

Quantifying the Effect of Synthetic Colour Fundus Photograph Scaling on Foundation Models

Bowen Xu

Chinese University of Hong Kong

Hongyang Jiang

Chinese University of Hong Kong

Anran Ran

Chinese University of Hong Kong

Paul Nderitu

NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust

Ke Zou

National University of Singapore

Yukun Zhou

NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust

Gabriel Dawei Yang

Singapore National Eye Centre

Jocelyn Hui Lin Goh

National University of Singapore

Sahana Srinivasan

National University of Singapore

Pearse A. Keane

NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust

Yih Chung Tham

National University of Singapore

Carol Y. Cheung

`carolcheung@cuhk.edu.hk`

Chinese University of Hong Kong

Article

Keywords:

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10970813/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. C.Y.C. is a cofounder and shareholder of i-Cognitio Sciences, which commercialises retinal imaging-based algorithms for brain-health assessment. C.Y.C. and A.R.R. are co-founders of Inno Eye Focus Ltd. P.A.K. is a cofounder of Cascader; has acted as a consultant for Retina Consultants of America, Roche, Boehringer Ingelheim and Bitfount; and holds equity in Big Picture Medical. P.A.K. has received speaker fees from Zeiss, Thea, Apellis and Roche, travel support from Bayer and Roche, and has served on advisory boards for Topcon, Bayer, Boehringer Ingelheim and Roche. P.N. has received speaker fees from Boehringer Ingelheim, Roche and Alimera. All other authors declare no competing interests.

Abstract

Scaling synthetic colour fundus photographs (CFPs) offers a tunable strategy for retinal foundation-model pretraining. From 28,430 real CFPs, we generated 568,600 synthetic CFPs and compared six one-pass strategies: real-only, synthetic-only, real-plus-1×synthetic, real-plus-5×synthetic, real-plus-10×synthetic and real-plus-20×synthetic. Average area under the receiver operating characteristic curve (AUROC) was highest in internal validation with real-plus-10×synthetic (0.906 versus 0.862 for real-only across 12 test datasets), but in external validation with real-plus-20×synthetic (0.777 versus 0.683 across eight datasets). The real-plus-1×synthetic strategy increased both averages with the lowest counted computation among real-plus-synthetic strategies. At a matched image count, synthetic-only achieved comparable average external AUROC to real-only. Generation accounted for approximately 99% of the estimated 9×10^{19} floating-point operations for generation and backbone pretraining. These findings identify complementary uses of synthetic scaling: lower-volume supplementation for a performance–computation balance and larger volumes for stronger external generalisation, with strategy selection guided by evaluation objectives and computational resources.

Introduction

Colour fundus photographs (CFPs) provide a non-invasive view of the retinal vasculature and optic nerve head, enabling task-specific deep-learning models to detect eye diseases, including glaucoma, diabetic retinopathy (DR), and age-related macular degeneration (AMD)^{1–3} and support systemic disease prediction (oculomics).^{4–7} Most deep-learning models trained on CFPs, however, use labelled data from specific development populations, which limits their generalisability across populations and settings. Foundation models (FMs), by learning transferable representations from large unlabelled collections, can be adapted to multiple downstream tasks using relatively little labelled data.⁸ For example, RETFound⁹ and VisionFM¹⁰ have demonstrated label-efficient adaptation across ocular and systemic prediction tasks, while EyeFM¹¹ extends the FM paradigm to multimodal clinical assistance. This versatility, however, shifts the data burden to pretraining, which requires large, diverse, and carefully curated collections of unlabelled CFPs and other ophthalmic images that are costly to assemble.^{12–14}

Synthetic image generation offers one way to expand pretraining corpora without acquiring additional real images, using approaches such as generative adversarial networks¹⁵ and diffusion models.^{16–17} Trained on real CFP datasets, these models can generate CFPs conditioned on available demographic, acquisition, and disease metadata. Synthetic medical images have augmented original datasets^{18–20} and increased minority-class representation in prognostic models.²¹ In ophthalmic FM pretraining, synthetic images have reduced the amount of real data required for pretraining.²² Although previous studies have compared more than one real-to-synthetic ratio,^{10,18,21,23} evidence remains limited on whether adding progressively larger synthetic volumes to a fixed real CFP corpus yields consistent downstream gains, particularly in external validation, in which models are tested on independent datasets not used to fine-tune the evaluated model. Because both image generation and pretraining over

larger datasets require computation, the utility of scale cannot be judged from performance alone. A scaling analysis must therefore examine downstream performance alongside floating-point operations (FLOPs) for generation and pretraining to determine whether gains persist on external datasets.

