20.079	29.139±22.089

Topcon, and KOWA has the smallest mean value of OD proportion. In terms of pixel values in RGB channels, images collected by Canon have larger pixel values than those collected by other machine models, while that collected by KOWA have larger pixel values in R channel and smaller pixel values in B channel. It can be seen that there are color deviations in images collected by different machine models.

2.2 Challenge Evaluation

Task1: Glaucoma Classification

Classification results will be compared to the clinical diagnosis of glaucoma. Receiver operating curve will be created across all the testing images and an area under the curve (AUC) will be calculated. Each team receives a rank (1=best) based on the obtained AUC value. This ranking forms the classification leaderboard.

Task 2: Optic disc/cup Segmentation

Submitted segmentation results will be compared to the reference standard. The Dice indices (DI) (Dice [1945]) for OD/OC separately, and the mean absolute error (MAE) (Willmott and Matsuura [2005]) of the vertical cup-to-disc ratio (vCDR) will be calculated as segmentation evaluation measures. Each team receives a rank (1=best) for each evaluation measure based on the mean value of the measure over the testing images. The segmentation score is then determined by adding the three individual ranks (OD DI, OC DI, and vCDR MAE). The team with the lowest score will be ranked $\#1$ on the segmentation leaderboard.

Task 3: Fovea Localization

Submitted fovea localization results will be compared to the reference standard. The evaluation criterion is the Average Euclidean Distance (AED) between the estimations and ground truth, which is the lower the better. Each team receives a rank (1=best) based on the obtained measure. This ranking forms the fovea localization leaderboard.

The final score of the challenge was calculated by the following equation:

$$S = 0.45 \times R_{cls} + 0.45 \times R_{seg} + 0.1 \times R_{loc}, \tag{1}$$

where R_{cls}, R_{seg} and R_{loc} represent the ranks of the aforementioned three leaderboards. This score then determines the online or onsite ranking (1=best) of the challenge. In case of a tie, the rank of the classification leaderboard has the preference.

After the online evaluation, 22 teams attended the final onsite challenge during MICCAI 2020. The onsite images were released for these teams and the final results were asked to be submitted. Both online and onsite evaluation rankings contributed the final ranking S_{final}:

$$S_{final} = 0.3 \times S_{online} + 0.7 \times S_{onsite}, \tag{2}$$

where a higher weight was assigned to the onsite evaluation ranking due that the results of onsite images can better reflect the generalization ability of the methods proposed by the participating teams, as same as other challenges [Orlando et al., 2020a, Fu et al., 2020]. The REFUGE2 challenge requires that the same model be used for both online and onsite sets. If there is any change in the model, the score on the onsite set will be considered invalid.

Table 4: Data processing and data augmentation overview of the participating teams.

Team	Glaucoma classification	Optic disc/cup segmentation	Fovea localization
VUNO	Rotation, flip, affine transformation, image perturbation (color, contrast, brightness, sharpness, RGB shift, Gamma), noise (blur, ISO noise, Gaussian noise, JPEG compression, sunflare), resize, elastic transform, and down-sampling.		
CBMIBrand	geometric transformation: horizontal flip, vertical flip, random rotation and scaling; color transformation: noise, blur, sharpening, RGB shift, HUE saturation conversion, CLAHE (Setiawan et al. [2013]), brightness contrast conversion.		
cheeron	scaling, rotation, flipping.		
MIG	-		
TeamTiger	random rotation, brightness, contrast, and saturation.		
MAI	data normalization, random crop, horizontal flip.	data normalization, random flip, random rotation, random scale.	
MIAG ULL	scaling, random rotations, vertical and horizontal flip, zoom.	random geometric and non-genmetric transformations.	
ALISR	random horizontal flip, color jitter (brightness, contrast, saturation, hue), normalized.	-	rotation, vertical flip, horizontal flip.
Pami-G	random shuffle, shift, rotation, resizing, horizontal flip and shearing.		
EyeStar	-	data normalization, randomly resize/crop, randomly flip and rotation, randomly change the brightness, contrast and saturation, converted to 50% grayscale.	data normalization, random shifting, scaling, rotating, random horizontal flipping.

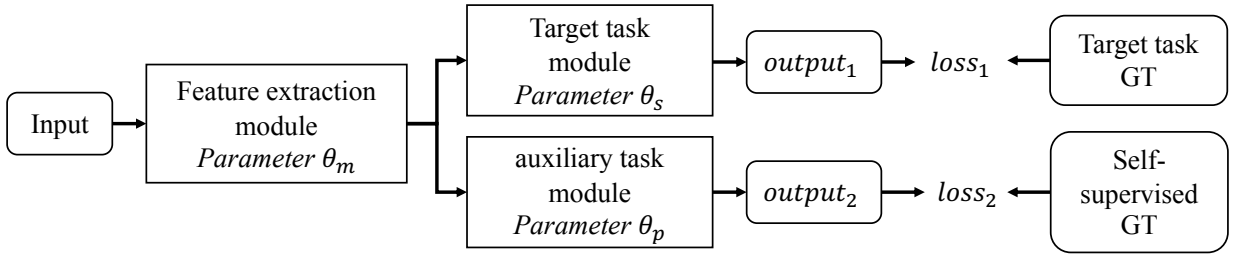


Figure 4: Schematic diagram of TTT strategy.

3 Summary of Challenge Solutions

Three clinically relevant tasks were proposed in the REFUGE2 challenge: classification of clinical glaucoma, segmentation of optic disc and cup, and localization of fovea, as shown in Fig. 1. This section summarizes the data processing technique (Table 4) and the methods (Table 5 - Table 7) used by 10 excellent participating teams* in these three tasks. In addition, we highlight the unsupervised domain adaptation strategy used by some teams for the multi-domain dataset.

3.1 Strategy about Unsupervised Domain Adaptation

The MAI team adopted Test-Time Training (TTT) strategy (Sun et al. [2020]) to ensure thier framework can be well generalized to multiple devices in three tasks. The concept of TTT is to enforce the framework to optimize itself with test data in the inference stage which is served as a plug-to-play strategy for model generalization. The key step for TTT is to construct a self-supervised auxiliary task for the original baseline. In the training stage, the losses of the target task and the auxiliary task simultaneously supervised the model training and optimized the parameters of θ_m, θ_s, and θ_p (as shown in Fig. 4). In the testing stage, the network parameters of target task branch and auxiliary task branch are fixed, and the self-supervised auxiliary task is used to fine-tune the public parameters θ_m of the feature extraction module on

*This paper summarizes the methods and results of VUNO EYE TEAM, cheeron, MAI, MIG, EyeStar, MIAG ULL and ALISR teams in the top 9 in each task of the overall leaderboard (the other 3 teams involved said they would not participate in the review paper). In addition, the methods and results of Pami-G, TeamTiger and CBMIBrand teams with better performance in online semi-final leaderboard are also summarized in this paper.

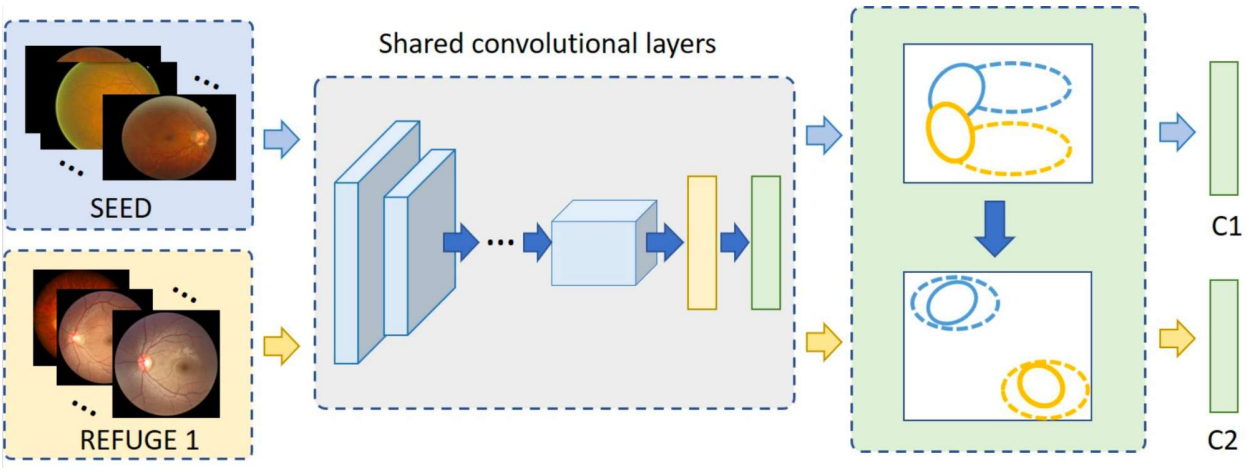


Figure 5: The framework of the EyeStar Team in task 1.

the test set, so as to minimize the $loss_2$. Then, the target task result is obtained by using fine-tuned public parameters θ_m and the target task branch parameters θ_s. In the framework of the MAI team, the auxiliary task is predicting the rotation angle $(0°, 90°, 180°, 270°)$.

In addition to the TTT strategy, the MAI team adopted another classical unsupervised domain adaptation strategy (Tsai et al. [2018]) to maintain the model performance on different devices in the optic disc/cup segmentation task. That is to propose a discriminator D to identify what datasets the predictions are derived from (as shown in stage 2 of the Fig. 8). The purpose of training is to make the discriminator can not distinguish which dataset the prediction comes from, that is, make the prediction of the data in the target domain close to the source domain.

The EyeStar team used a new domain adaptation method to transfer the knowledge from their bigger private dataset, called the Singapore Epidemiology of Eye Disease (SEED) dataset (Guidoboni et al. [2020]), to improve the performance of the REFUGE dataset. Specifically, they treated the SEED dataset as the source domain and the REFUGE2 training dataset as the target domain. They proposed to align distributions of the two domains progressively by utilizing the label information from the two domain. First, they trained the models to predict glaucoma from source and target datasets, and these two classifiers shared the feature extract layers. Then, they aligned the distributions of the samples from the two datasets in the shared feature space. They defined a distribution from the feature space as conditional distribution D and a distribution based on the label space as an ideal distribution P. Then, enforced the conditional distribution D be similar to the ideal distribution P (as shown in Fig. 5).

3.2 Classification of clinical Glaucoma

The aim is to predict the probability of a given color fundus image having glaucoma. Three of 10 participating teams proposed classification methods based on the whole color fundus images. Local information in the optic disc region is critical for glaucoma prediction due to the significant variety of optic disc structure and texture caused by glaucoma, hence, the remaining teams all used the information in the optic disc region.

The VUNO team used an architecture of a commercial system that outputs fifteen ophthalmologic findings (hemorrhage, hard exudate, cotton wool patch, drusen, retinal pigmentary change, vascular abnormality, membrane, fluid accumulation, chorioretinal atrophy, choroidal lesion, myelinated nerve fiber, retinal nerve fiber layer defect, glaucomatous disc change, non-glaucomatous disc change, and macular hole) based on the whole color fundus image, as introduced in Son et al. [2020] with an architectural modification reflecting the tenet of EfficientNet (Tan and Le [2019]). All feature maps of findings are concatenated and a fully connected layer follows with a Sigmoid activation to output the score of glaucoma in the range of 0 to 1. The last fully connected layer is the only part that is trainable. The EyeStar also used the whole color fundus images, while they used the domain adaptation method described in Section *Strategy about Unsupervised Domain Adaptation* to transfer the knowledge from their bigger private dataset to improve the performance of the REFUGE dataset. The TeamTiger team trained Efficientnet family architectures, namely EfficientnetB4, EfficientnetB5, EfficientnetB6, and EfficientnetB7 to conduct the glaucoma classification. Lastly, the results from these four architectures are ensembled by averaging.

Table 5: Methods overview of the 10 participating teams in glaucoma classification task.

