

Eye Disease Diagnosis and Fundus Synthesis: A Large-Scale Dataset and Benchmark

Xue Xia^{*†}, Kun Zhan^{*§}, Ying Li^{*}, Guobei Xiao[†], Jinhua Yan[†], Zhuxiang Huang^{*}, Guofu Huang[†], Yuming Fang^{*}

^{*}Jiangxi University of Finance and Economics, [†]The First Hospital of Nanchang

Email: [†]hbxue@i.shu.edu.cn, [§]2201601740@stu.jxufe.edu.cn

Abstract—As one of the most common imaging modalities, retinal fundus imaging offers images of interior surface of eyes for initial examination of disorders. Data-driven machine learning methods, especially deep learning models in recent years, provide automatic ophthalmological disease diagnosis techniques from color fundus images. Data with high quality, diversity and balanced distribution supports deep model-based eye disease diagnosis. However, many existing datasets focus on a specific kind of eye disease, and some suffer from label noise or quality degeneration, which hinders automatic screening algorithms from dealing with multiple eye diseases. To solve this, we propose a high-quality dataset containing 28877 color fundus images for deep learning-based diagnosis. Except for 15000 healthy samples, the dataset consists of 8 eye disorders including diabetic retinopathy, age-related macular degeneration, glaucoma, pathological myopia, hypertension, retinal vein occlusion, LASIK spot and others. Based on this, we propose a co-attention network for disease diagnosis, establish benchmark on screening and grading tasks, and demonstrate that the proposed dataset supports generative adversarial network-based image synthesis. The dataset will be made publicly available.

Index Terms—dataset, eye disease diagnosis, fundus synthesis

I. INTRODUCTION

Eye diseases are the leading causes of visual disturbances or blindness if left untreated. Nowadays, eye healthy has become a global priority [1]. Ophthalmoscopic examination [2] and machine learning techniques brought groundbreaking progress to enable non-invasive automatic diagnosis. Data matters in computer-aided biomedical diagnostics since either ophthalmologists or algorithms acquire disease-related observations from retinal fundus images. Especially as data-driven instruments, deep networks rely on versatile data to achieve significant gains in terms of effectiveness.

Although some researchers aim at dealing with scarce data [3], [4], high-quality data plays a substantial role in providing complex anatomical details and tiny retinal structures for deep models. Some works established datasets for specific medical tasks like cell detection and counting [5], retinal image quality assessment [6] and registration [7] to tackle the problem of data-hungry.

Existing ocular disease diagnosis strategies can be coarsely divided into two categories, which are single task diagnosis

This work was supported in part by the National Natural Science Foundation under Grant 62132006 and 62162029, in part by the Shenzhen Municipal Science and Technology Innovation Council under Grant 2021Szzvup051, and in part by the Natural Science Foundation of Jiangxi Province under Grant 20202ACB202007 and 20202BABL212007.

TABLE I
COMPARISON OF SOME AVAILABLE OCULAR DATASETS.

Dataset name	Sample numbers	Disease(s) and tasks(s)
STARE [14]	397	14 diseases
E-OPHTHA [16]	463	DR Lesion Detection
Messidor [9]	1200	DR Grading
Messidor2 [10]	1748	DR Grading
FGADR [8]	2842	DR Grading, Lesion Segmentation
RFMiD [15]	3200	46 diseases
ODIR [12]	10000	7 diseases
EDDFS(Ours)	28877	8 diseases

and multi-task one. The former presents disease screening or grading by classification, while the latter involves assistant tasks such as lesion identification or localization as the improvements. Both categories rely on multi-class or binary classification. Furthermore, for multi-disease diagnosis, as is often the case in fundus image analysis, multi-label classification offers the solution. Accordingly, commonly used ocular datasets consist of either samples of one specific disease [8]–[11] or multi-disease samples [12]–[15]. Datasets with multi diseases or multi-label annotations mainly endure data unbalanced, poor quality or difficulty bias, i.e., specific diseases are too easy to be identified. Therefore, we present a multi-disease dataset with multi-label and high quality for ocular disease screening, grading and retinal fundus image synthesis. Our main contributions are as follows:

- We collect and annotate a high-quality retina fundus dataset that may support data-driven eye disease diagnosis and fundus image synthesis.
- We propose a co-attention network with cross-stage feature fusion that forces global self-attention to cooperate with local attention information for more accurate disease diagnosis.
- We present benchmark and comparison experiments on our dataset among commonly used deep networks to support further research.

