Supplementary materials

Appendix 1. Detailed calculation of FID and KID.

FID is an indicator proposed by Martin Heusel et al.[1] that uses the deep neural network Inception to extract feature vectors from the real dataset and the generated dataset and obtain the mean, denoted as μ_1 and μ_2, respectively, using the definition of Fréchet distance to calculate FID:

$$FID = ||\mu_1 - \mu_2||_2^2 + Tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2})$$

where $\|\cdot\|_2$ denotes Euclidean norm, Tr denotes trace operation, Σ_1 and Σ_2 represent covariance matrices representing the real image and the Inception feature representation of the generated image, respectively.

KID is an evaluation metric proposed by Binkowski et al.[2], which is also based on Inception. Unlike FID, KID uses a kernel matrix to measure similarity between the real and generated datasets. We use a polynomial kernel for the squared Maximum Mean Difference estimate, which is unbiased and gives the final KID representation:

$$Kid = MMD_u^2(X,Y) = \frac{1}{m(m-1)}\Sigma_{i\neq j}^m k(x_i,x_j) + \frac{1}{n(n-1)}\Sigma_{i\neq j}^n k(y_i,y_j) - \frac{2}{mn}\Sigma_{i=1}^m\Sigma_{j=1}^n k(x_i,y_j)$$

where $k(x,y) = \left(\frac{1}{d}x^T y + 1\right)^3$, the generated dataset is represented as X, at the same time, the real data set is represented as Y. x and y are the feature vectors from both datasets. m is the sample size of X, and n is the sample size of Y. In general, smaller FID and KID values both correspond to more realistic, diverse, and higher quality generated images.

Table S1. Details about three experiments in clinical evaluation.

Experiment	Question	Answer
(1) Quality	Which answer best describes the quality of this image?	Excellent: Almost no quality defects. All retinopathies clearly visible. Adequate: Notable quality defects. All retinopathies clearly visible. Inadequate: Severe quality defects. Some retinopathies unidentifiable.
(2) Authenticity	Is this image realistic or synthetic?	Realistic: Consistent with characteristics of real images collected in clinical settings. Synthetic: Not consistent with characteristics of real images, unnatural.
(3) Diagnostic efficacy	Which condition is present in this image?	For CFP images: Normal, AMD, DR, PM, and Others. For UWF images: Normal, Glaucoma, Lattice/hole, RD, and Others.
Definition of retinal diseases: Normal: No significant retinopathies. Answers other than “Normal” are all considered as “Retinopathies” in subsequent analysis. AMD: An age-related macular degeneration severity level of “Early AMD” or worse[3]. DR: A diabetic retinopathy severity level of “Moderate DR” or worse[4]. PM: Positive for “Pathologic myopia” according to META-PM classification[5]. Glaucoma: Classified into “Suspect” or “Certain” according to an established classification for glaucomatous optic neuropathy[6 7]. Lattice/hole: Classified into “NPRLs” according to an established classification for notable peripheral retinal lesions[8]. RD: Retinal detachment from all causes.		

CFP, colour fundus photography; UWF, ultra-wide field; AMD, age-related macular degeneration; DR, diabetic retinopathy; PM, pathological myopia; RD, retinal detachment.

Table S2. Baseline clinical characteristics.

	Development set	Test set	P value
Number of subjects	510	200	N/A
Paired images (CFP)	959	100	N/A
Paired images (UWF)	1009	100	N/A
Age (mean±SD, years)	64.5±12.6	64.6±11.5	0.945
Female gender (%)	295 (55.9)	123 (61.5)	0.175
Right eye (%)	274 (53.7)	114 (57)	0.427
Preoperative BCVA (LogMAR, mean±SD)	0.882±0.603	0.969±0.642	0.103
LOCS II classification (mean±SD)			
C	2.53±0.86	2.52±0.91	0.777
N	3.03±0.98	3.00±0.95	0.698
P	2.12±1.02	2.02±0.97	0.196

CFP, colour fundus photography; UWF, ultra-wide field; BCVA, best corrected distance visual acuity; LogMAR, logarithm of the minimum angle of resolution; LOCS II, lens opacities classification system II.

Table S3. Quality evaluation using pre-operative and generated CFP and UWF images in the test dataset.