Team	Inputs	Archiecture	Additional dataset	Highlight
VUNO	whole images	EfficientNet	Pre-training data	Pre-trained model was trained by additional samples with 15 lesion labels.
CBMIBrand	cropped multi-scale optic disc regions	ResNet18,ResNer34, ResNet50, ResNet101, DenseNet121, DenseNet169	-	Multi-scale optic disc regions were used in the input module
MIG	whole images and cropped optic disc regions	Variants of ResNet50	ORIGA, Drishti-GS1, RIMONE_r3, ACRIMA	
TeamTiger	whole images	EfficientnetB4, EfficientnetB5, EfficientnetB6, EfficientnetB7	-	
cheeron	the cropped optic disc region	Res2NeXt	-	Attention module
MAI		ResNet50	-	Test-Time Training stragegy
MIAG ULL		VGG19	-	
ALISR		CSPNet	-	A cross-stage partial network
Pami-G		The encoder of U-shaped FCN	RIMONE	Transferm learning and polar transformation
EyeStar	whole images	convolutional layers	Singapore Epidemiology of Eye Disease	A novel deep distribution alignment method

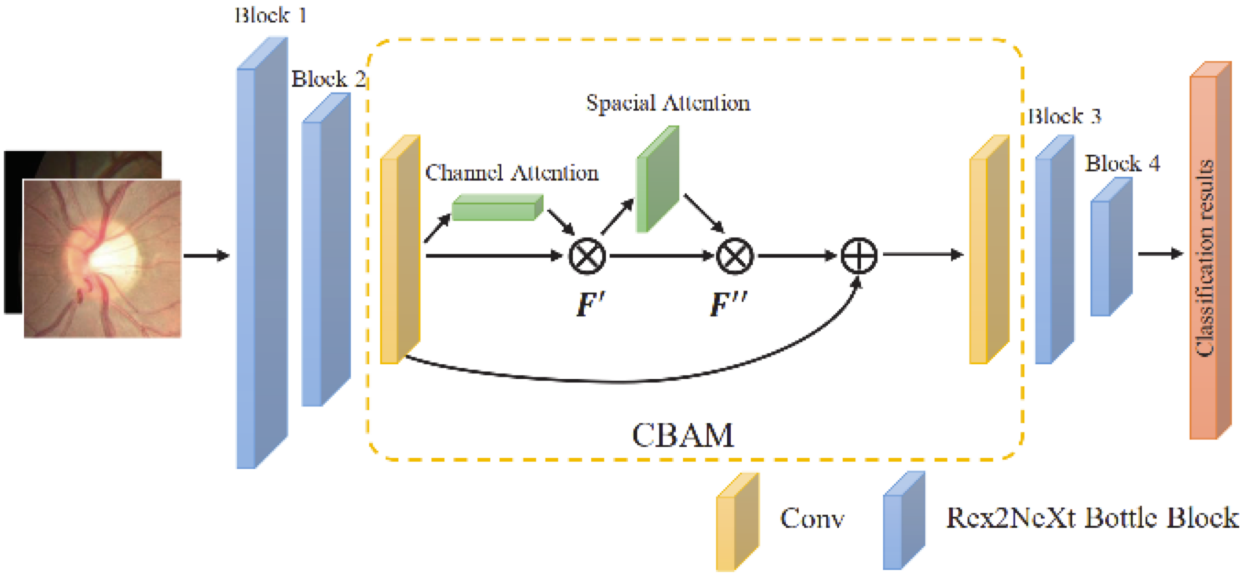


Figure 6: The pipeline of the method of the cheeron team in task 1.

There are six teams predicted the glaucoma based on the local optic disc region only. They all first segmented the optic disc (OD) region coarsely, and cropped the OD patch with designed length, then they adopted different models to predict the glaucoma based on the patch. In particular, the cheeron team cropped the OD patches with 3 disc diameter and adopted Res2NeXt, which is the combination of ResNeXt (Xie et al. [2017]) and Res2Net (Gao et al. [2019]), to predict glaucoma probability. As shown in Fig. 6, to increase the quality and efficacy of the extracted feature maps, they apply attention module on the feed-forward convolution outputs in block 2 to increase the network's attention to the relevant features and suppress the unnecessary features. The architecture sequentially infers attention maps along two separate dimensions, channel and spatial. Afterwards, the attention maps are multiplied to the input feature map

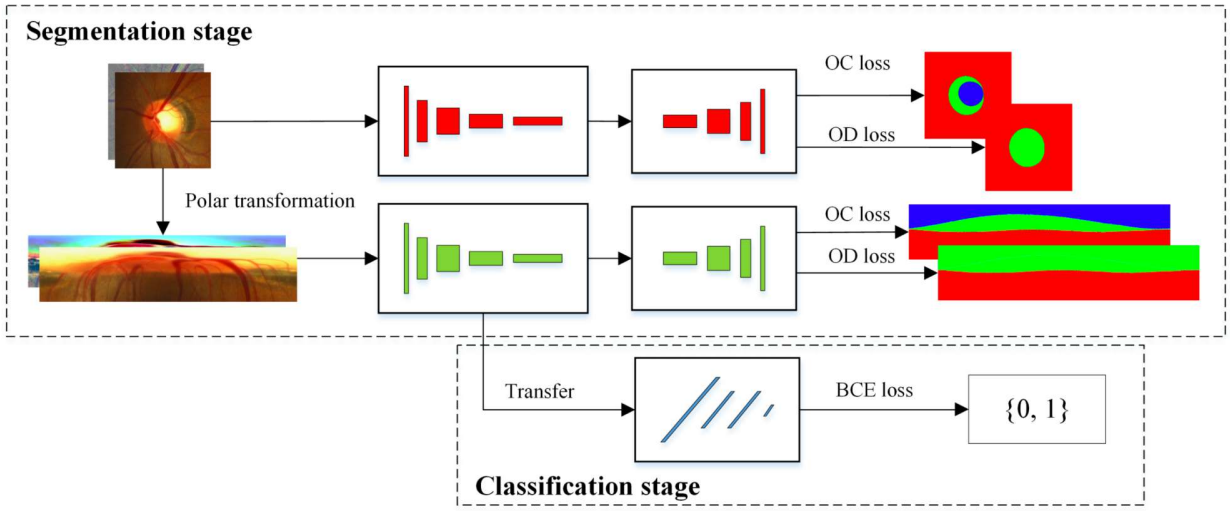


Figure 7: Segmentation and Classification stages procedure of the Pami-G team.

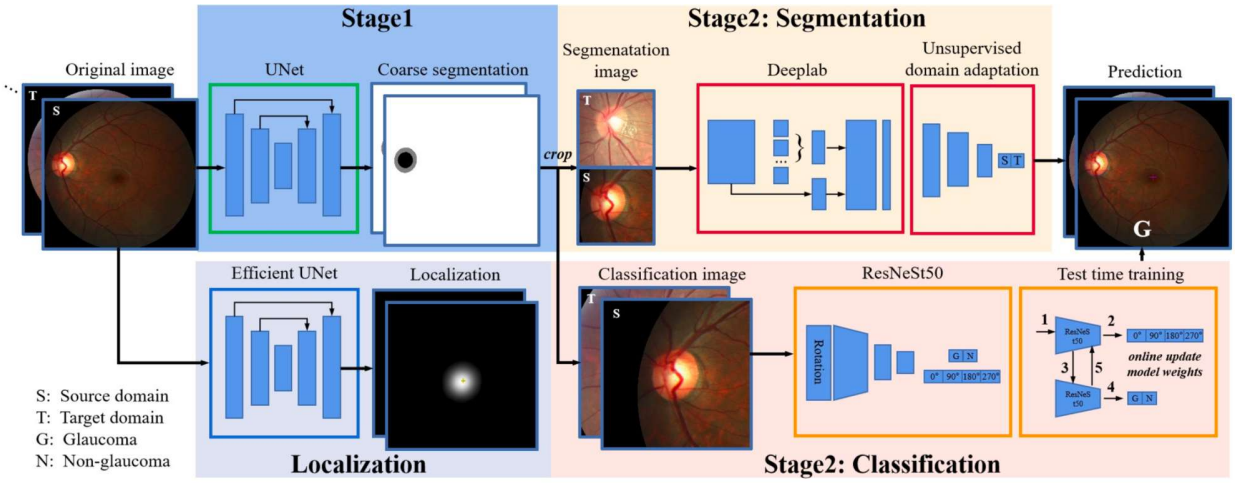


Figure 8: An overview of the proposed framework of the MAI team in the REFUGE2 Challenge.

for adaptive feature refinement. The ALISR team extracted the OD patches by using Mask R-CNN and classified the patches by using cross-stage partial network (CSPNet) (Wang et al. [2020]). The MIAG-ULL team used the method proposed in Sigut et al. [2017] to localize the optic disc and cropped the OD patches, then used VGG19 to predict the glaucoma probability. The Pami-G team introduced the thought of transfer learning in classification task. They changed the structure of the FCN in segmentation stage, removed the layers behind the first upsampling layer, then connected 3 fully connected (FC) layers, and copy the weights of the layers in FCN, as shown in Fig. 7.

The CBMIBrand team first detected the center of OD by using Mask R-CNN, and then cropped multiple regions with various scales enclosing the OD (384x384, 416x416, 448x448, 480x480, 512x512 pixels) to solve the problem of inconsistent object size caused from inconsistent image size from the training and validation set. As for the classification, they devised an ensemble strategy to train the ResNet18, ResNer34, ResNet50, ResNet101, DenseNet121 and DenseNet169 separately and finally average the results of the models which obtain the top performances on the validation set. Fig. 8 shows the overview of the framework proposed by the MAI team. They first acquired the coarse OD segmentation mask with a standard UNet (Ronneberger et al. [2015]). Then, in their classification framework, the ROIs are cropped as the size of 2.5 times that of the disc diameter and resized to 256×256, and ResNet50 (Zhang et al. [2020]) was chosen as the baseline. To increase the generalization capability of the classification network for different devices, they proposed to integrate the TTT strategy (see *Strategy about Unsupervised Domain Adaptation*), which was first proposed by Sun et al. (Sun et al. [2020]).

The remaining teams MIG used both the full fundus images and the cropped optic disc patches. They used five models based on ResNet50 to predict the glaucoma classification, and used ensemble learning to merge their results. Specifically, one model was for the full images, and the other four were for the cropped patches.

3.3 Segmentation of Optic Disc and Cup

Table 6: Methods overview of the 10 participating teams in optic disc and cup segmentation task.

Team	Stategy	Archiecture	Additional dataset	Highlight
ALISR	whole image as input	DeeplabV3 with EfficientNet-B1 as backbone	-	-
VUNO	Whole image and corresponding vessel mask as input	Unet with Efficient-Net B0 as encoder, depth-wise separable convolutions as decoder	RIGA, IDRiD, PALM	using clinical prior knowledge: the position of optic disc is relative to the fundus vessels
CBMIBrand	First detect the OD region, then using mask branch to realize the segmentation	Mask R-CNN	RIGA	using the prior knowledge: position relationship of the optic disc and cup, and using detection model
cheeron	2 stages: first coarse segment the OD, then to realize the fine OD and OC segmentation	UNet with ResNet for OD coarse segmentation; ResUNet for fine segmentation	-	using ASPP and deep supervised
MIG		CENet	ORIGA, Drishti-GS1, RIMONE_r3	texture encoder module
TeamTiger		GAN based network, the generate part using U-shaped network and the EfficientNet as block	-	add GAN loss
MAI		DeeplabV3+	-	using unsupervised domain adaptation strategy
MIAG ULL		ResNet50 PSPNet	-	-
Pami-G		U-shaped FCN	AMD, DRIONS, IDRiD(a), Messidor, PALM, BinRushed, Drishti-GS1, Magrabia, RIM_ONE	Polar transformation
EyeStar		Vision Transformer	Drishti-GS, RIM-ONE-r3	using vision transformer-based method

The aim of this task is to segment the optic disc and the optic cup region from a color fundus image. The framworks proposed by the 10 participanting teams can be divided in two categories: segmenting optic disc and cup directly, and segmenting from coarse to fine by using two-stage strategy.