II. OUR DATASET

A. Data Collection and Annotation

We collected a multi-label multi-class dataset for eye disease diagnosis and fundus synthesis (EDDFS), which contains 28877 color retinal fundus images. The images were collected from 5 different fundus cameras, two of which share

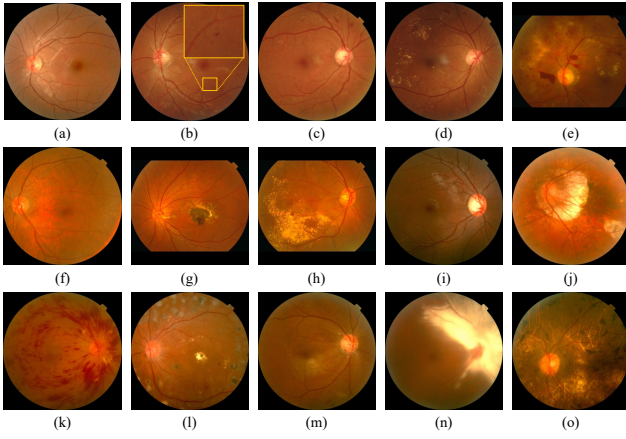


Fig. 1. Examples of fundus images in EDDFS. (a) Normal, (b) mild NPDR, (c) moderate NPDR, (d) severe NPDR, (e) proliferative DR, (f) early&intermediate AMD, (g) advanced dry AMD, (h) advanced wet AMD, (i) glaucoma, (j) myopia, (k) RVO, (l) LASIK spot, (m) hypertension, (n-o) other diseases.

the same resolution. The original fundus image resolutions 768×576 , 1956×1934 , 2976×3158 and 3264×2448 , and the corresponding image numbers are 9176, 3602, 6266 and 9833, respectively. To present professional and reasonable annotations, the whole annotation was organized by an ophthalmologist together with 6 experts from the same hospital. Four participants from our team were involved as assistants, who underwent a training program lasting for 6 months. The trainee who failed in the final test will be weeded out as unsuitable. Finally, 10 annotators conducted three rounds of annotation on more than forty thousand real fundus images. Annotators were split into a team of two, and each team was responsible for the same sub-set of images. Only the annotations reaching consistency between the two annotators in a team were considered available. In addition, spot checks and discussions were carried out after every round of annotation to improve the efficiency and effectiveness of labeling.

CNNs always require a fixed size of inputs in a batch, hence we adopt center crop and zero-padding to obtain non-wrapped square images, and then resize them to the size of 1024×1024 . It's worth noting that we do not involve any content wrapping, color correction or illumination modification. Although data augmentation is a conventional instrument, we keep images original, as illustrated in Fig. 1.

Considering the fact that a majority of samples are healthy fundus, we randomly selected 15 thousand images as normal samples and screened out the category with scarce samples for the sake of data balance. Furthermore, to guarantee high availability of data, images of an outer eye or those that suffer from low quality were also considered unusable. As a result, the presented EDDFS contains 28877 high-quality color retinal fundus images with image-level annotations, some of which are multi-label samples. Comparison of some public ocular datasets is shown in Table I. In particular, in the local enlarged view of picture (b), a small focal microaneurysm at the stage of mild non-proliferative DR (mild NPDR) can be observed.

TABLE II
SAMPLE NUMBERS AND AVAILABLE TASKS OF EACH OCULAR DISEASE.