In this study, using a fixed real CFP corpus as the reference, we compared real-only pretraining, synthetic-only pretraining with the same number of images, and real-plus-synthetic pretraining with synthetic-to-real image ratios of 1:1, 5:1, 10:1, and 20:1. We evaluated the resulting backbones across multiple classification tasks using internal validation on held-out test sets and external validation. For each strategy, we also related the observed unweighted mean area under the receiver operating characteristic curve (AUROC) to the total FLOPs required for synthetic-image generation and one DINOv2 pretraining run, separately for internal and external validation.

Results

Study overview

Figure 1 summarises the study design; source and downstream datasets are described in Tables 1 and 2, respectively. Implementation settings are summarised in Supplementary Table S1. Figure 2 shows representative real CFPs from each of the three source datasets used to train low-rank adaptation (LoRA) adapters alongside synthetic CFPs produced by the corresponding adapter.

Internal validation across synthetic volumes

Across the 12 downstream datasets, mean AUROC increased by $\Delta = +0.043$ from real-only to real-plus-10×synthetic, reaching the highest observed internal mean (**Figure 3**). Generally, the largest gains were found between adjacent strategies from real-only to real-plus-1×synthetic and from real-plus-5×synthetic to real-plus-10×synthetic. Only a subtle change in mean AUROC was observed between the 1× and 5× strategies. Across the 12 internal datasets, no synthetic-only versus real-only comparison met the Benjamini–Hochberg (BH)-adjusted significance criterion of $q < 0.05$. Nevertheless, the magnitude of improvement varied by dataset. For example, APTOS and AMDNet23, which had high real-only AUROCs, showed comparatively small gains. Per-dataset AUROC estimates are in **Table S2**. Threshold-dependent metrics (sensitivity, specificity, precision, negative predictive value, and F1 score) at the locked operating threshold are reported in **Figure S1** (Supplementary Section S1).

The Friedman test across the 12 downstream datasets and six scaling strategies yielded $\chi^2(5) = 44.05$, $p < 0.001$, indicating that strategy ranks were not all equal. The Nemenyi post hoc test had a critical difference of 2.18 at $\alpha = 0.05$ ($k = 6$, $N = 12$; **Figure 4**). The mean-rank differences between real-only and real-plus-10×synthetic and between real-only and real-plus-20×synthetic were 3.00 and 3.58, respectively, and therefore exceeded the critical difference.

Of the 60 pairwise comparisons against real-only (12 datasets $\times$ 5 non-baseline strategies), 10 met the within-dataset BH criterion ($q < 0.05$); no real-plus-synthetic strategy was significantly worse than real-only. The 10 significant improvements occurred in five datasets: DeepDRiD, AITS, HRDC, ORIGA, and HYAMD. All comparisons that remained significant after BH correction are reported in **Table S3**.

Area under the precision–recall curve (AUPRC) showed a similar overall scaling pattern: mean AUPRC across the 12 downstream datasets increased by $\Delta = +0.062$ from real-only to real-plus-20 $\times$ synthetic, which had the highest observed mean AUPRC. The Friedman test on AUPRC ranks yielded $\chi^2(5) = 37.90$, $p < 0.001$, indicating that strategy ranks for AUPRC were not all equal. Real-plus-20 $\times$ synthetic also had the highest internal mean AUPRC, marginally exceeding real-plus-10 $\times$ synthetic, which had the highest mean AUROC. Per-dataset AUPRC values are tabulated in **Table S4** with the corresponding synthetic CFP scaling curve in **Figure S2**.

External validation across synthetic volumes

Across the eight same-disease external validations, mean AUROC increased by $\Delta = +0.093$ from real-only to real-plus-20 $\times$ synthetic, the highest observed external mean, compared with a maximum internal gain of +0.043 (**Figure 5**). Mean external AUROC was 0.012 higher with synthetic-only than with real-only pretraining, but results varied across external validation datasets. Of the 40 non-baseline-strategy comparisons, 22 showed significantly higher AUROC than real-only after applying the BH criterion within each external validation at $q < 0.05$ (18 of 32 real-plus-synthetic comparisons and 4 of 8 synthetic-only comparisons). One additional synthetic-only comparison, AMDNet23 $\rightarrow$ HYAMD, showed significantly lower AUROC; no real-plus-synthetic strategy showed a significant decrease. AUROC estimates for each external validation are in **Table S5**; AUPRC values are reported in **Table S6** and visualised in **Figure S3**. In the selected true-positive cases shown for the eight external validations, the attention overlays were spatially localised and sometimes overlapped the optic disc in glaucoma, visible retinal lesions in DR, or the macular region in AMD. The heatmaps for each external validation are shown in **Figure S4** (Supplementary Section S7).