The VUNO EYE TEAM used a framework to segment the optic disc and the optic cup, respectively. In the framework, they incorporated retinal vessels during training, so the network consisted of two branches as encoders, EfficientNet-B4 and EfficientNet-B0 to respectively deal with the fundus image and the vessel image. The penultimate feature maps of the fundus branch were concatenated to those of the vessel branch. In the decoder module, they up-scaled the feature maps using 1×1 convolutions, depth-wise separable convolutions (Howard et al. [2017]), swish activation functions (Ramachandran et al. [2017]), and depth-wise concatenation and then scaled the feature maps to yield those with the same size of the input. The final segmentation layer is generated using a 1×1 convolution followed by a Sigmoid function. They imposed a loss weight of 0.1 on the vessel branch as the vessel shape can give a good indication of the optic disc region, and the loss weight of the last layer was set to 1. The ALISR team designed a variant of the DeepLab-v3 (Chen et al. [2017]), which used a retrained EfficientNet-B1 (Tan and Le [2019]) to replace the ResNet architecture for the DCNN backbone of the DeepLab-v3. Then the network was used to segment the optic disc and cup directly.

The EyeStar team adopted a transformer method, and to increase efficiency, the transformer takes coarse feature maps from a CNN backbone as input, on which each unit corresponds to a small patch within the original image. As shown in Fig. 9, the output feature maps of the transformer are upsampled with a feature pyramid network (FPN) (Lin et al.

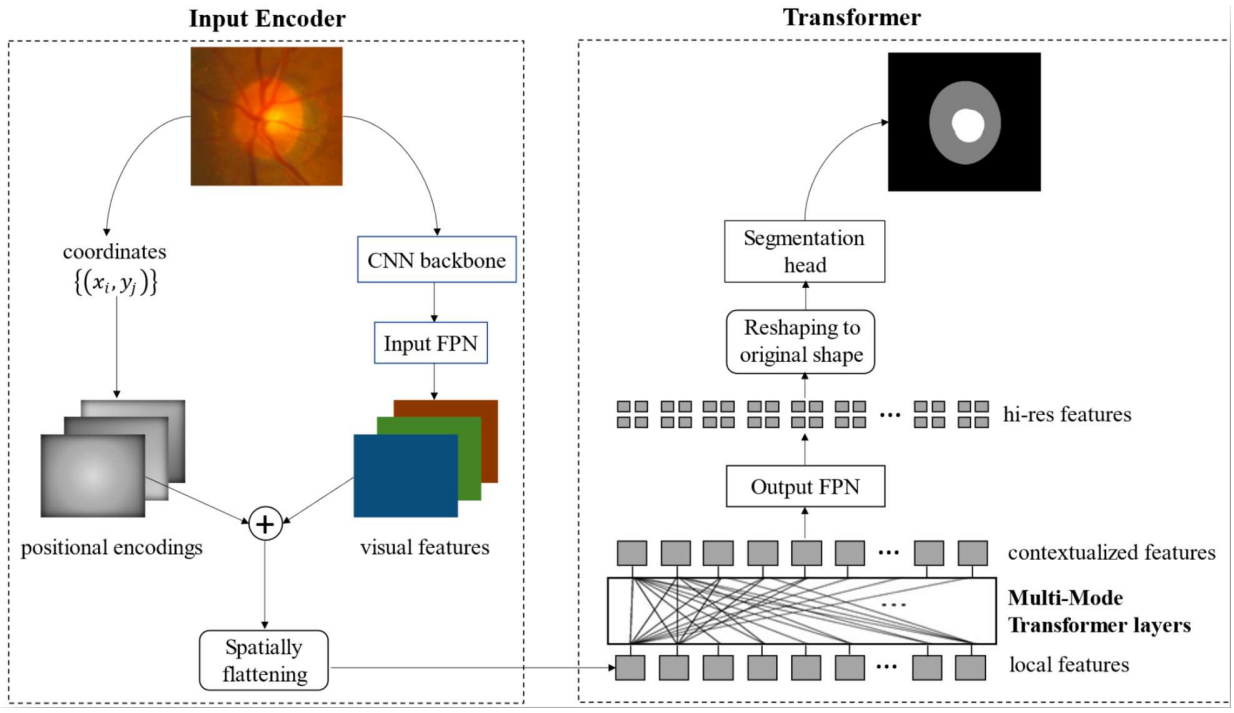


Figure 9: The framework of the EyeStar Team in task 2.

[2017]) before being classified by a segmentation head. In their experiments, EfficientNet-D4 (Tan and Le [2019]) is used as the CNN backbone; accordingly, the number of feature channels was set as 1792. To increase the spatial resolution of feature maps, they adopted an input FPN and an output FPN that upsample the feature maps at the transformer input end and output end, respectively. The positional encoding in their framework was learnable sinusoidal. Given a pixel coordinate (x, y), the C-dimension positional encoding vector $p(x, y)$ is:

$$p_i(x,y) = \begin{cases} sin(a_i x + b_i y + c_i), if i < C/2 \\ cos(a_i x + b_i y + c_i), if i \geq C/2 \end{cases}$$

where C is the same as the channel of the CNN backbone, i indexes the elements in p, $\{a_i, b_i, c_i\}$ are learnable weights. The (x, y) was normalized into [0,1] to maintain a consistent behavior across different image sizes. The visual features and positional encodings of the whole image are added up before being fed to the transformer: $X_{vol} = f(X_0) + p(X_0)$, where each image unit (a small downsampled patch) corresponds to a C-dimensional vector. The transformer model has a few transformer layers. Each layer fuses the features of input units based on their correlations, and outputs contextualized features. The core of a transformer layer is self attention, to improve self-attention for image applications, they proposed a novel Multi-Mode Transformer Layer. At last, the segmentation head is simply a 1 × 1 convolutional layer, whose number of output channels is 3 for the optic disc/cup segmentation task.

Since optic cup and optic disc occupy relatively small proportions in color fundus images, and the optic cup is inside the optic disc, many teams detected the coarse area of optic disc from the whole image first, and then segmented the fine optic disc and optic cup in this area. The CBMIBrand team devise the detection branch of Mask R-CNN to detect the region of optic disc and devise the mask branch with U-Net to segment both the optic disc and cup. The TeamTiger team and the cheeron team both segmented the optic disc region first using U-Net, and then the TeamTiger team adopted a generative adversarial netwrok, while the cheeron team adopted a ResU-Net to further segment the optic disc and cup. When segment the fine optic disc and cup from the coarse optic disc region, the MIG team adopted a CE-Net, which is based on U-Net model and ResNet34 module replace the encode path. To alleviate the consecutive pooling and strided convolutional operation led to reduce the feature resolution, loss of some spatial information, a context encoder module is proposed in CENet, which is composed of dense atrous convolution (DAC) block and a residual multi-kernel pooling (RMP) block. Finally, in the decode path, the decoder module first combines the features in the encode path by skip connection to obtain some spatial information, and then recovers the high-resolution features in the decoder by 1×1 convolution and 3×3 deconvolution respectively. Similar to the teams above, the MAI team utilized DeeplabV3+ framework to achieve the precise segmentation of OC/OD after obtaining a rough optic disc area using U-Net. Noticed that the OD and OC segmentation tasks are viewed as two independent binary tasks. In addition to the general binary

cross-entropy loss, MAI team used 1-IOU as an additional loss item. Moreover, they adopted a classic UDA strategy (see Section Strategy about Unsupervised Domain Adaptation) to maintain the segmentation performance on different devices.

The MIAG ULL team and the Pami-G team both considered the relationship of the optic disc and the fovea. The MIAG ULL team trained the model to learn the segmentation of optic disc and fovea simultaneously. And the Pami-G team designed two branches to train the model learning the segmentation and localization of the optic disc and fovea simultaneously (as shown in Fig. 7).

3.4 Localization of Fovea

This task is to localise the coordinate of fovea, which is the center of the macula. Table 7 shows the methods overview of the 10 participating teams. We can divide these fovea localization methods into the following three categories: direct determining by using the original image, determining by using the relative position strategy of the optic disc and the fovea, and determining by using the coarse-to-fine two-stage strategy.

Table 7: Methods overview of the participating teams in fovea localization task.

Team	Strategy	Archiecture	Additional dataset	Highlight
VUNO	whole image and corresponding vessel mask as input	EfficientNet	private dataset	feature extracted model was pre-trained by using samples with 15 lesion labels; and using the prior relationship between fovea and fundus vessel
CBMIBrand	coarse-to-fine localization	DenseNet121	IDRiD	coarse crop the obtained multi-scale macular region by using the position relationship between the optic disc and fovea
cheeron	objection detection	Yolo5	-	-
MIG	distance map regression	U-Net	-	designed Bi-Distance map for predicting the minimum distance to the optic disc or the fovea
TeamTiger	coarse-to-fine localization	EfficientNet for coarse localization, U-Net with EfficientNet as encoder for fine localization	-	
MAI	classic keypoint detection	U-Net with EfficientNet-B5 as encoder	-	test-time training
MIAG ULL	change to fovea segmentaiton task	ResNet50 PSPNet	-	simultaneously segment the optic disc and the fovea
ALISR	distance map regression	U-Net	MESSIDOR AMD, IDRiD(c), Messidor, PALM, EyePACS(part)	
Pami-G	simultaneously realize the segmentation and localization	FCN for segmentation; CNN for localization		segment and localise the optic disc and fovea simultaneously
EyeStar	coarse-to-fine localization	HRNet, StemNet, Mask RCNN	-	using the coarse and fine feature to fix the fovea position

The VUNO team designed a single pixel of fovea segmentation mask, and two deviation masks on the x and y axis, which was used to deal with the problem of the image scaling. Then, they used U-Net framework to predict the above three masks. In their framework, the fundus image and its corresponding vessel mask were input. The network consisted of two branches, one was EffcientNet-B4 to process the fundus image, and the other was EffcientNet-B0 to operate the vessel segmentation mask. The penultimate feature maps of the fundus branch were concatenated to those of the vessel branch. Then, decoding layers were appended to the concatenated feature maps using depth-wise separable convolutions. The final layer consisted of a confidence map, a map for x-offset, and a map for y-offset. Along with the loss functions at the last layers, they also gave losses to the final feature maps of the vessel branch. The MAI team utilized the classic keypoint detection technique (Yuan et al. [2019], Zhou et al. [2019]), which is frequently used in medical detection challenges in recent years with a potential to capture tiny regions accurately. In this regard, they perform the fovea localization based on an EfficientUNet (Tan and Le [2019]), where EfficientNet-B5 is utilized as the feature extractor. The ALISR team transformed the localization task to regress the distance map and utilized a pre-trained U-Net (Meyer et al. [2018]) to achieve the prediction. The cheeron team transformed the landmark localization of fovea into objection

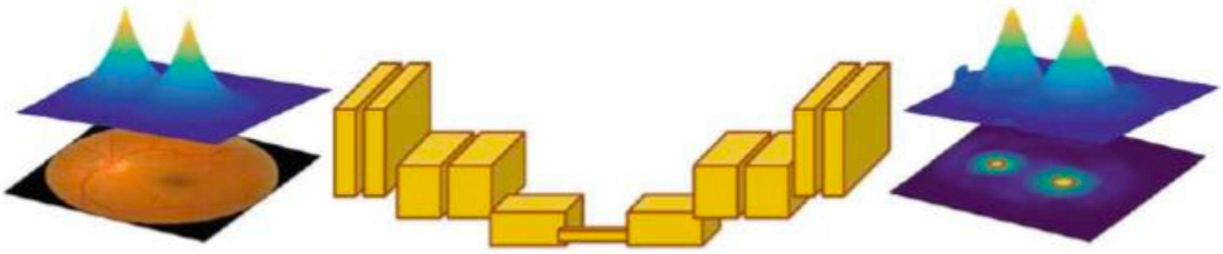


Figure 10: The method of the MIG team in task 3 for joint fovea and optic disc localization via regressing a distance map.