Disease	Numbers	Split1		Split2		
		Train	Test	Train	Val	Test
Normal (N)	15000	9000	6000	/	/	/
DR (D)	1660	996	664	814	135	409
AMD (A)	7095	4257	2838	3987	663	1997
Glaucoma (G)	815	489	326	379	63	191
Myopia (M)	1901	1141	760	1093	181	545
RVO (RV)	227	166	111	137	24	68
LASIK Spot (LS)	1308	785	523	624	104	313
Hypertension (H)	719	431	288	313	53	158
Others (O)	924	554	370	492	81	247
Total	28877	17326	11551	7839	1304	3928

B. Data Split and Preprocessing

We adopt two versions of data split methods to conduct diagnosis experiments. For multi-label disease diagnosis and single disease grading, the data is split into a training set containing 60% of samples and testing one consisting of the rest samples. While for single-label multi-disease diagnosis, the data with only disease samples are split into training, validating and testing sets by 6:1:3. Both versions of splits are implemented by random selection from every category to keep data balance.

As demonstrated in Table II, 8 diseases are involved in our data including diabetic retinopathy (DR), age-related macular degeneration (AMD), glaucoma, pathological myopia (Myopia), retinal vein occlusion (RVO), laser-assisted in situ keratomileusis spot (LASIK spot), hypertension and other diseases. Noting that the number of RV sample is limited, not all following experiments involve this class. In addition, the total number does not equal the sum of disease-wise sample numbers due to multi-label annotations.

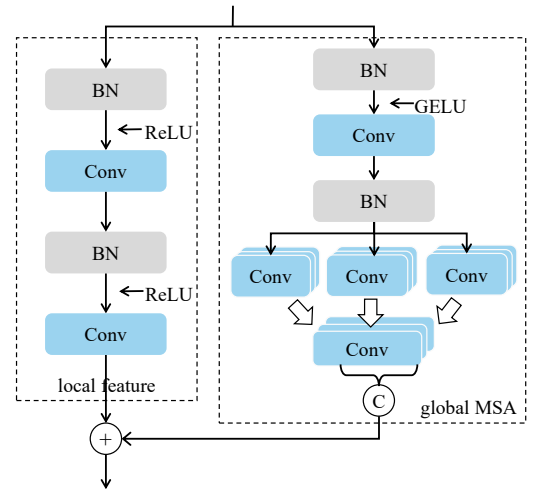


Fig. 2. The proposed co-attention structure involving local feature extraction path (left) and global multi-head self-attention (MSA) path (right).

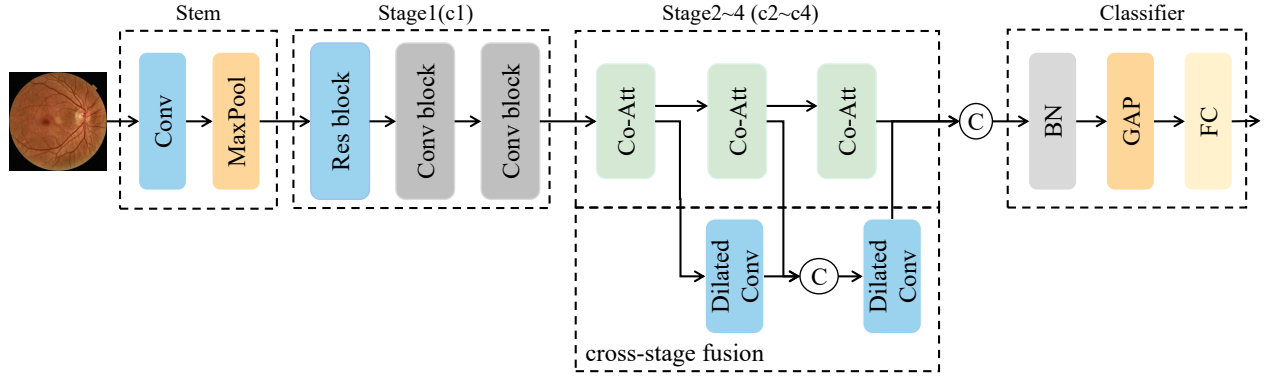


Fig. 3. The proposed disease diagnosis network with stacked co-attention (Co-Att) structure and cross-stage feature fusion. The stem is a convolution stem, and the classifier consists of batchnorm, global average pooling and fully connected layers.