Observed AUROC versus counted FLOPs

Generating one synthetic CFP was estimated to require 156.45 trillion FLOPs, whereas one image presentation during the specified single-pass DINOv2 pretraining protocol using a base-size Vision Transformer (ViT) with 14-pixel patches (ViT-B/14) was estimated to require 601 billion FLOPs. Across these two counted stages, generating the shared synthetic pool once and pretraining all six backbones was estimated to require approximately 90 EFLOPs (1 EFLOP = 10^{18} floating-point operations), of which approximately 99% arose from synthetic-image generation.

Figure 6 shows the empirical Pareto frontiers for mean AUROC versus counted FLOPs. Using the strategy-specific FLOP accounting and Equation 1 in Methods, the internal empirical frontier comprised

real-only (0.017 EFLOPs, 0.862), real-plus-1×synthetic (4.482 EFLOPs, 0.881), real-plus-5×synthetic (22.342 EFLOPs, 0.885), and real-plus-10×synthetic (44.667 EFLOPs, 0.906). Synthetic-only and real-plus-20×synthetic were empirically dominated in the internal summary. The external empirical frontier comprised real-only, synthetic-only, real-plus-1×synthetic, real-plus-5×synthetic, and real-plus-20×synthetic; real-plus-10×synthetic was empirically dominated in the external summary by real-plus-5×synthetic.

Under the study-defined rule, real-plus-1×synthetic was designated because it had the lowest counted FLOPs among real-plus-synthetic strategies, lay on both empirical frontiers, and exceeded real-only in both observed mean AUROCs. Its observed internal AUROC gain was $\Delta = +0.018$, 42% of the maximum observed internal gain of +0.043, while using approximately 10% of the counted FLOPs of real-plus-10×synthetic. Its observed external AUROC gain was $\Delta = +0.051$, 55% of the maximum observed external gain of +0.093, while using approximately 5% of the counted FLOPs of real-plus-20×synthetic. This designation did not imply statistical superiority or a unique optimum.

At higher volumes, the internal and external summaries diverged. Real-plus-20×synthetic accounted for 89.316 EFLOPs, approximately twice the 44.667 EFLOPs of real-plus-10×synthetic and produced no higher observed mean internal AUROC (0.905 versus 0.906); however, it achieved the highest observed mean external AUROC (0.777 versus 0.683 for real-only). **Table 3** summarises the strategy-specific values. The corresponding Pareto analysis using mean AUPRC versus counted FLOPs, reported in **Figure S5**, designated the same strategy in both panels.

Discussion

Real-plus-synthetic pretraining achieved higher mean downstream AUROC than real-only pretraining, with the largest gains observed in external validation. Internal and external validation showed different performance patterns as synthetic volume increased. Selecting this scale according to the intended downstream tasks and available computation may provide a practical strategy for FM development.

Prior work has shown that synthetic medical images can augment classification datasets^{18–20}, provide training data for downstream classifiers²⁴, increase representation of minority outcomes in prognostic modelling²¹, and expand FM pretraining corpora.²² Four studies have compared multiple real-to-synthetic ratios.^{10,18,21,23} Across the evaluated range, lower synthetic-data volumes delivered a substantial share of the performance improvement with modest corpus expansion, whereas higher-volume strategies achieved the strongest external validation performance. This contrast suggests two useful operating regimes: efficient supplementation at lower volumes and scaling for external generalisation at higher volumes.

Moreover, we assessed performance–compute efficiency and consistency by the Pareto and rank analyses. We found that under the study decision rule, real-plus-1×synthetic provided a promising starting point for synthetic supplementation among the evaluated real-plus-synthetic strategies. It

achieved higher mean AUROCs than real-only pretraining in both evaluation settings, used the lowest counted FLOPs among these strategies, and lay on both empirical frontiers.

We also found that at a matched image count, synthetic-only pretraining achieved comparable mean external AUROC to real-only pretraining, supporting the use of synthetic CFPs as a potential standalone pretraining resource in data-constrained settings. The treatment of retinal imaging as biometric data remains debated.^{25–26} Synthetic-only pretraining offers a potential route to privacy-conscious foundation-model development by reducing the need to share original patient CFPs for backbone training. Nebbia et al. showed that retinal foundation-model features can support patient re-identification.²⁷ Our generator was conditioned on disease labels and demographic and acquisition metadata, and the resulting synthetic images may support learning shared disease-relevant features with less reliance on patient-specific retinal cues. Our findings support the disease-prediction utility of this approach; its potential to reduce re-identification risk requires further evaluation.

A principal strength of this study was the consistent evaluation of synthetic CFP scaling under a one-pass training protocol. All strategies used the same architecture, random initialisation, pretraining method, and evaluation pipeline. This comparison measured the practical consequences of increasing synthetic-data volume, including the additional training updates and computation associated with larger corpora. Isolating the contribution of synthetic-image content at matched training updates represents a complementary experimental question beyond the scope of this study.