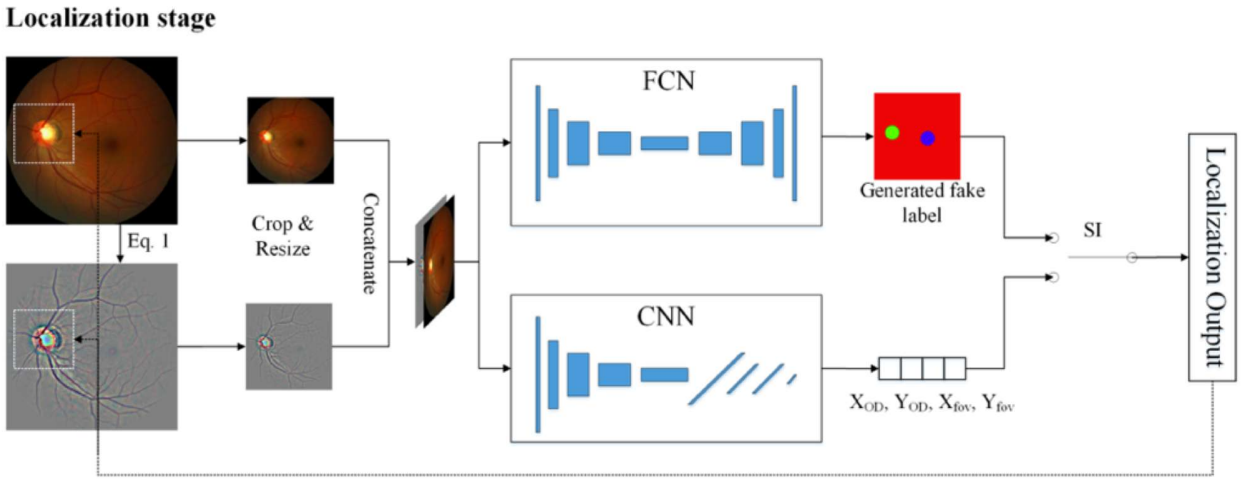


Figure 11: Localization stage of the Pami-G team.

detection task. They used the fovea location as the center of the object and used the disc radius as the width and height of the object. Then, the latest YOLO5 framework (Jocher) was adopted for the detection of fovea.

Referring to the strategy of Meyer's paper (Meyer et al. [2018]), the MIG team believed that joint learning of the position of each pixel associated with optic disc and fovea helps to automatically understand the overall anatomical distribution. At the same time, they transformed the localization problem into a distance map regression problem. As shown in Fig. 10, they utilized a Bi-Distance map for each pixel location (x, y), and adopted U-Net to solve the regression problem. In the Bi-Distance map, pixel value B(x,y) is the distance from the nearest landmark of interest (the optic disc center or the fovea). The MIAG ULL also considered the relationship between the optic disc and the fovea, they adopted ResNet50 and PSPNet to segment the optic disc and fovea simultaneously. Similarly, the Pami-G team also segment these two fundus structures, but they utilized an FCN framework. In addition, they also used another CNN branch to localise the center of the optic disc and the fovea (see Fig. 11). They designed an shape index (SI) to determine which fovea prediction is the final result. SI is defined as follow:

$$SI = \frac{C^2}{S} \tag{3}$$

where S and C denote the area and the perimeter of a region respectively in prediction. The shape of output with very small or large SI was considered as a failure case. They set a threshold that $SI \in (11, 12.2)$. When $SI \in (11, 12.2)$, they used the center of FCN output region as the localization result and they used the CNN output when $SI \notin (11, 12.2)$.

The TeamTiger team designed a coarse-to-fine fovea segmentation method. The initial model was an EfficientNet based model to get a tentative position of the fovea, and the next model worked on the patch around the coarse fovea area. The second model consists of encoder-decoder architecture having EfficientNet as encoder and U-Net like decoder to achieve the fovea segmentation. The center coordinate of the fovea segmentation mask is the fovea localization result. The CBMIBrand team also adopted the coarse-to-fine strategy. They first cropped the coarse fovea patches according to the positional relationship between the optic disc center and the fovea center. Specifically, they obtained the center of

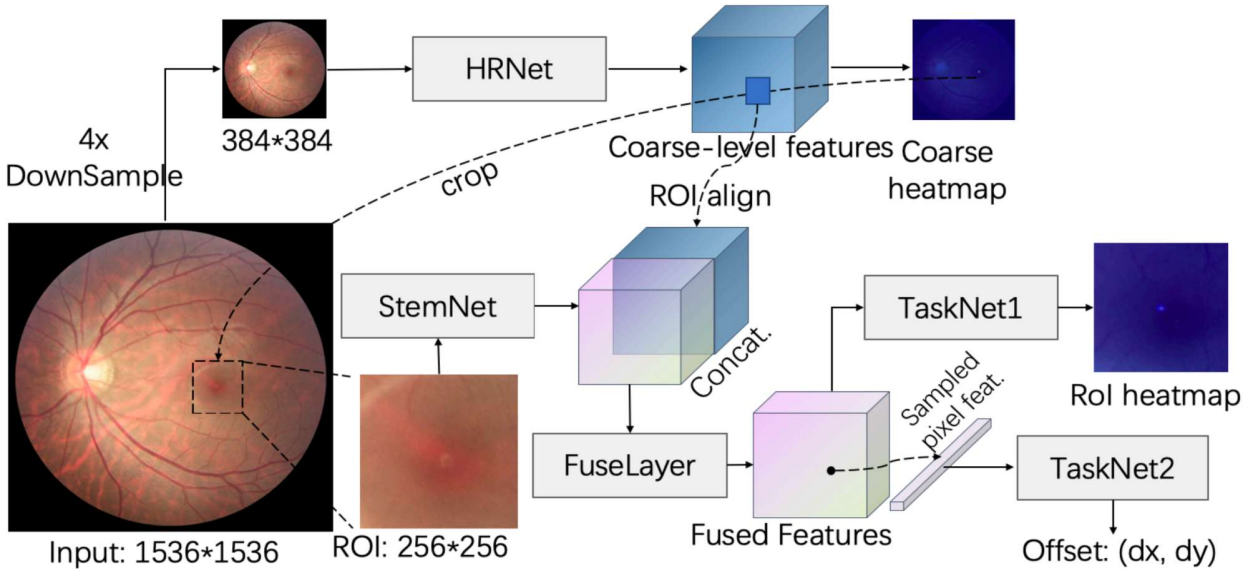


Figure 12: The architecture of the EyeStar team method in task 3.

OD based on task 2. Second, starting from the OD center, they approximated the fovea region by searching the region directing down of 1/6 OD diameter (ODD) and right/left of 2.5 ODD (right eye and left eye can be distinguished). Finally, they cropped multi-scale squared fovea ROI (384x384, 448x448, 480x480, 512x512, 576x576, 640x640 pixels), and resized them into 512x512 pixels. They concatenated the multi-scale patches at the channel dimension and finally adjusted DenseNet to perform a fovea coordinate regression. Similarly, the EyeStar team adopted the coarse-to-fine strategy, but they fused the local feature of the whole image and the global feature of the cropped patch for the fovea localization. As can be seen in Fig. 12, the input image was down-sampled by 4x and fed into pre-trained HRNet (Sun et al. [2019]) to get per-pixel coarse heatmaps. The peak pixel was then located. A ROI region on the original input image, centered at the predicted peak pixel at coarse scale, was then cropped and fed into the StemNet (Sun et al. [2019]) to obtain fine-scale features. The fine-scale features were concatenated with the pixel-aligned coarse-level features (using the ROIAlign2 layer), and processed by FuseLayer. The fused features are finally passed through the TaskNet1 to get fine-scale heatmap and through the TaskNet2 to predict the offset of the sampling location to the ground truth. In their framework, the FuseLayer was a convolutional block that outputs 32 feature channels. The TaskNet1 consisted of two convolutional blocks with 32 and 1 channel. TaskNet2 was a multi-layer perceptron with 32, 16 and 2 channels.

4 Results

This section presents the results on the REFUGE2 online and onsite set of the above teams. All these 10 teams scored valid points in all three tasks on the online set, whereas 7, 6, and 6 teams had valid scores in task 1, 2, and 3 on the onsite set. The evaluation on the onsite set in this paper only contained the teams with valid scores. The full tournament rankings according to the online and onsite sets of the REFUGE2 challenge is available at `https://refuge.grand-challenge.org/Final_Leaderboards/`. The tables displaying the results in our paper present the actual rankings of the online and onsite set of these discussed teams in this article.

4.1 Classification of clinical Glaucoma

The evaluations of the glaucoma classification task on the online and onsite set, in term of AUC, are presented in Table 8. From the table, we can see that on the online set, the AUC for all teams are more than 0.93, and the first three teams are Pami-G, MAI, and VUNO EYE TEAM with the AUC of 0.9841,0.9840,and 0.983, respectively. On the onsite set, VUNO EYE TEAM reported the best performance with AUC achieving 0.883. And the second and third ranked teams are MIG and MAI with AUC of 0.876 and 0.861, respectively. To illustrate the effectiveness of AI technology in glaucoma classification based on fundus color photography, we also included an additional approach based on using the vCRD values of the ground truth as a likelihood for glaucoma classification. On the online and the onsite sets, the AUC of the approach based on the vCRD values are respectively 0.8817, and 0.7571. Fig. 13 and Fig. 14 respectively present

Table 8: Evaluation in term of AUC of the results of 10 teams in the glaucoma classification task on the online and onsite datasets.

Team Name	Online	Online Rank	Onsite	Onsite Rank
Pami-G	0.9841	2	-	-
MAI	0.9840	3	0.861	3
VUNO EYE TEAM	0.983	5	0.883	1
cheeron	0.980	6	0.856	4
TeamTiger	0.968	8	-	-
CBMIBrand	0.962	9	-	-
ALISR	0.947	12	0.844	6
EyeStar	0.944	14	0.820	9
MIG	0.943	16	0.876	2
MIA GULL	0.939	17	0.847	5

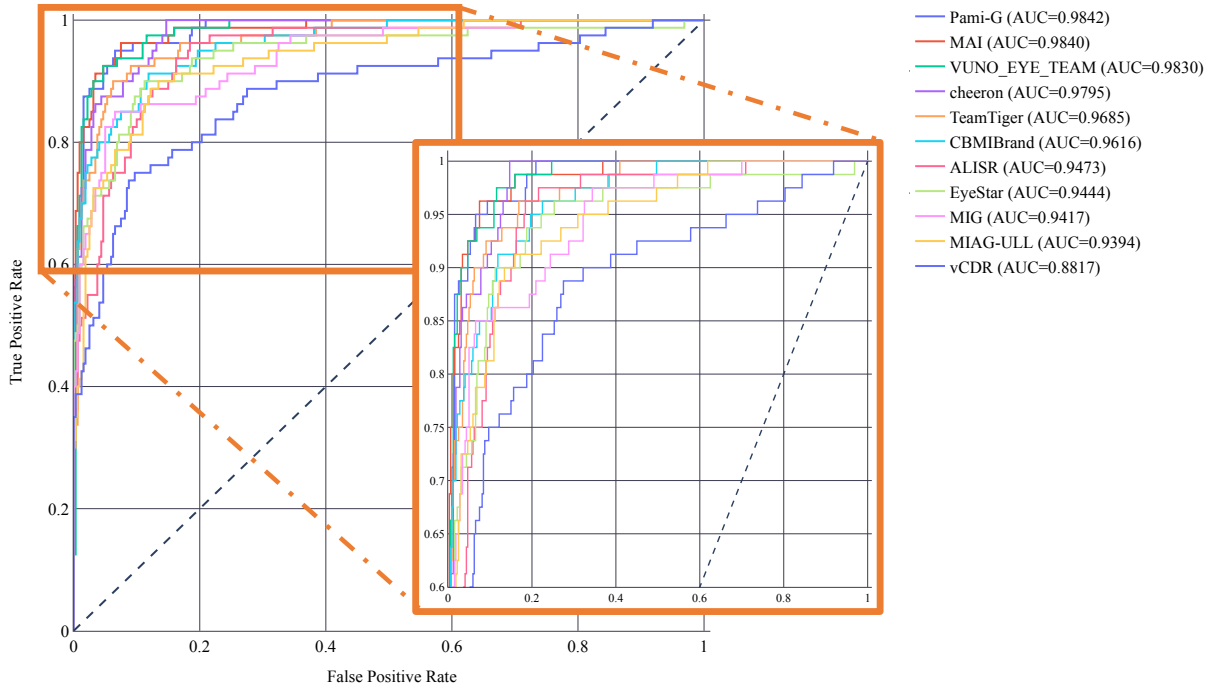


Figure 13: ROC of the glaucoma classification results on the online dataset of the 10 teams.

the ROC curves of the methods of the teams with valid scores and the vCRD values-based method on the online and onsite sets.