TABLE III
COMPARISON RESULTSON MULTI-LABEL MULTI-DISEASE DIAGNOSIS.

Metrics	Models	D	A	G	M	RV	LS	H	O
Pre	Res18	0.8268	0.8403	0.6220	0.9039	0.7564	0.9444	0.5286	0.5951
	Dense121	0.8359	0.8363	0.6409	0.9151	0.7229	0.9521	0.5172	0.6524
	In3	0.7992	0.8381	0.6570	0.8967	0.7263	0.9249	0.5326	0.6222
	AlterNet-T	0.8109	0.8561	0.7095	0.9155	0.7011	0.9375	0.5032	0.6226
	OURS	0.8086	0.8576	0.7045	0.9114	0.7612	0.9260	0.5333	0.6684
AUC	Res18	0.9334	0.9212	0.9450	0.9924	0.9673	0.9877	0.8752	0.8975
	Dense121	0.9340	0.9156	0.9461	0.9917	0.9644	0.9857	0.8736	0.8973
	In3	0.9293	0.9227	0.9500	0.9905	0.9729	0.9855	0.8578	0.9011
	AlterNet-T	0.9325	0.9293	0.9408	0.9931	0.9757	0.9854	0.8773	0.9079
	OURS	0.9415	0.9307	0.9466	0.9931	0.9729	0.9899	0.8811	0.9034
MSE	Res18	0.0511	0.1133	0.0344	0.0199	0.0108	0.0171	0.0409	0.0445
	Dense121	0.0478	0.1179	0.0333	0.0209	0.0107	0.0194	0.0412	0.0430
	In3	0.0509	0.1228	0.0344	0.0226	0.0104	0.0162	0.0431	0.0431
	AlterNet-T	0.0485	0.1065	0.0315	0.0190	0.0112	0.0171	0.0408	0.0441
	OURS	0.0468	0.1035	0.0326	0.0193	0.0108	0.0157	0.0401	0.0434

III. THE PROPOSED METHOD

A. Co-Attention Structure

Inspired by the global and local feature extractions in traditional machine learning, we proposed a cooperating attention structure to simultaneously obtain local information and global multi-head self-attention, termed as co-attention structure.

Conventional convolutions acquire features within local areas, while the self-attention structure depicts long-range information. Existing networks combine the two features sequentially, while we try to present them in parallel to preserve information and to conduct feature fusion in summation. The proposed co-attention structure is shown in Fig. 2.

B. Attention Disease Diagnosis Network with Cross-stage Feature Fusion

Based on the proposed co-attention structure, we embedded the stacked structure into AlterNet [17] for substituting the alternating pattern of convolution and MSA operations. Besides, we design a simple yet efficient cross-stage feature fusion to preserve features in different attention stages. The proposed fusion method simply involves dilated convolution for larger

active areas, through which multi-resolution features caused by down-sampling in different stages can be adapted.

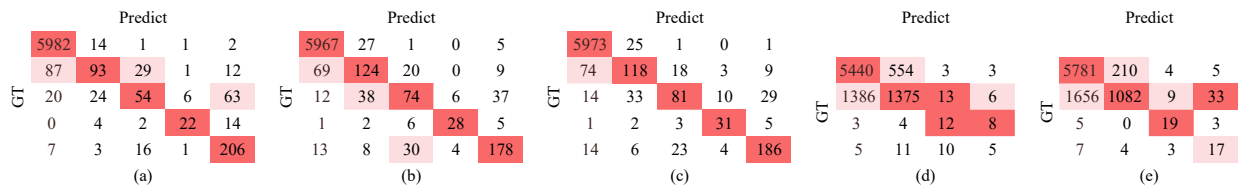
The overall framework of the proposed network is illustrated in Fig. 3. The blue rectangle marked with “Conv” stands for convolution with ReLU and batchnorm (BN), while the “Conv block” contains several layers of “Conv.” GAP stands for global average pooling, FC represents a fully connected layer and “C” stands for concatenation.