The findings apply to the tested generation and pretraining pipeline, binary CFP classification tasks, and populations and imaging devices represented in the source datasets. The synthetic-only comparison at a matched image count provides a useful benchmark; future matched-computation comparisons with additional real data or conventional augmentation could build on these findings. Further studies should assess generated-image quality, diversity and generator-specific bias, and evaluate the comparative performance and clinical utility of synthetic-only pretraining. The bootstrap confidence intervals provide a descriptive summary of the eight external validations using three source-trained models, with unequal disease representation. Extending evaluation to biomarker extraction and segmentation would broaden the assessment of synthetic-data pretraining beyond classification.

In summary, our findings support synthetic CFPs as a valuable resource for retinal FM pretraining, either as an extension of real-image corpora or as a potential standalone resource when access to real CFPs is limited. Synthetic-data scale should be aligned with the intended evaluation objective and available computation, with external validation used to guide and confirm strategy selection.

Methods

Study design

Figure 1 summarises the study design. For each source dataset, we trained a separate image-generation adapter and used it to generate synthetic CFPs. We then pretrained a separate ViT backbone for each pretraining strategy, fine-tuned each backbone on each binary classification dataset, and evaluated the resulting models on held-out images from the same dataset (internal validation) and on independent datasets for the same disease (external validation).

Synthetic CFP generation

We adapted Stable Diffusion XL (SDXL) 1.0,²⁸ a latent-diffusion text-to-image model with a U-Net backbone,¹⁷ using LoRA for CFP generation.^{29,30} For each source CFP dataset, we trained only a LoRA adapter in the U-Net self-attention and cross-attention projections; the SDXL text encoders and all base-model weights remained frozen (Supplementary Table S1).

From the public ODIR,³¹ BRSET,³² and mBRSET³³ datasets, we included 7,000, 16,266, and 5,164 real CFPs, respectively (28,430 total; Table 1). For image generation, we constructed a prompt from the available metadata for each source image, including camera type, demographic characteristics (age, ethnicity, and sex), ophthalmic disease (DR, AMD, and glaucoma), and systemic disease (hypertension and diabetes), using the following template:

"A fundus image collected using {camera} from a {age} year old {ethnicity} {sex} with {DR}, {AMD}, {Glaucoma/Hypertension/Diabetes}."

Each field was populated from the corresponding source-image metadata, and fields with missing data were omitted. Each prompt therefore encoded the available camera and demographic characteristics and recorded disease labels. Demographic metadata were included because they were available in the source datasets and may influence image appearance; their effect on generated-image diversity was not independently evaluated. We sampled prompts to match the distribution of available metadata and recorded labels in each source dataset. Using the three adapters, we generated a shared pool of 568,600 synthetic CFPs at a resolution of 512×512 pixels. Generation used 50 Euler scheduler steps and a classifier-free guidance scale of 5.0³⁴. Each image was generated with a distinct random seed.

Pretraining scaling strategies and backbone

We generated up to 568,600 synthetic CFPs (20× the real-set size) and, by sampling from this shared pool, compared six scaling strategies: real-only (28,430 real CFPs), synthetic-only (28,430 synthetic CFPs), and four real-plus-synthetic strategies combining the real set with 1×, 5×, 10×, or 20× the real-set size of synthetic CFPs. Each strategy was used to pretrain a separate backbone for one complete pass over its dataset under otherwise identical hyperparameters.

Each backbone used the ViT-B/14 architecture³⁵ and was trained from random initialisation with DINOv2³⁶, a self-distillation framework that learns without labels. Because our pretraining corpora ranged from tens to hundreds of thousands of images, we selected DINOv2, which is better suited to this data scale than DINOv3, a model developed using a curated corpus of 1.689 billion images.³⁷ Additional generation, preprocessing, pretraining, and fine-tuning settings are summarised in Supplementary Table S1. Training from random initialisation prevented inherited natural-image weights from influencing the backbone comparison.