4.2 Segmentation of Optic Disc and Cup

Table 9 and Table 10 summarize the OD and OC Dice and vCDR MAE metrics of each team on the online and onsite set. From the two tables, we can see that the Dice values of OD are greater than that of OC, which is mainly because OC are smaller and more difficult to segment. Specifically, on the online set, MAI reached the best segmentation performance according to the rules reported in Challenge Evaluation with OD Dice of 0.966, OC Dice of 0.880, and vCDR MAE of 0.037. The cheeron and CBMIBrand teams were the second and third places. To compare the statistical significance of the differences in the metric values of the top three teams, we adopted Mann-Whitney U hypothesis test with $\alpha = 0.05$. For OD segmentation, compared with cheeron and CBMIBrand, MAI was not statistical significantly different with respect to them (cheeron p=0.8037, CBMIBrand p=0.3715). No statistical significant different results also occurred among the first team respect to the second and third teams for OC segmentation (cheeron p=0.1821, CBMIBrand p=0.0889) and vCDR evaluation (cheeron p=0.2521, CBMIBrand p=0.1814). On the onsite set, the first place was

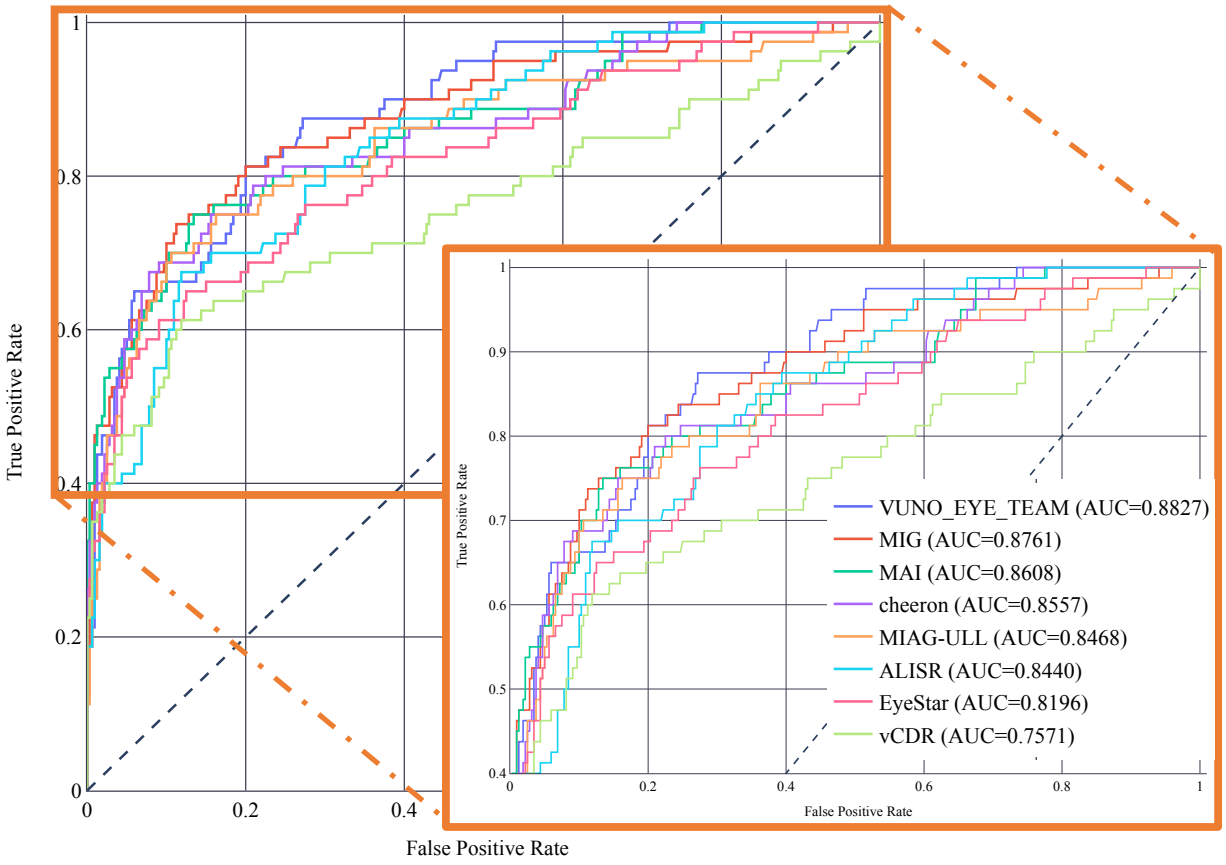


Figure 14: ROC of the glaucoma classification results on the onsite dataset of the 7 teams.

Table 9: Evaluation in term of OC Dice, OD Dice and vCDR MAE of the results of 10 teams in the segmentation of optic disc and cup task on the online dataset.

Team Name	OC Dice	OD Dice	vCDR MAE	Rank
MAI	0.880	0.966	0.037	1
cheeron	0.874	0.965	0.038	2
CBMIBrand	0.874	0.964	0.039	3
VUNO EYE TEAM	0.870	0.966	0.040	4
EyeStar	0.873	0.961	0.039	5
Pami-G	0.865	0.967	0.042	6
TeamTiger	0.871	0.961	0.042	7
MIAG ULL	0.854	0.934	0.044	16
MIG	0.825	0.959	0.060	19
ALISR	0.826	0.942	0.056	21

cheeron with OD Dice of 0.961, OC Dice of 0.865, and vCDR MAE of 0.055. For OD segmentation, compared with MAI and VUNO EYE TEAM-the second and third teams, respectively-the differences were not significant (MAI p=0.5879, CBMIBrand p=0.5075). For OC segmentation, the differences in the OC Dice values achievedy by cheeron were statistically significant with respect to VUNO EYE TEAM (p=1.0047×10^{-6}), except to MAI (p=0.1057). For vCDR estimation, cheeron was with no significant differences with respect to the MAI and VUNO EYE TEAM (MAI p0.0715, VUNO EYE TEAM p=0.0795). The distribution of OD Dice, OC Dice and vCDR MAE values of each team on the online and onsite sets are represented as boxplots in Figs. 15 and 16.

Table 10: Evaluation in term of OC Dice, OD Dice and vCDR MAE of the results of 6 teams in the segmentation of optic disc and cup task on the onsite dataset.

Team Name	OC Dice	OD Dice	vCDR MAE	Rank
cheeron	0.865	0.961	0.055	1
MAI	0.854	0.960	0.060	2
VUNO EYE TEAM	0.845	0.960	0.058	2
MIG	0.846	0.949	0.055	4
EyeStar	0.831	0.939	0.054	5
MIAG ULL	0.851	0.918	0.064	8

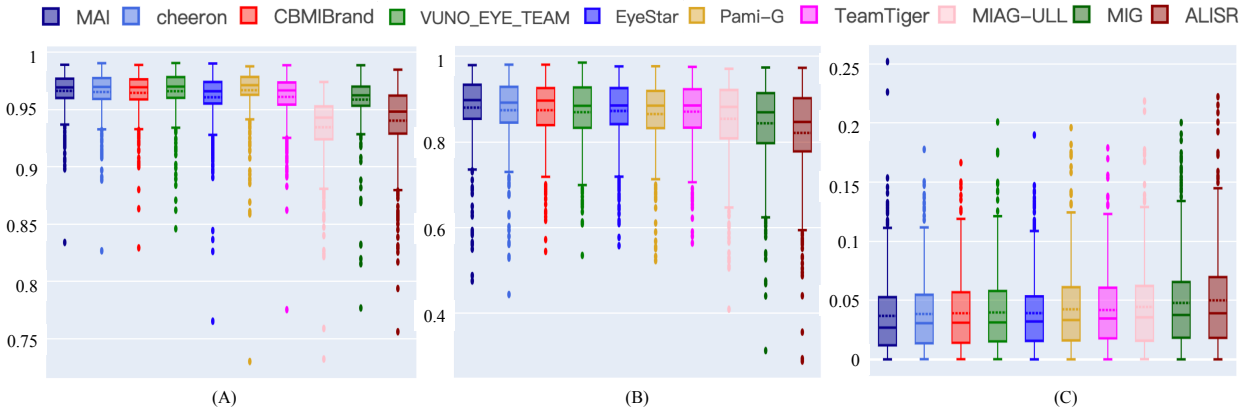


Figure 15: Box-plots illustrating the performance of each optic disc/cup segmentation method in the online dataset. Distribution of Dice values for (A) optic disc, (B) optic cup, and (C) mean absolute error (MAE) of the estimated vertical cup-to-disc-ratio (vCDR).

Fig. 17 shows the edges of the optic disc and cup segmentation results on the online and onsite sets of the top 3 teams, respectively. In Figs. 17(A) and (B), green, blue, red and yellow lines respectively present the ground truth, and the segmentation results of MAI, cheeron, and CBMIBrand. Similarly, in Figs. 17(C) and (D), the four color lines respectively present the ground truth, the results of cheeron, MAI, and VUNO EYE TEAM. From the figure, we can see that the segmentation results from both glaucoma and non-glaucoma images can cover the target region, but the effect at the edge needs to be improved. Fig. 18 shows the overlapping display of the optic disc and cup segmentation

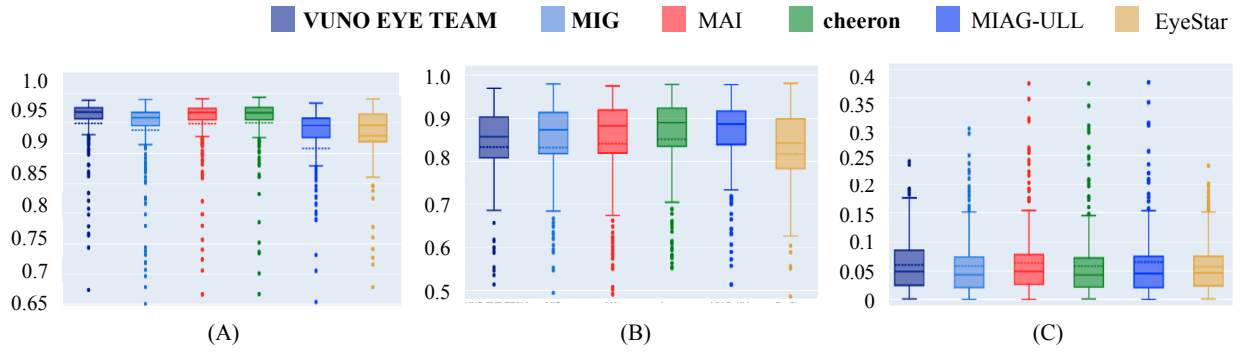


Figure 16: Box-plots illustrating the performance of each optic disc/cup segmentation method in the onsite dataset. Distribution of Dice values for (A) optic disc, (B) optic cup, and (C) mean absolute error (MAE) of the estimated vertical cup-to-disc-ratio (vCDR).The top 3 teams are highlighted in bolds (cheeron, MAI, and VUNO EYE TEAM).