IV. BENCHMARK

The experiments were conducted on the computer with an AMD Ryzen 5950X Processor, 64GB RAM and a Nvidia GeForce RTX 3090 GPU. Comparisons in the same table followed the same configuration to ensure fairness.

A. Disease Screening

To establish a benchmark on the EDDFS dataset and evaluate the performance of some popular deep models. We involve ResNet-18 (abbreviated as Res18) [18], Densenet-121 (abbreviated as Dense121) [19], Inception-v3 (abbreviated as In3) [20], AlterNet-T [17] and our model in the experiments. In order to adapt to the comparative experiment, the pre-trained



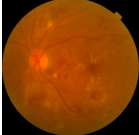


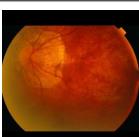
	GT: D	Res18: RV (0.581) Dense121: D (0.877) In3: A (0.971) AlterNet-T: D (0.602) OURS: D (0.874)
	GT: A	Res18: M (0.864) Dense121: A (0.661) In3: A (0.975) AlterNet-T: A (0.349) OURS: A (0.826)
	GT: G	Res18: G (0.737), A (0.717) Dense121: G (0.782) In3: A (0.947) AlterNet-T: G (0.829) OURS: G (0.833)
	GT: M	Res18: M (0.476) Dense121: D (0.451) In3: M (0.766), A (0.730) AlterNet-T: M (0.974) OURS: M (0.828)

Fig. 5. The visualization of several samples. The values in parentheses are the prediction confidence scores of each model.

models are not adopted. The performance is listed in Table III, in which the best results are bolded. For classification of data with multi-label, we compute metrics, i.e., precision (Pre), area under curve (AUC) and mean square error (MSE), globally by counting the total true positives, false negatives and false positives. Furthermore, several samples are shown in Fig. 5.

Based on this, the analysis can be three folds. 1) Deeper models do not always acquire better results. We attribute this to the motivation of these models, i.e., they were designed for natural images. While medical images hold some domain-related semantic information that should be captured by a specific feature extractor rather than simply deeper layers. 2) However, for hypertension (H) samples, deeper models exhibit more true positives because signs such as tortuosity of the retinal vessels or indentation of veins are not visually obvious for models. Thus left it difficult to recognize hypertension from samples with different diseases. 3) Different from the results reported in papers about natural image analysis, ViT demonstrates bad performance on our data even with pre-trained model. The reason may be that the global attention modules in ViT bring a large number of weights, thus millions

TABLE IV
DATA DISTRIBUTION OF DISEASE GRADING TASK.

Diseases	Grades	Sample Number	Total Number
DR	0: no DR	15000	15000
	1: mild NPDR	756	1850
	2: moderate NPDR	407	
	3: sever NPDR	166	
	4: proliferative DR	521	
AMD	0: no AMD	15000	15000
	1: early&intermediate	6910	7095
	2: advanced dry	93	
	3: advanced wet	92	

of diverse samples are always required.

B. Disease Grading

According to the International Classification of Diabetic Retinopathy (ICDR), DR severity is coarsely divided into non-proliferative DR (NPDR) and proliferative DR (PDR), the former of which can be further graded into mild NPDR, moderate NPDR and sever NPDR, as shown in Table IV. While AMD stages are generally early, intermediate and advanced ones, and advanced stage exhibits dry AMD (non-neovascular AMD) and wet AMD (neovascular AMD) [21]. Considering the difficulties of labeling, the ophthalmologist adopted the criterion in Table IV for AMD grading, which leads to unbalanced data distribution.

Single-disease grading is intrinsically a multi-class classification on single-label data, in which each sample belongs to exactly one class. Therefore, weighted averages of precision (Pre), recall (Recall) and f1-measure (F1) across disease grades are reported in Table V, considering the case of unequally distributed data. Models that performed well in the above experiments were involved in DR and AMD grading, and pre-training was discarded to present the ability of the original models. The experiments in Table V followed the same training configuration, *i.e.*, the Adamw optimizer with a learning rate of 9×10^5 was involved, and focal loss was used as the criterion. Besides, the batch size was set to 32 and the epoch was 51.