		Pre-operative images			Generated images			P1	P2	P3
		Excellent	Adequate	Inadequate	Excellent	Adequate	Inadequate			
CFP	Residents	9.3% (6.5%, 13.2%)	51.7% (46.0%, 57.3%)	39.0% (33.7%, 44.6%)	17.0% (13.2%, 21.7%)	60.7% (55.0%, 66.0%)	22.3% (18.0%, 27.4%)	0.008	0.032	<0.001
	Seniors	12.7% (9.4%, 16.9%)	48.7% (43.1%, 54.3%)	38.7% (33.3%, 44.3%)	21.3% (17.1%, 26.3%)	52.3% (46.7%, 57.9%)	26.3% (21.7%, 31.6%)	0.007	0.414	0.002
	Experts	15.7% (12.0%, 20.2%)	66.0% (60.5%, 71.1%)	18.3% (14.4%, 23.1%)	30.7% (25.7%, 36.1%)	61.7% (56.1%, 67.0%)	7.7% (5.2%, 11.2%)	<0.001	0.308	<0.001
UWF	Residents	14.3% (10.8%, 18.8%)	62.0% (56.4%, 67.3%)	23.7% (19.2%, 28.8%)	24.0% (19.5%, 29.1%)	56.7% (51.0%, 62.2%)	19.3% (15.3%, 24.2%)	0.004	0.213	0.233
	Seniors	25.0% (20.4%, 30.2%)	55.3% (49.7%, 60.9%)	19.7% (15.6%, 24.5%)	32.7% (27.6%, 38.2%)	52.7% (47.0%, 58.3%)	14.7% (11.1%, 19.1%)	0.047	0.566	0.130
	Experts	36.0% (30.8%, 41.6%)	51.3% (45.7%, 56.9%)	12.7% (9.4%, 16.9%)	42.0% (36.6%, 47.7%)	48.0% (42.4%, 53.6%)	10.0% (7.1%, 13.9%)	0.155	0.462	0.367

CFP, colour fundus photography; UWF, ultra-wide field. P1 indicates the p value calculated between the proportion of excellent images in pre-operative and generated images using McNemar's test. P2 indicates the p value calculated between the proportion of adequate images in pre-operative and generated images using McNemar's test. P3 indicates the p value calculated between the proportion of inadequate images in pre-operative and generated images using McNemar's test.