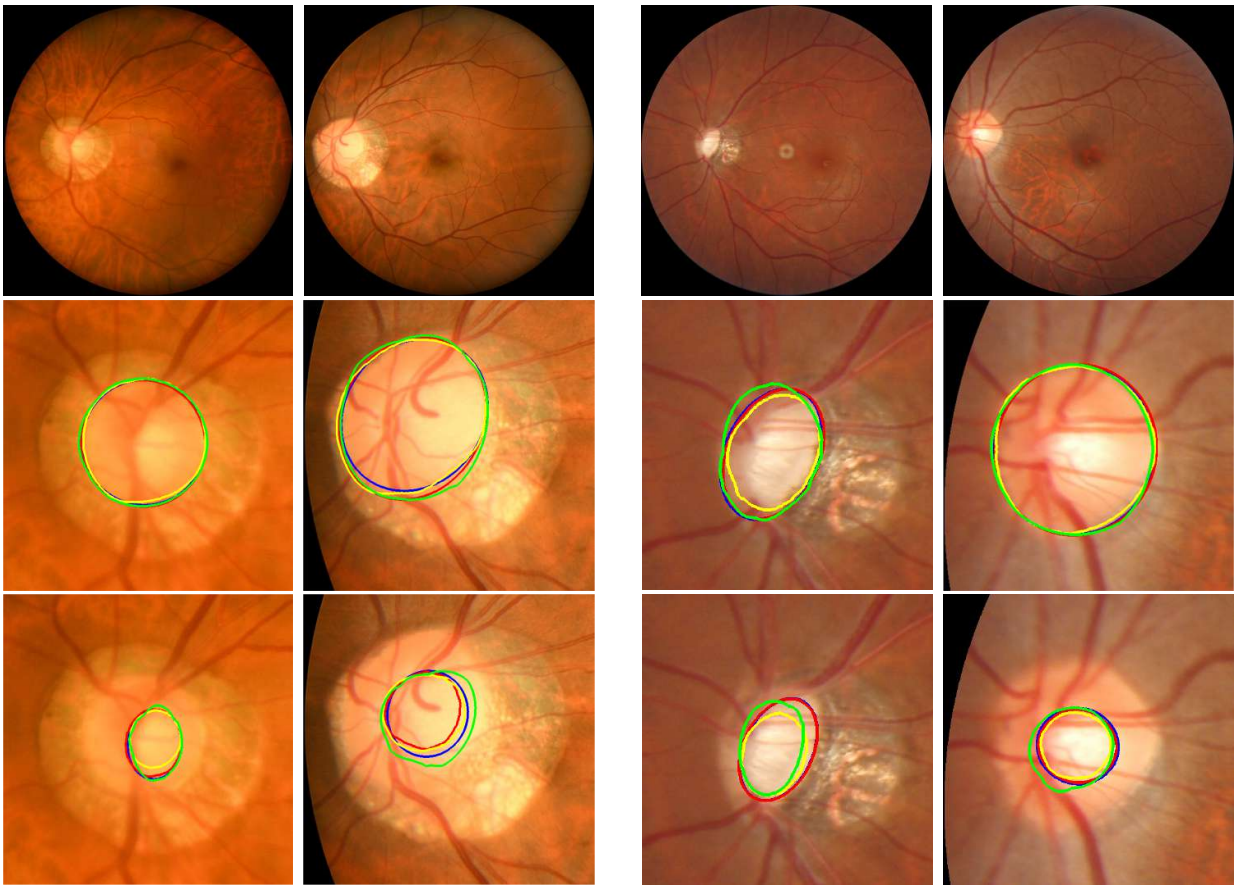


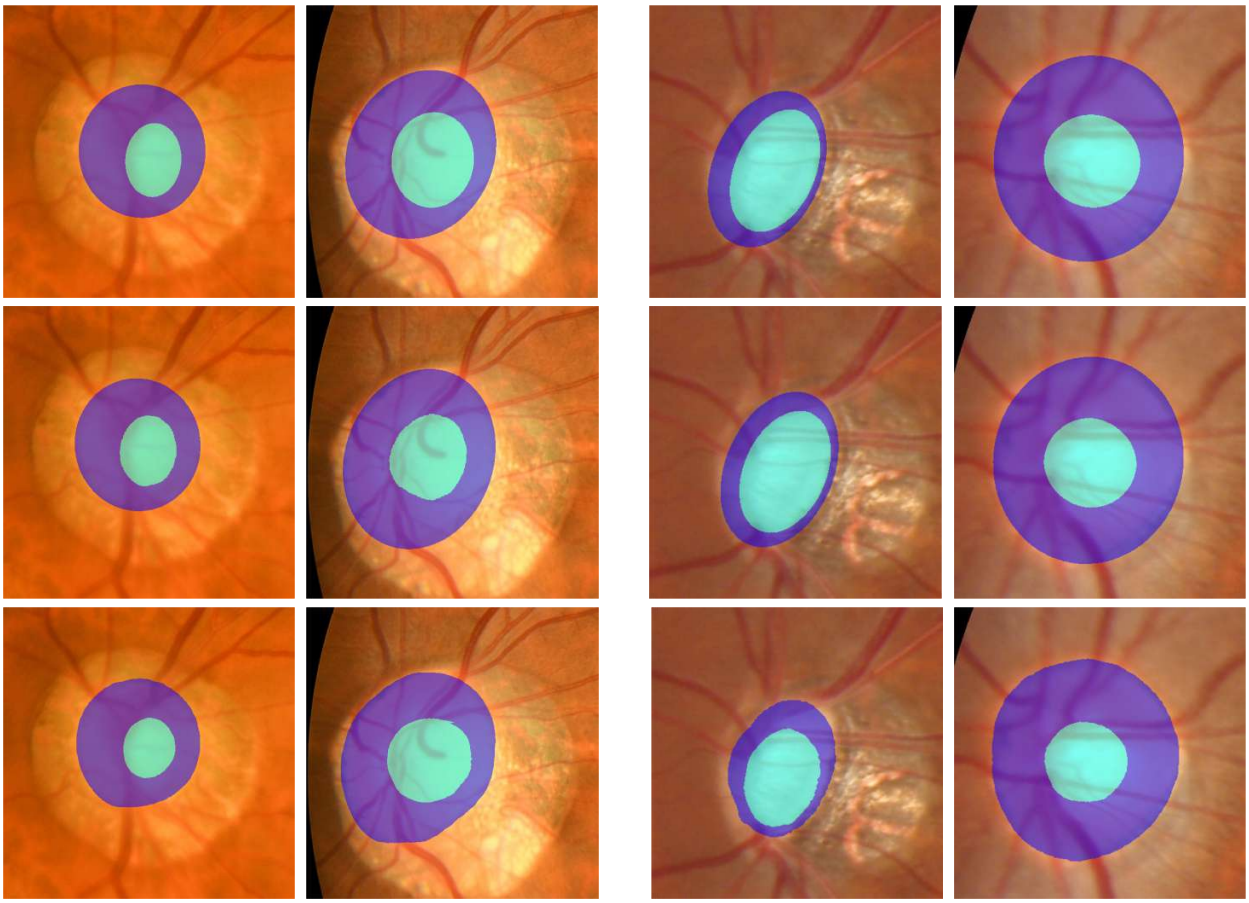
Figure 17: Demonstrate of the edges of the optic disc and cup segmentation results on the online and onsite datasets of the top 3 teams. (A) Glaucoma sample in the online dataset, (B) non-glaucoma sample in the online dataset; Green: Ground truth, Blue:MAI, Red: cheeron, Yellow: CBMIBrand. (C) glaucoma sample in the onsite dataset, (D) non-glaucoma sample in the onsite dataset. Green: Ground truth, Blue:cheeron, Red:MAI, Yellow: VUNO EYE TEAM.

results and the original images. Examples from the first row to the last are corresponding to the top 3 teams on the online and onsite sets, respectively.

4.3 Localization of Fovea

Table 11 summaries the results of the 10 teams in the fovea localization task on both online and onsite sets. From the table, we can see that on the online set, the first 3 teams are MAI, VUNO EYE TEAM, and Pami-G with the AED of 8.412, 8.727, and 9.457 pixels. Compare with VUNO EYE TEAM and Pami-G, the performance of MAI was only statistical significantly different with respect to Pami-G (VUNO EYE TEAM p=0.6356, Pami-G p=0.0003). On the onsite set, the first 3 teams with valid scores are MAI, VUNO EYE TEAM, and cheeron with AED of 21.841, 27456, and 28.344 pixels. The performance of MAI was not statistical significantly different with respect to VUNO EYE TEAM and cheeron (VUNO EYE TEAM p=0.3764, cheeron p=0.3283). The distribution of AED values obtained by the participating teams are represented as boxplots in Fig.19.

Fig.20 shows the fovea localization results of the top 3 teams on the online and onsite sets, respectively. In the figure, the first row show the original images, and the second row show the corresponding fovea localization results in the form of cross mark. Similarly as Fig. 17, we show the results on the glaucoma and non-glaucoma samples. From the figure, it can be seen that the localization results of the top 3 teams on online and onsite sets are near the ground truths on both glaucoma and non-glaucoma samples.



(A) (B) (C) (D)

Figure 18: The optic disc and cup segmentation results on the online and onsite datasets of the top 3 teams. (A) Glaucoma sample in the online dataset, (B) non-glaucoma sample in the online dataset, examples from the first row to the last are corresponding to the teams MAI, cheeron, and VUNO EYE TEAM; (C) glaucoma sample in the onsite dataset, (D) non-glaucoma sample in the onsite dataset, examples from the first row to the last are corresponding to the teams cheeron, MAI, and VUNO EYE TEAM.

5 Discussion

In this section, we discuss the findings from the challenge results, and discuss the significance of the REFUGE2 challenge to the AI technology and clinical applications in ophthalmology, as well as the future work.

5.1 Findings

Considering the results of the three sub-tasks, it can be seen that the prediction results of each team on the online set are better than those on the onsite set. The main reason for this is that teams can tune their models on the online set. In the three tasks, the MAI team using UDA strategy had excellent performances on both online and onsite set (fundus images collected from KOWA and Topcon), which the scores of the three tasks were all in the top three. This indicates that the UDA strategy designed for training and inference processes with different data distributions makes the model more robust than the model using conventional training and inference methods. In addition, most of the participating teams used data augmentation in their training processing to suppress overfitting. The data augmentation methods of both geometric transformation and color transformation aim to expand the data distribution and make the training model more capable of generalization. The different solutions are discussed below for the specific tasks.

Task 1: Classification of Clinical Glaucoma

Table 11: Evaluation in term of AED of the results of 10 participating teams in the fovea localization task on both online and onsite datasets.