It is clear in Table V that our co-attention network outperforms others in DR grading while slightly inferior to Inception-v3 in terms of recall and F1. Still, we achieve the top 2 in these grading tasks.

To gain insights into disease grading, confusion matrices on several baselines are recorded in Fig. 4. It’s worth noting that deep models achieving good results in DR grading task do not

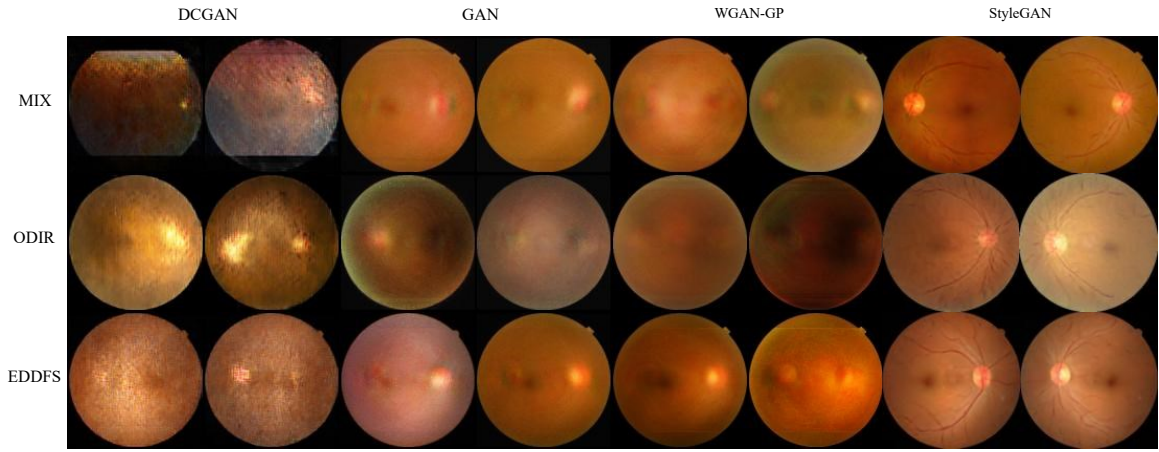


Fig. 6. Examples of images synthesized by the same model through RFMiD and APTOS2019 mixed dataset(MIX), EDDFS and ODIR at the same training epoch.

TABLE V
EXPERIMENTAL RESULTS OF DR AND AMD GRADING. THE BEST RESULTS ARE BOLD AND THE SECOND ONES ARE IN BLUE.

Diseases	Metrics	Res18	Dense121	In3	AlterNet-T	OURS
DR	Pre	0.9508	0.9475	0.9485	0.9492	0.9522
	Recall	0.9550	0.9518	0.9530	0.9533	0.9550
	F1	0.9523	0.9492	0.9501	0.9509	0.9532
AMD	Pre	0.8098	0.8134	0.8151	0.8013	0.8152
	Recall	0.8126	0.8176	0.8191	0.8054	0.8184
	F1	0.8027	0.8114	0.8132	0.7962	0.8116

TABLE VI
QUALITY ASSESSMENT ON 4 DATASETS.

“Good” rate(%)	ODIR	RFMiD	FADGR	EDDFS
healthy samples	70.99	92.71	-	88.31
all samples	49.92	63.03	62.45	77.24

perform that good in AMD task. We argue that this is because of the extremely unbalanced AMD data. Most true positives occur in grade 0 (no AMD) that holds the most samples. This kind of model bias happens in both DR and AMD grading, but more severe in AMD one. To solve this, learning strategy or specifically designed models should be explored in the future.

V. FUNDUS SYNTHESIS AND DATASET ANALYSIS

To present that our dataset is in high quality, we conduct image quality assessment on both original images through a “Good” rate [6] and synthesized images in terms of PSNR, a commonly used no reference metric.