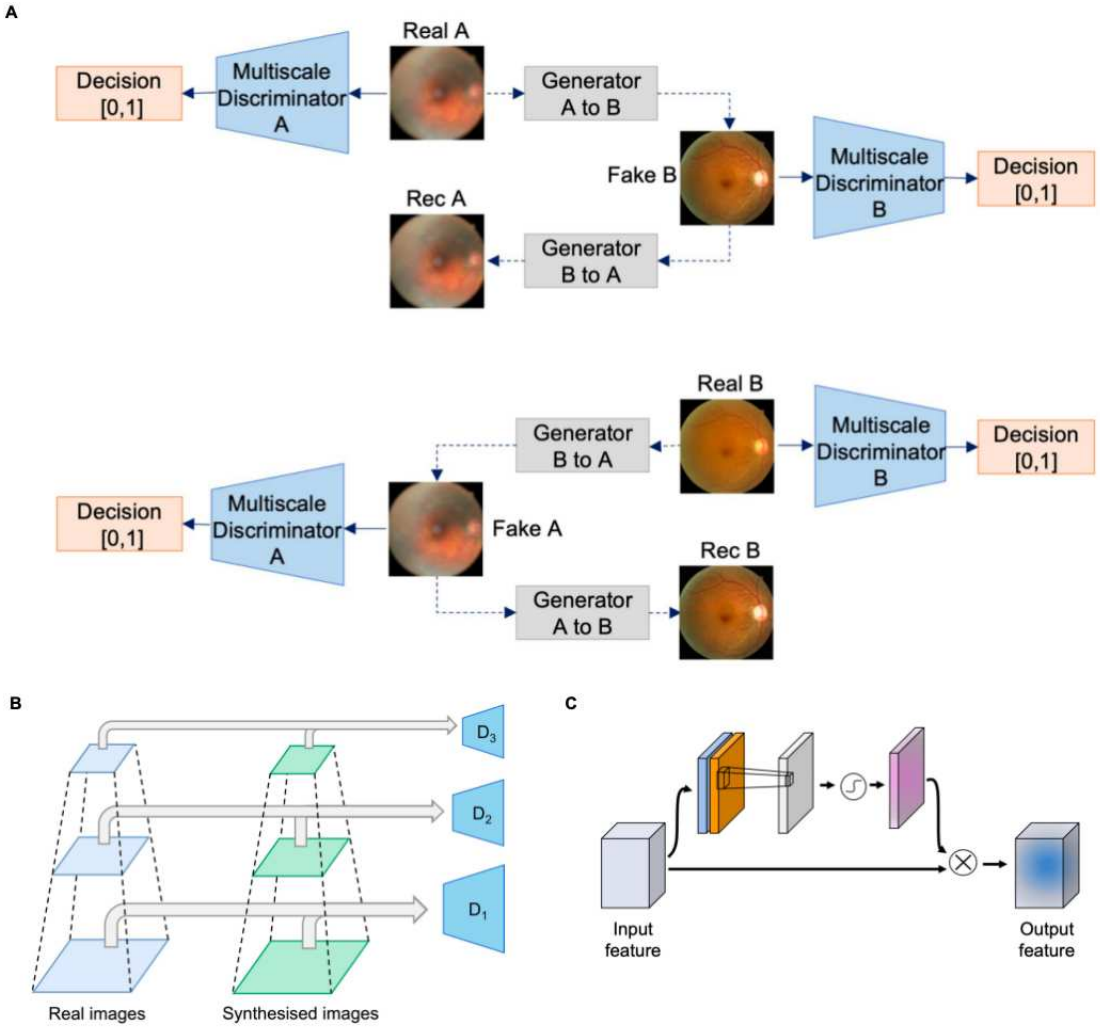

Figure S1. A conceptual illustration of the CycleGAN structure and modification.

This figure demonstrated the overall structure of CycleGAN(A), the multiscale discriminators (D_1, D_2 and D_3) (B) as well as the spatial attention mechanism (C) used in C^{2}ycleGAN. Our objective is to learn a mapping for transforming images from a source domain A, characterized by cloudy retinal images caused by cataracts, into a target domain B, where the post-operative retinal images are sharp and clear. In general, CycleGAN algorithm is designed to iteratively refine image generation networks through a competitive process. Alongside the target generating networks, discriminative networks are also trained during the process. The generating networks aim to produce post-operative images that closely resemble real ones, while the discriminative networks work to differentiate between

generated and real images. These two sets of networks continuously update themselves until the training is completed, resulting in generating networks capable of producing high-quality images.