Downstream tasks and external validation

Each pretrained backbone was fine-tuned on 12 binary classification tasks spanning glaucoma, DR, AMD, and systemic hypertension (Table 2). For glaucoma detection tasks, the Bangladesh Eye Hospital (BEH) dataset,³⁸ REFUGE,³⁹ ORIGA,⁴⁰ and AITS were used. Reference standards differed among the glaucoma datasets. BEH images were labelled for glaucoma by two eye specialists. REFUGE labels were assigned from comprehensive clinical-record review requiring glaucomatous optic nerve head damage and reproducible visual field defects. ORIGA images originated from the Singapore Malay Eye Study and were released with expert glaucoma labels and optic disc/cup annotations. AITS was an internal prospective ocular-imaging cohort at the Chinese University of Hong Kong Eye Centre, with clinical glaucoma diagnoses labelled by eye specialists. The Research Ethics Committee (Kowloon Central / Kowloon East) approved the AITS cohort study (KC/KE-20-0241/ER-1), and participants provided written informed consent. All four datasets were analysed as binary glaucoma versus non-glaucoma tasks (Table 2). For DR detection, APTOS,⁴¹ IDRiD,⁴² MESSIDOR-2,^{43–44} and DeepDRiD datasets were used.⁴⁵ All four datasets were graded on the five-category International Clinical Diabetic Retinopathy severity scale and, for this study, binarised as any DR (grades 1–4) versus no DR (grade 0). For AMD detection, AMDNet23,⁴⁶ HYAMD,⁴⁷ and ADAM⁴⁸ were used, all binarised as present versus absent: AMDNet23 was binarised from four source classes into AMD versus non-AMD; HYAMD used classes 1 and 2 (early and intermediate-to-late AMD) as positive and class 0 (non-AMD) as negative; and ADAM used its native binary AMD-versus-non-AMD label (Table 2). HRDC⁴⁹ was used to predict systemic hypertension from CFPs. For each dataset, we used a fixed, stratified image-level split of 50% for training, 20% for validation, and 30% for testing; we used identical image splits for all pretraining strategies. We fine-tuned all model parameters for 50 epochs using AdamW⁵⁰ at a constant learning rate with focal loss⁵¹ ($\gamma = 2.0$; $\alpha = 1.0$) to address class imbalance, and selected the checkpoint with the highest exponentially smoothed validation AUROC. Full hyperparameter settings are given in Supplementary Table S1. To visualise the spatial distribution of attention for each fine-tuned backbone, we extracted the class-token ([CLS]) self-attention weights from the last transformer block, averaged them across heads, reshaped the CLS-to-patch attention to the patch grid, resized it to the input image dimensions, and overlaid the resulting map on the input CFP.

To assess external generalisation, we fine-tuned each backbone on the largest dataset within each disease group as the source and evaluated it on all images in each remaining same-disease dataset as an independent target. The eight external validations were REFUGE→AITS, REFUGE→BEH, and REFUGE→ORIGA for glaucoma; APTOS→IDRiD, APTOS→MESSIDOR-2, and APTOS→DeepDRiD for DR; and AMDNet23→ADAM and AMDNet23→HYAMD for AMD.

Statistical analysis

Model discriminative performance was evaluated separately in each internal held-out test cohort and across all images in each external target dataset. AUROC was prespecified as the primary metric, and AUPRC as a secondary metric. For each downstream task and pretraining strategy, we drew 1,000 bootstrap samples from the source-dataset validation split, selected the threshold that maximised Youden's J statistic in each sample, averaged the selected thresholds, and locked that mean threshold before evaluation. The threshold was applied without recalibration to the corresponding internal held-out test cohort or, for external validation, to all images in the target dataset. Sensitivity, specificity, precision (positive predictive value), negative predictive value, and F1 score were calculated at this threshold.

Overall differences among the six strategies were assessed across the 12 internal downstream datasets using a Friedman test,⁵² with dataset as the block and strategy as the within-block factor; Nemenyi post hoc comparisons of mean ranks were applied to the primary AUROC analysis.⁵³ Within each evaluation unit, each of the five alternative strategies was compared with real-only using a two-sided DeLong test⁵⁴ on paired predictions from the same evaluation images—the held-out test partition for internal validation or all images in the target dataset for external validation—and the five resulting p values were adjusted within that unit using the BH procedure, with $q < 0.05$ used as the significance criterion.⁵⁵ The evaluation unit was the individual dataset for internal comparisons and the source–target dataset pair for external comparisons. AUROCs for each internal or external validation and paired AUROC differences (Δ AUROC, defined as the evaluated strategy's AUROC minus the real-only AUROC) were reported with two-sided 95% confidence intervals from DeLong variance estimates.⁵⁴ Confidence intervals for the unweighted mean AUROC and AUPRC across internal or external validations were estimated with 1,000 paired bootstrap resamples of the evaluation units. These intervals describe evaluation uncertainty for the fitted models; repeated pretraining and fine-tuning runs would additionally characterise training variability. All differences and across-dataset summaries were calculated from full-precision estimates before rounding. Benjamini–Hochberg adjustment therefore controlled the five comparisons within each internal or external validation, not the false-discovery rate across all internal or external validations; aggregate counts of significant comparisons are descriptive.