The “Good” rate is designed for fundus images by [6], and it’s a “Good” sample number over the total sample number. The evaluation results are recorded in Table VI, where our data achieves the best average score over all samples. However, some lesions (e.g., retinal detachment or choroidal neovascularization caused by myopia) may decrease fundus quality, thus we also compute a “Good” rate for healthy samples. Although RFMiD exceeds EDDFS in the “Good” rate

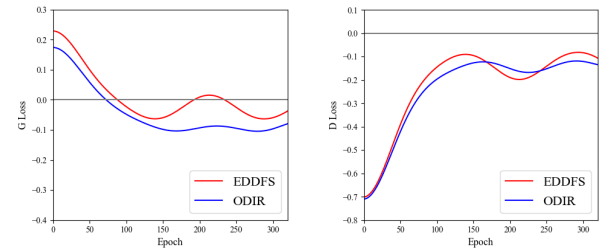


Fig. 7. Comparison between EDDFS and ODIR in terms of loss curves of fundus synthesis.

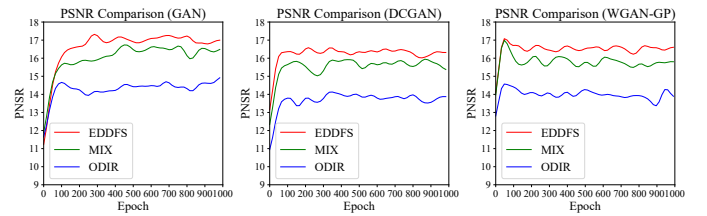


Fig. 8. Comparison among different datasets in terms of loss curves and PSNRs of synthetic images. The red, green and blue curves stand for PSNRs obtained through our EDDFS, MIX and ODIR respectively.

on health samples, it contains less than 600 healthy images, which is about 1/30 of our healthy fundus number.

To prove that our data can better support image synthesis, we generate fundus images from different datasets and evaluate the synthesized images. Existing basic GANs are always not able to generate high-resolution fundus images with all the complex components, hence many works either involve extra module(s) to deal with vessel generation or simply use segmented vessels. For simplification, we focus on basic fundus generation by DCGAN [22], WGAN-GP [23], and StyleGAN [24]. The quality metric is PSNR since the generated 128×128 images are so small that they suffer from a very low “Good” rate. Our future work may involve higher-

resolution image synthesis evaluation.

Images in size of 128×128 are leveraged since abundant details are not necessary for generating basic fundus. The loss curves demonstrated in Fig. 7 indicate that our data is able to offer a faster convergence for training. In addition, the PSNR curves in Fig. 8 reflect that our data is able to support generation with better quality.

The synthesized images are shown in Fig. 6. Although three of the involved GANs fail in generating fundus details like vessels, the generated fundus generally exhibit differences. The generators trained on our dataset preliminarily obtained better appearances than those trained on ODIR, RFMiD and APTOS2019 mixed datasets, e.g., the images trained on ODIR suffer from low exposure. Especially, the generated fundus images based on our EDDFS are more clear in optic disc and cup. Moreover, the StyleGAN [24] successfully generates more clear vessels than other models from the perspective of subjective perception. And, its generated images are with higher visual quality.

VI. CONCLUSION

We collect and annotate a high-quality retinal fundus dataset for disease diagnosis and fundus synthesis. Most samples are 45 degrees and centered on the fovea and optic disc. This dataset contains both disease-level and grade-level annotations with multi-label, and includes some challenging samples with tiny lesions and subtle signs. Based on this dataset, we benchmark several baseline deep models on disease diagnosis and verify its capability of supporting fundus synthesis. In addition, we propose a co-attention structure to extract fundus features using both global and local means, and designed an attention network based on this structure together with a proposed cross-stage feature fusion module. We hope this work to assist and inspire further research on learning-based fundus image analysis and ocular disease diagnosis. Our future work are mainly three folds, presenting finer-grained annotations for “difficult” samples, balancing data through collecting or generating more fundus images with specific diseases, improving the proposed co-attention network.