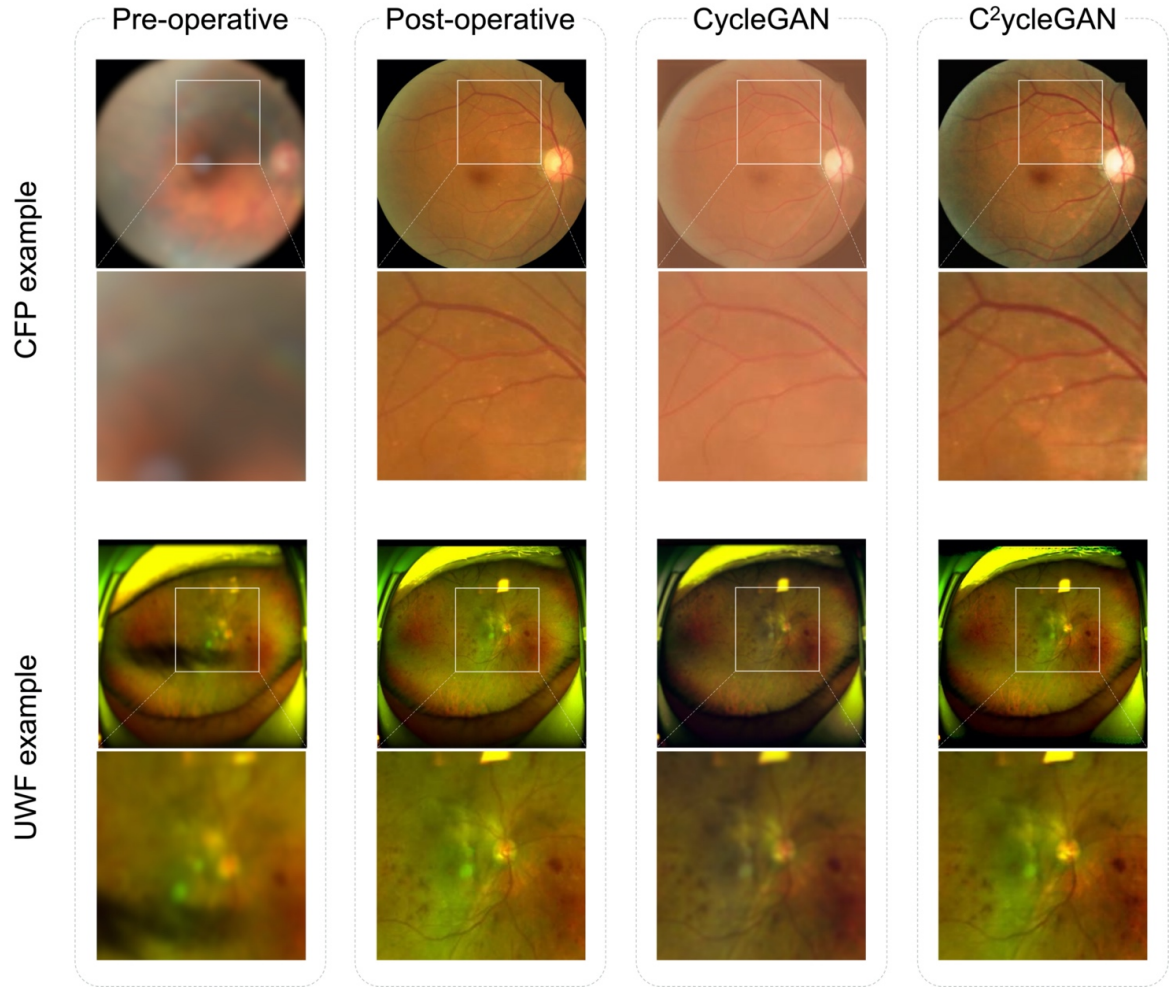

Figure S2. Examples of pre-operative and post-operative fundus image pairs and corresponding generated images using CycleGAN and C^{2}ycleGAN.

CFP, colour fundus photography; UWF, ultra-wide field.

REFERENCES

1. Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 2017;30
2. Bińkowski M, Sutherland DJ, Arbel M, et al. Demystifying mmd gans. *arXiv preprint arXiv:180101401* 2018
3. Klein R, Davis MD, Magli YL, et al. The Wisconsin age-related maculopathy grading system. *Ophthalmology* 1991;98(7):1128-34. doi: 10.1016/s0161-6420(91)32186-9 [published Online First: 1991/07/01]
4. Wilkinson CP, Ferris FL, 3rd, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology* 2003;110(9):1677-82. doi: 10.1016/S0161-6420(03)00475-5 [published Online First: 2003/09/18]
5. Ohno-Matsui K, Kawasaki R, Jonas JB, et al. International photographic classification and grading system for myopic maculopathy. *Am J Ophthalmol* 2015;159(5):877-83 e7. doi: 10.1016/j.ajo.2015.01.022 [published Online First: 2015/01/31]
6. Li Z, He Y, Keel S, et al. Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. *Ophthalmology* 2018;125(8):1199-206. doi: 10.1016/j.ophtha.2018.01.023 [published Online First: 2018/03/07]
7. Li Z, Guo C, Lin D, et al. Deep learning for automated glaucomatous optic neuropathy detection from ultra-widefield fundus images. *Br J Ophthalmol* 2021;105(11):1548-54. doi: 10.1136/bjophthalmol-2020-317327 [published Online First: 2020/09/18]
8. Li Z, Guo C, Nie D, et al. A deep learning system for identifying lattice degeneration and retinal breaks using ultra-widefield fundus images. *Ann Transl Med* 2019;7(22):618. doi: 10.21037/atm.2019.11.28 [published Online First: 2020/01/14]