FLOP estimation and empirical Pareto analysis

To quantify computational requirements, we used PyTorch FlopCounterMode to estimate the FLOPs required for synthetic-image generation and DINOv2 backbone pretraining. Counts were reported in EFLOPs; only these two stages were included.

We compared observed unweighted mean AUROC with counted FLOPs separately for internal and external validation. For each strategy s , $C(s)$ denoted the combined FLOPs required to generate its synthetic images and pretrain one backbone, and $A(s)$ denoted its unweighted mean AUROC. Mean AUROC was calculated across internal test datasets for the internal analysis and across external validations for the external analysis. A strategy lay on the empirical Pareto frontier if no other strategy Pareto-dominated it ($s' \succ s$, requiring $C(s') \leq C(s)$ and $A(s') \geq A(s)$ with at least one strict inequality). With S denoting the set of evaluated strategies, the empirical Pareto frontier P was:

$$P = \{s \in S : \{s' \in S : s' \succ s, s \neq s'\} = \emptyset\}$$

1

To provide a transparent descriptive designation, we defined the selected strategy as the lowest-FLOP real-plus-synthetic strategy that was nondominated in both separately calculated AUROC summaries and whose observed internal and external mean AUROCs both exceeded those of real-only. As an exploratory secondary analysis, we repeated the Pareto analysis using mean AUPRC in place of mean AUROC. The results are reported in Supplementary Section S8.

Declarations

Data availability

A 1× release of the synthetic CFPs is publicly available in Figshare at <https://doi.org/10.6084/m9.figshare.33452863.v1>. ODIR, BRSET, and mBRSET are available from their original repositories. AITS is an internal clinical cohort and is not publicly available.

Code availability

Code for metadata-to-prompt construction, synthetic CFP generation, image preprocessing, study-protocol specification, and aggregate-result figure reconstruction is publicly available at <https://github.com/PolskieStronk/synthetic-fundus-scaling>. The three dataset-specific SDXL LoRA adapters are publicly available in Figshare at <https://doi.org/10.6084/m9.figshare.33452371.v1>.

Acknowledgements

This work was supported by the Health and Medical Research Fund, Hong Kong SAR (ref. 19201991) and The Chinese University of Hong Kong Direct Grant (ref. 2026.327).

Author contributions

B.X. and H.J. developed the methodology, conducted the experiments and analysed the data. A.R.R., P.N., K.Z. and Y.Z. contributed to study design and interpretation. G.D.Y., J.H.L.G. and S.S. contributed to interpretation of the findings. B.X. drafted the manuscript. All authors critically revised the manuscript and approved the final version. B.X., H.J., A.R.R., P.N., K.Z. and Y.Z. contributed equally to this work. P.A.K., Y.C.T. and C.Y.C. jointly supervised the study.

Competing interests

C.Y.C. is a cofounder and shareholder of i-Cognitio Sciences, which commercialises retinal imaging-based algorithms for brain-health assessment. C.Y.C. and A.R.R. are co-founders of Inno Eye Focus Ltd. P.A.K. is a cofounder of Cascader; has acted as a consultant for Retina Consultants of America, Roche, Boehringer Ingelheim and Bitfount; and holds equity in Big Picture Medical. P.A.K. has received speaker fees from Zeiss, Thea, Apellis and Roche, travel support from Bayer and Roche, and has served on advisory boards for Topcon, Bayer, Boehringer Ingelheim and Roche. P.N. has received speaker fees from Boehringer Ingelheim, Roche and Alimera. All other authors declare no competing interests.

References

1. Li, Z. et al. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology* 125, 1199–1206 (2018). <https://doi.org/10.1016/j.ophtha.2018.01.023>
2. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316, 2402–2410 (2016). <https://doi.org/10.1001/jama.2016.17216>
3. Peng, Y. et al. DeepSeeNet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from colour fundus photographs. *Ophthalmology* 126, 565–575 (2019). <https://doi.org/10.1016/j.ophtha.2018.11.015>
4. Zhu, Z. et al. Oculomics: current concepts and evidence. *Prog Retin Eye Res* 106, 101350 (2025). <https://doi.org/10.1016/j.preteyeres.2025.101350>
5. Cheung, C. Y. et al. A deep learning model for detection of Alzheimer’s disease based on retinal photographs: a retrospective, multicentre case-control study. *Lancet Digit Health* 4, e806–e815 (2022). [https://doi.org/10.1016/S2589-7500\(22\)00169-8](https://doi.org/10.1016/S2589-7500(22)00169-8)
6. Meng, Z. et al. Non-invasive biopsy diagnosis of diabetic kidney disease via deep learning applied to retinal images: a population-based study. *Lancet Digit Health* 7, 100868 (2025). <https://doi.org/10.1016/j.landig.2025.02.008>
7. Jiang, N. et al. A deep learning system for detecting silent brain infarction and predicting stroke risk. *Nat Biomed Eng* 9, 1907–1919 (2025). <https://doi.org/10.1038/s41551-025-01413-9>
8. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* 616, 259–265 (2023). <https://doi.org/10.1038/s41586-023-05881-4>

9. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* 622, 156–163 (2023). <https://doi.org/10.1038/s41586-023-06555-x>
10. Qiu, J. et al. Development and Validation of a Multimodal Multitask Vision Foundation Model for Generalist Ophthalmic Artificial Intelligence. *NEJM AI* 1(12), Aloa2300221 (2024). <https://doi.org/10.1056/Aloa2300221>
11. Wu, Y. et al. An eyecare foundation model for clinical assistance: a randomized controlled trial. *Nat Med* 31, 3404–3413 (2025). <https://doi.org/10.1038/s41591-025-03900-7>
12. Zhou, Y. et al. Understanding pre-training data effects in retinal foundation models using two large fundus cohorts. *Nat Commun* 17, 3309 (2026). <https://doi.org/10.1038/s41467-026-70077-z>
13. Willeminck, M. J. et al. Preparing medical imaging data for machine learning. *Radiology* 295, 4–15 (2020). <https://doi.org/10.1148/radiol.2020192224>
14. Tham, Y. C. et al. Building the world's first truly global medical foundation model. *Nat Med* 31, 3580–3585 (2025). <https://doi.org/10.1038/s41591-025-03859-5>
15. Goodfellow, I. J. et al. Generative adversarial nets. *Adv Neural Inf Process Syst* 27, 2672–2680 (2014). <https://doi.org/10.48550/arXiv.1406.2661>
16. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst* 33, 6840–6851 (2020). <https://doi.org/10.48550/arXiv.2006.11239>
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. *Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR)*, 10684–10695 (2022). <https://doi.org/10.1109/CVPR52688.2022.01042>
18. Sagers, L. W. et al. Augmenting medical image classifiers with synthetic data from latent diffusion models. *arXiv:2308.12453* (2023). <https://doi.org/10.48550/arXiv.2308.12453>
19. Wang, J. et al. Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nat Med* 31, 609–617 (2025). <https://doi.org/10.1038/s41591-024-03359-y>
20. Chen, R. et al. Boosting foundation models for rare eye disease diagnosis via a multimodal text-to-image generative framework. *npj Digit Med* 9, 371 (2026). <https://doi.org/10.1038/s41746-026-02560-2>
21. Nderitu, P. et al. Augmenting a prognostic deep learning system for referable diabetic retinopathy and maculopathy with synthetic retinal images. *Commun Med* 6, 57 (2026). <https://doi.org/10.1038/s43856-025-01316-5>
22. Sun, Y. et al. A data-efficient strategy for building high-performing medical foundation models. *Nat Biomed Eng* 9, 539–551 (2025). <https://doi.org/10.1038/s41551-025-01365-0>
23. Lin, Y. et al. Improving generalization of polyp detection via conditional StyleGAN augmented training. *npj Digit Med* 9, 113 (2026). <https://doi.org/10.1038/s41746-025-02293-8>
24. Bluethgen, C. et al. A vision-language foundation model for the generation of realistic chest X-ray images. *Nat Biomed Eng* 9, 494–506 (2025). <https://doi.org/10.1038/s41551-024-01246-y>