REFERENCES

- [1] International Agency for the Prevention of Blindness, *Rethinking Global Health Means Including Eye-Health*. IAPB, 2020. [Online]. Available: <https://www.iapb.org/news/rethinking-global-health-means-including-eye-health/>
- [2] H. K. Walker, W. D. Hall, and J. W. Hurst, *Clinical methods: the history, physical, and laboratory examinations, 3rd edition*. Boston: Butterworths, 1990. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK201/>
- [3] M. Kim, J. Zuallart, and W. De Neve, “Few-shot learning using a small-sized dataset of high-resolution fundus images for glaucoma diagnosis,” in *Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care*, 2017, pp. 89–92.
- [4] Y. Zhou, X. He, L. Huang, L. Liu, F. Zhu, S. Cui, and L. Shao, “Collaborative learning of semi-supervised segmentation and classification for medical images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2079–2088.
- [5] Z. Huang, Y. Ding, G. Song, L. Wang, R. Geng, H. He, S. Du, X. Liu, Y. Tian, Y. Liang *et al.*, “Bcdata: A large-scale dataset and benchmark for cell detection and counting,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020, pp. 289–298.
- [6] H. Fu, B. Wang, J. Shen, S. Cui, Y. Xu, J. Liu, and L. Shao, “Evaluation of retinal image quality assessment networks in different color-spaces,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 48–56.
- [7] C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, S. Douma, and A. A. Argyros, “FIRE: fundus image registration dataset,” *Modeling and Artificial Intelligence in Ophthalmology*, vol. 1, pp. 16–28, 2017.
- [8] Y. Zhou, B. Wang, L. Huang, S. Cui, and L. Shao, “A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 818–828, 2021.
- [9] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, “Feedback on a publicly distributed database: the messidor database,” *Image Analysis & Stereology*, vol. 33, no. 3, pp. 231–234, 2014.
- [10] M. D. Abràmoff, J. C. Folk, D. P. Han, J. D. Walker, D. F. Williams, S. R. Russell, P. Massin, B. Cochener, P. Gain, L. Tang *et al.*, “Automated analysis of retinal images for detection of referable diabetic retinopathy,” *JAMA Ophthalmology*, vol. 131, no. 3, pp. 351–357, 2013.
- [11] C. Carvalho, J. Pedrosa, C. Maia, S. Penas, Â. Carneiro, L. Mendonça, A. M. Mendonça, and A. Campilho, “A multi-dataset approach for dme risk detection in eye fundus images,” in *International Conference on Image Analysis and Recognition*, 2020, pp. 285–298.
- [12] N. Li, T. Li, C. Hu, K. Wang, and H. Kang, “A benchmark of ocular disease intelligent recognition: one shot for multi-disease detection,” in *International Symposium on Benchmarking, Measuring and Optimization*, vol. 12614, 2021, pp. 177–193.
- [13] “Peking university international competition on ocular disease intelligent recognition.” 2020. [Online]. Available: <https://odir2019.grand-challenge.org/>
- [14] “Stare database: Structure analysis of the retina.” 2004. [Online]. Available: <http://cecas.clemson.edu/~ahoover/stare/>
- [15] S. Pachade, P. Porwal, D. Thulkar, M. Kokare, G. Deshmukh, V. Sahasrabudhe, L. Giancardo, G. Quellec, and F. Mériaudeau, “Retinal fundus multi-disease image dataset (rfmid): a dataset for multi-disease detection research,” *Data*, vol. 6, no. 2, p. 14, 2021.
- [16] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, D. Elie, P. Massin, Z. Viktor, A. Erginay, B. Laÿ, and A. Chabouis, “Teleophta: Machine learning and image processing methods for teleophthalmology,” *Irbm*, vol. 34, no. 2, pp. 196–203, 2013.
- [17] N. Park and S. Kim, “How do vision transformers work?” in *International Conference on Learning Representations*, to appear, 2022.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.
- [20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
- [21] A. García-Layana, F. Cabrera-López, J. García-Arumí, L. Arias-Barquet, and J. M. Ruiz-Moreno, “Early and intermediate age-related macular degeneration: update and clinical review,” *Clinical Interventions in Aging*, vol. 12, pp. 1579–1587, 2017.
- [22] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015.
- [23] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 5767–5777, 2017.
- [24] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.