Team Name	Online	Online Rank	Onsite	Onsite Rank
MAI	8.412	1	21.841	1
VUNO EYE TEAM	8.727	3	27.456	3
Pami-G	9.457	4	-	-
cheeron	10.086	6	28.344	4
EyeStar	10.096	7	43.982	6
CBMIBrand	11.699	9	-	-
MIAG ULL	24.439	16	-	-
TeamTiger	33.778	19	-	-
MIG	35.741	20	105.812	8
ALISR	111.239	22	173.405	9

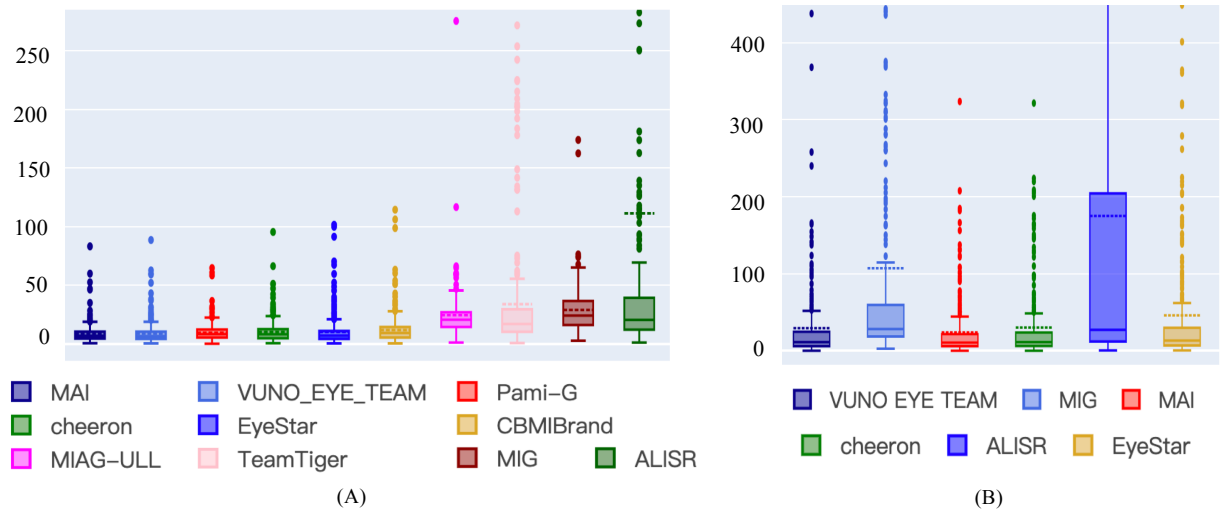
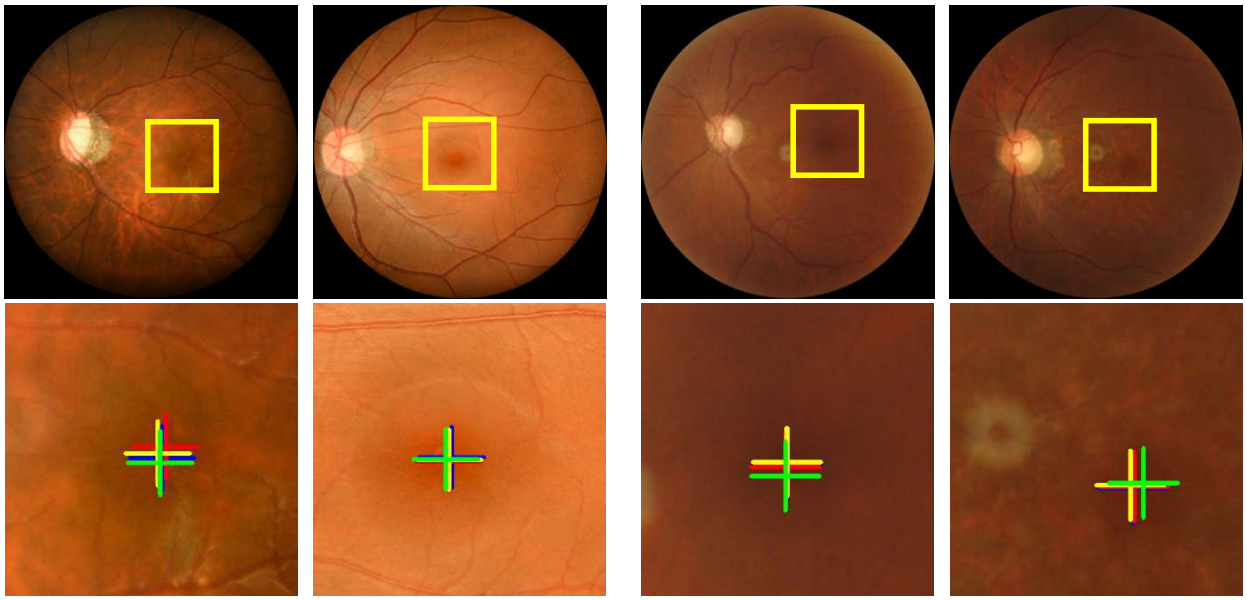


Figure 19: Box-plots illustrating the performance of each fovea localization method in the online (A) and onsite (B) dataset.

Glaucoma will lead to degeneration around the optic disc region, such as vertical cup-disk ratio expansion, optic disc bleeding, optic nerve rim notching and other signs (Aung and Crowston [2016]). Hence 7 teams predicted the glaucoma using the local optic disc region. Among them, the CBMIBrand team cropped multi-scale patches around the optic disc to simulate the different sizes of optic discs collected in different sample images. The MIG team utilized not only the local patch of the optic disc area, but also a whole image, which condering both local and global information of the fundus images. The network archiectures used of these teams in glaucoma classification task include EfficinetNet, ResNet, DenseNet, VGG, encoder module of U-Net, which are the common and efficient convolutional neural networks for image classification. In addition, the ALISR team adopted a CSPNet to mitigate the problem that previous works require heavy inference computations from the network architecture perspective. The solution to alleviate computational stress in the application of AI technology is a valuable topic (Wang et al. [2020]). In addition to discussing the results of the participating teams, we also calculated the predicted results for glaucoma by vCDR values. From Figs. 13 and 14, we can see that the glaucoma prediction of vCDR-based method is worse than that of AI methods based on the images. In KOWA data (online set), the AUC of Pami-G team with the best performance among the participating teams is 0.1025 higher than that of vCDR-based method. Similarly, in Topcon data (onsite set), the AUC of the best team-VUNO EYE TEAM is 0.1256 higher than that of vCDR-based method.

Task 2: Segmentation of Optic Disc and Cup

From the first two rows of the Table 3, we can see that the OD and OC areas account for a small proportion in the whole fundus image. Hence, for the tiny region segmentation task, there are 8 teams adopted the solutions with coarse-to-fine strategy. When building the model, The VUNO EYE TEAM took into account the position relationship between optic disc and blood vessels, so the vessel segmentation mask was added into the input module to supplement information.



(A) (B) (C) (D)

Figure 20: Fovea localization results of the top 3 teams and the enlarged display. (A) Glaucoma sample in the online dataset, (B) non-glaucoma sample in the online dataset, Green: Ground truth, Blue: MAI, Red: VUNO EYE TEAM, Yellow: Pami-G; (C) glaucoma sample in the onsite dataset, (D) non-glaucoma sample in the onsite dataset, Green: Ground truth, Blue: MAI, Red: VUNO EYE TEAM, Yellow: cheeron.

While conduct the OD/OC segmentation by using the original patch of the optic disc area, the Pami-G team also used polar transformation to obtain spatial constraint and augment the cup/disc region proportion (Fu et al. [2018]) for easy segmentation. From the aspect of the network architecture used by the teams, U-Net and its variants are popular for the segmentation task. Notably, the EyeStar team used vision transformer technology, which emerged as a competitive alternative to convolutional neural networks that are currently state-of-the-art in computer vision (Khan et al. [2021]). However, this technique has high requirements on the number of training samples and computing resources. In terms of segmentation results, MAI team used UDA strategy and achieved top three in the KOWA and Topcon data. In addition, the cheeron team, which utilized Atrous Spatial Pyramidal Pooling, deep supervision, and test-time augmentation strategies, also achieved excellent segmentation results on the datasets collecting from different machines. Based on the qualitative results in Fig. 18, we can see that the vCDR of the glaucoma samples shown in Fig. 18(A) was not significantly amplified, indicating that it is not enough to apply only vCDR value as the screening standard for glaucoma. This conclusion is consistent with the conclusion of the glaucoma classification experiment based on vCDR value in Task 1. In addition, we have calculated the differences between the initial annotations, which were delineated by different glaucoma specialists, and the ground truth of the optic cup and disc segmentation task in the onsite set. The calculation results showed that the best performance of the manual delineation was $OC\ Dice = 0.865$, $OD\ Dice = 0.952$, $vCDR\ MAE = 0.049$, the worst one was $OC\ Dice = 0.742$, $OD\ Dice = 0.817$, $vCDR\ MAE = 0.084$. As can be seen from Table 10, the optic disc and cup segmentation results obtained automatically by each team are all better than the worst one of the initial annotation. Meanwhile, the performance of the cheeron team (Rank 1) are closest to the best performance of the manual delineation. This indicates that the automatic segmentation method can achieve similar or even better segmentation effect than manual, which can assist doctors to quickly observe the structure of optic cup and optic disc in clinical practice in the future, which is of great significance.

Task 3: Localization of Fovea

In the fovea localization task, we observed that the proposed solutions mostly transformed the localization task to segmentation, distance map regression, and object detection tasks. The corresponding ground truths were also converted into the corresponding forms. In the distance map regression and target segmentation tasks, common network frameworks such as U-Net and ResNet were mainly used. For object detection task, cheeron team utilized YOLO5, the latest version of the classic and effective YOLO series. It can be seen that the models in computer vision field is well applied to the medical image processing. In addition, the VUNO EYE TEAM considered vessel information into the network together, and the Pami-G team simultaneously processed the segmentation and localization of optic disc and fovea, all of which applied the clinical knowledge of the location correlation between optic disc, macula and

blood vessels in the fundus. According to the fovea localization results, we can see that the MAI team using UDA stragegy, the VUNO EYE TEAM and Pami-G team that combined clinical knowledge, and the cheeron team that used the latest object detection framework achieved excellent results in this task. This indicates that the training strategy and network framework designed for specific medical tasks are helpful to improve the performance of the model. We also calculated the differences between the initial annotations, which were labeled by different specialists, and the ground truth of the fovea localization. The AED of the manual annotation result closest to the ground truth is 22.936 pixels, and the AED of the worst performance is 27.408 pixels. The average AED of the results of each annotator is 24.961 pixels. As we can see from Table 11, only the 1st team MAI outperformed all the manually annotated results, and better than the best manually annotated results. This was mainly due to the fact that the localization task was harder than the segmentation task, so not all teams work well on the onsite dataset. Still, the performances of the VUNO EYE TEAM and cheeron teams can be close to the worst manual labeling result on the onsite dataset. This shows that it is hopeful for automatic localization methods to achieve similar results to manual annotation in the fovea localization task, which plays an important role in clinical observation of macular structure.

5.2 Challenge significance, limitations and future

REFUGE2 has released the first multi-model and multi-label color fundus image dataset, and the labels of different tasks combine the knowledge and experiences of 8 physicians. This dataset can be used for researchers to study the application of AI algorithm in glaucoma classification, optic cup and disc segmentation and fovea localization, and to study the performance migration of AI algorithm on different machine models. Moreover, the REFUGE2 dataset can provide the multi-annotation for multi-rater studies in the future (Ji et al. [2021]). At present, Li et al.(Li et al. [2021b]) had studied few-shot domain adaptation using REFUGE2 dataset, and their work was published in MICCAI 2021. Previous datasets published via iChallage, such as ADAM, PALM, and REFUGE1, have been used by many researchers and the REFUGE1 challenge review paper (Orlando et al. [2020b]) has been cited more than 140 times.

The limitation of the REFUGE2 dataset is the lack of demographic information, such as age distribution and source scenarios (clinic, community). Besides, the data are all collected from the Chinese population, which lacks ethic diversity. In the future data preparation, we will pay attention to the supplement of the above information. In addition, as the clinical diagnosis of glaucoma involves not only fundus color examination, but also OCT, visual field test and other examinations, we will pay more attention to the topics related to multi-modality data analysis in the future challenge. Apart from the discrimination on whether the samples are glaucoma, clinical attention is also paid to the degree of disease. Therefore, in the future, we will also design competitions to focus on the grading of glaucoma.

6 Conclusion

In this paper, we summarized the released dataset, methods and results of the REFUGE2 challenge. We analyzed the multi-models dataset provided by REFUGE2 challenge, and it was observed that there were differences in the distribution of fundus images collected by different models. These data are valuable for studying unsupervised domain adaptation. We summarized the solutions adopted by 10 teams for the three sub-tasks (glaucoma classification, optic disc/cup segmentation, fovea localization) of the challenge, focusing on the training and inference strategies designed for different data distribution, as well as the strategy of combining AI technology with clinical prior knowledge.

We analyzed the valid performances of the teams participating in the online and onsite challenge at MICCAI 2020. We observed that the UDA strategy and the use of large amount of datasets had a good effect in our challenge. The combination of clinical prior knowledge (such as the position relationship between optic disc, macula, vessels in the fundus, and the importance of optic disc area for the diagnosis of glaucoma) and AI technology gave promising results on all three tasks designed by the competition. In addition, the glaucoma prediction results of AI based on image are superior to those of vCDR value based detection, indicating that AI technology can be applied to glaucoma screening and has advantages.

In summary, the dataset released in the REFUGE2 challenge is the first open multi-model fundus dataset focusd on glaucoma classification, optic disc/cup segmentation, and fovea localization. In addition, REFUGE2 challenge provided a unified evaluation framework for the above three tasks. The data and evaluation framework are publicly accessible through the Grand Challenge website at `https://refuge.grand-challenge.org/Home2020/`. Future participants are welcome to use our dataset and submit their results on the website and use it for benchmarking their methods. REFUGE2 is designed to advance AI research on color fundus photography and help researchers further explore its clinical implications. In the future, we will continue to promote this kind of competition and welcome the active participation of scholars.

References

Yong Han, Weiming Li, Mengmeng Liu, Zhiyuan Wu, Feng Zhang, Xiangtong Liu, Lixin Tao, Xia Li, Xiuhua Guo, et al. Application of an anomaly detection model to screen for ocular diseases using color retinal fundus images: Design and evaluation study. *Journal of medical Internet research*, 23(7):e27822, 2021.

T Aung and J Crowston. *Asia pacific glaucoma guidelines*. Kugler Publications, 2016.

Meindert Niemeijer, Michael D Abràmoff, and Bram Van Ginneken. Fast detection of the optic disc and fovea in color fundus photographs. *Medical image analysis*, 13(6):859–870, 2009.

Tao Li, Wang Bo, Chunyu Hu, Hong Kang, Hanruo Liu, Kai Wang, and Huazhu Fu. Applications of deep learning in fundus images: A review. *Medical Image Analysis*, 69:101971, 2021a.

Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *arXiv preprint arXiv:2201.09873*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Artem Sevastopolsky. Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network. *Pattern Recognition and Image Analysis*, 27(3):618–624, 2017.

Shuang Yu, Di Xiao, Shaun Frost, and Yogesan Kanagasingam. Robust optic disc and cup segmentation with deep learning for glaucoma detection. *Computerized Medical Imaging and Graphics*, 74:61–71, 2019.

Shujun Wang, Lequan Yu, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Patch-based output space adversarial learning for joint optic disc and cup segmentation. *IEEE transactions on medical imaging*, 38(11):2485–2495, 2019a.

Huazhu Fu, Jun Cheng, Yanwu Xu, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE transactions on medical imaging*, 37(7):1597–1605, 2018.

Baidaa Al-Bander, Waleed Al-Nuaimy, Bryan M Williams, and Yalin Zheng. Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc. *Biomedical Signal Processing and Control*, 40:91–101, 2018.

Md Kamrul Hasan, Md Ashraful Alam, Md Toufick E Elahi, Shidhartho Roy, and Robert Martí. Drnet: Segmentation and localization of optic disc and fovea from diabetic retinopathy image. *Artificial Intelligence in Medicine*, 111: 102001, 2021.

Zhixi Li, Yifan He, Stuart Keel, Wei Meng, Robert T Chang, and Mingguang He. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*, 125(8):1199–1206, 2018.

Muhammad Naseer Bajwa, Muhammad Imran Malik, Shoaib Ahmed Siddiqui, Andreas Dengel, Faisal Shafait, Wolfgang Neumeier, and Sheraz Ahmed. Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning. *BMC medical informatics and decision making*, 19(1):1–16, 2019.

Yuming Jiang, Lixin Duan, Jun Cheng, Zaiwang Gu, Hu Xia, Huazhu Fu, Changsheng Li, and Jiang Liu. Jointrcnn: a region-based convolutional neural network for optic disc and cup segmentation. *IEEE Transactions on Biomedical Engineering*, 67(2):335–343, 2019.

Ruben Hemelings, Bart Elen, Joao Barbosa-Breda, Sophie Lemmens, Maarten Meire, Sayeh Pourjavan, Evelien Vandewalle, Sara Van de Veire, Matthew B Blaschko, Patrick De Boever, et al. Accurate prediction of glaucoma from colour fundus images with a convolutional neural network that relies on active and transfer learning. *Acta ophthalmologica*, 98(1):e94–e100, 2020.

Richard M Felder and Rebecca Brent. Active learning: An introduction. *ASQ higher education brief*, 2(4):1–5, 2009.

Yalin Zheng, Mohd Hanafi Ahmad Hijazi, and Frans Coenen. Automated "disease/no disease" grading of age-related macular degeneration by an image mining approach. *Investigative ophthalmology & visual science*, 53(13):8310–8318, 2012.

T Kauppi, V Kalesnykiene, et al. Diaretdb0-standard diabetic retinopathy database, calibration level 0. imageret project 2007.

Tomi Kauppi, Valentina Kalesnykiene, Joni-Kristian Kamarainen, Lasse Lensu, Iiris Sorri, A Raninen, R Voutilainen, J Pietilä, H Kälviäinen, and H Uusitalo. Diaretdb1—standard diabetic retinopathy database calibration level 1, 2007.

Enrique J. Carmona, Mariano Rincón, Julián Garcí a Feijoó, and José M. Martínez-de-la Casa. Identification of the optic nerve head with genetic algorithms. *Artif. Intell. Med.*, 43(3):243–259, July 2008. ISSN 0933-3657. doi:10.1016/j.artmed.2008.04.005. URL http://dx.doi.org/10.1016/j.artmed.2008.04.005.

Jayanthi Sivaswamy, SR Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A Ujjwaft Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, pages 53–56. IEEE, 2014.

Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004.

L Giancardo et al. The hamilton eye institute macular edema dataset (hei-med). *GitHub https://github.com/lgiancaUTH/HEI-MED*, 2012.

Attila Budai, Rüdiger Bock, Andreas Maier, Joachim Hornegger, and Georg Michelson. Robust vessel segmentation in fundus images. *International journal of biomedical imaging*, 2013, 2013.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.

Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014.

Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. 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*, pages 3065–3068. IEEE, 2010.

Ahmed Almazroa, Sami Alodhayb, Essameldin Osman, Eslam Ramadan, Mohammed Hummadi, Mohammed Dlaim, Muhannad Alkatee, Kaamran Raahemifar, and Vasudevan Lakshminarayanan. Retinal fundus images for glaucoma analysis: the riga dataset. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, volume 10579, page 105790B. International Society for Optics and Photonics, 2018.

Francisco Fumero, Silvia Alayón, José L Sanchez, Jose Sigut, and M Gonzalez-Hernandez. Rim-one: An open retinal image database for optic nerve evaluation. In *2011 24th international symposium on computer-based medical systems (CBMS)*, pages 1–6. IEEE, 2011.

Mani Baskaran, Reuben C Foo, Ching-Yu Cheng, Arun K Narayanaswamy, Ying-Feng Zheng, Renyi Wu, Seang-Mei Saw, Paul J Foster, Tien-Yin Wong, and Tin Aung. The prevalence and types of glaucoma in an urban chinese population: the singapore chinese eye study. *JAMA ophthalmology*, 133(8):874–880, 2015.

Michael Goldbaum. The stare project, structured analysis of the retina database. *Zuletzt abgerufen am*, 27, 2013.

José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020a.

Chonglin Chen and Gang Wang. Iosuda: an unsupervised domain adaptation with input and output space alignment for joint optic disc and cup segmentation. *Applied Intelligence*, 51(6):3880–3898, 2021.

Haijun Lei, Weixin Liu, Hai Xie, Benjian Zhao, Guanghui Yue, and Baiying Lei. Unsupervised domain adaptation based image synthesis and feature alignment for joint optic disc and cup segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2021.

Shreya Kadambi, Zeya Wang, and Eric Xing. Wgan domain adaptation for the joint optic disc-and-cup segmentation in fundus images. *International Journal of Computer Assisted Radiology and Surgery*, 15:1205–1213, 2020.

Peng Liu, Bin Kong, Zhongyu Li, Shaoting Zhang, and Ruogu Fang. Cfea: collaborative feature ensembling adaptation for domain adaptation in unsupervised optic disc and cup segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 521–529. Springer, 2019.

Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Boundary and entropy-driven adversarial learning for fundus image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 102–110. Springer, 2019b.

Cheng Chen, Quande Liu, Yueming Jin, Qi Dou, and Pheng-Ann Heng. Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 225–235. Springer, 2021.

R GeethaRamani and Lakshmi Balasubramanian. Macula segmentation and fovea localization employing image processing and heuristic based clustering for automated retinal screening. *Computer methods and programs in biomedicine*, 160:153–163, 2018.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008.

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005.

Huazhu Fu, Fei Li, et al. AGE challenge: Angle Closure Glaucoma Evaluation in Anterior Segment Optical Coherence Tomography. *Medical Image Analysis*, 66:101798, dec 2020. ISSN 13618415.

Agung W Setiawan, Tati R Mengko, Oerip S Santoso, and Andriyan B Suksmono. Color retinal image enhancement using clahe. In *International Conference on ICT for Smart Society*, pages 1–3. IEEE, 2013.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, pages 9229–9248. PMLR, 2020.

Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018.

Giovanna Guidoboni, Rachel Shujuan Chong, Nicholas Marazzi, Miao Li Chee, Jessica Wellington, Emily Lichtenegger, Ching-Yu Cheng, and Alon Harris. A mechanism-driven algorithm for artificial intelligence in ophthalmology: Understanding glaucoma risk factors in the singapore eye diseases study. *Investigative Ophthalmology & Visual Science*, 61(7):619–619, 2020.

Jaemin Son, Joo Young Shin, Hoon Dong Kim, Kyu-Hwan Jung, Kyu Hyung Park, and Sang Jun Park. Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology*, 127(1):85–94, 2020.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020.

Jose Sigut, Omar Nunez, Francisco Fumero, Marta Gonzalez, and Rafael Arnay. Contrast based circular approximation for accurate and robust optic disc segmentation in retinal images. *PeerJ*, 5:e3763, 2017.

Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

Andrew G. Howard, Menglong Zhu, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv:1704.04861 [cs]*, April 2017.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for Activation Functions. *arXiv:1710.05941 [cs]*, October 2017.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

Chenglang Yuan, Cheng Bian, Hongjian Kang, Shu Liang, Kai Ma, and Yefeng Zheng. Identification of primary angle-closure on as-oct images with convolutional neural networks. *arXiv preprint arXiv:1910.10414*, 2019.

Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

Maria Ines Meyer, Adrian Galdran, Ana Maria Mendonça, and Aurélio Campilho. A pixel-wise distance regression approach for joint retinal optical disc and fovea detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 39–47. Springer, 2018.

Glenn Jocher. Yolov5. `https://github.com/ultralytics/yolov5`. Accessed: Aug, 2020.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021.

Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12341–12351, 2021.

Shaohua Li, Xiuchao Sui, Jie Fu, Huazhu Fu, Xiangde Luo, Yangqin Feng, Xinxing Xu, Yong Liu, Daniel SW Ting, and Rick Siow Mong Goh. Few-shot domain adaptation with polymorphic transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 330–340. Springer, 2021b.

José Ignacio Orlando, Huazhu Fu, et al. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, page 21, 2020b.