25. McCallister, E., Grance, T. & Scarfone, K. A. Guide to protecting the confidentiality of personally identifiable information (PII). NIST Special Publication 800 – 122. National Institute of Standards and Technology (2010). <https://doi.org/10.6028/NIST.SP.800-122>
26. American Academy of Ophthalmology Board of Trustees. Special commentary: balancing benefits and risks: the case for retinal images to be considered as nonprotected health information for research purposes. *Ophthalmology* 132, 115–118 (2025). <https://doi.org/10.1016/j.ophtha.2024.07.031>
27. Nebbia, G. et al. Re-identification of patients from imaging features extracted by foundation models. *npj Digit Med* 8, 469 (2025). <https://doi.org/10.1038/s41746-025-01801-0>
28. Podell, D. et al. SDXL: improving latent diffusion models for high-resolution image synthesis. *Int Conf Learn Represent (ICLR)* (2024). <https://doi.org/10.48550/arXiv.2307.01952>
29. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. *Int Conf Learn Represent (ICLR)* (2022). <https://doi.org/10.48550/arXiv.2106.09685>
30. Zhong, M. et al. Multi-LoRA composition for image generation. *Trans Mach Learn Res (TMLR)* (2024). <https://doi.org/10.48550/arXiv.2402.16843>
31. Li, N., Li, T., Hu, C., Wang, K. & Kang, H. A benchmark of ocular disease intelligent recognition: one shot for multi-disease detection. In: *Benchmarking, Measuring, and Optimizing (Bench 2020). Lecture Notes in Computer Science, vol 12614*, 177–193. Springer (2021). https://doi.org/10.1007/978-3-030-71058-3_11
32. Nakayama, L. F. et al. BRSET: a Brazilian multilabel ophthalmological dataset of retina fundus photos. *PLOS Digit Health* 3, e0000454 (2024). <https://doi.org/10.1371/journal.pdig.0000454>
33. Wu, C. et al. A portable retina fundus photos dataset for clinical, demographic, and diabetic retinopathy prediction. *Sci Data* 12, 323 (2025). <https://doi.org/10.1038/s41597-025-04627-3>
34. Ho, J. & Salimans, T. Classifier-free diffusion guidance. arXiv:2207.12598 (2022). <https://doi.org/10.48550/arXiv.2207.12598>
35. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. *Int Conf Learn Represent (ICLR)* (2021). <https://doi.org/10.48550/arXiv.2010.11929>
36. Oquab, M. et al. DINOv2: learning robust visual features without supervision. *Trans Mach Learn Res (TMLR)* (2024). <https://doi.org/10.48550/arXiv.2304.07193>
37. Siméoni, O. et al. DINOv3. arXiv (2025). <https://doi.org/10.48550/arXiv.2508.10104>
38. Islam, M. T. et al. Deep learning-based glaucoma detection with cropped optic cup and disc and blood vessel segmentation. *IEEE Access* 10, 2828–2841 (2022). <https://doi.org/10.1109/ACCESS.2021.3139160>
39. Orlando, J. I. et al. REFUGE Challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med Image Anal* 59, 101570 (2020). <https://doi.org/10.1016/j.media.2019.101570>

40. Zhang, Z. et al. ORIGA(-light): an online retinal fundus image database for glaucoma analysis and research. *Conf Proc IEEE Eng Med Biol Soc* 2010, 3065–3068 (2010).
<https://doi.org/10.1109/IEMBS.2010.5626137>
41. Karthik, Maggie & Dane, S. APTOS 2019 Blindness Detection. Kaggle (2019).
<https://www.kaggle.com/competitions/aptos2019-blindness-detection>.
42. Porwal, P. et al. Indian Diabetic Retinopathy Image Dataset (IDRiD): a database for diabetic retinopathy screening research. *Data* 3, 25 (2018). <https://doi.org/10.3390/data3030025>
43. Abràmoff, M. D. et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol* 131, 351–357 (2013).
<https://doi.org/10.1001/jamaophthalmol.2013.1743>
44. Krause, J. et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology* 125, 1264–1272 (2018).
<https://doi.org/10.1016/j.ophtha.2018.01.034>
45. Liu, R. et al. DeepDRiD: diabetic retinopathy—grading and image quality estimation challenge. *Patterns* 3, 100512 (2022). <https://doi.org/10.1016/j.patter.2022.100512>
46. Ali, M. A. et al. AMDNet23: hybrid CNN-LSTM deep learning approach with enhanced preprocessing for age-related macular degeneration (AMD) detection. *Intell Syst Appl* 21, 200334 (2024).
<https://doi.org/10.1016/j.iswa.2024.200334>
47. Meisel, M. et al. HYAMD: high-resolution fundus image dataset for age-related macular degeneration (AMD) diagnosis (version 1.0.0). PhysioNet (2025). <https://doi.org/10.13026/ydf1-z238>
48. Fang, H. et al. ADAM Challenge: detecting age-related macular degeneration from fundus images. *IEEE Trans Med Imaging* 41, 2828–2847 (2022). <https://doi.org/10.1109/TMI.2022.3172773>
49. Qian, B. et al. HRDC challenge: a public benchmark for hypertension and hypertensive retinopathy classification from fundus images. *Vis Comput* 41, 1061–1077 (2025).
<https://doi.org/10.1007/s00371-024-03384-5>
50. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *Int Conf Learn Represent (ICLR)* (2019). <https://doi.org/10.48550/arXiv.1711.05101>
51. Lin, T. Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. *Proc IEEE Int Conf Comput Vis (ICCV)*, 2980–2988 (2017). <https://doi.org/10.1109/ICCV.2017.324>
52. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc* 32, 675–701 (1937). <https://doi.org/10.1080/01621459.1937.10503522>
53. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7, 1–30 (2006).
54. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837–845 (1988). <https://doi.org/10.2307/2531595>

55. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 57, 289–300 (1995). <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>

Tables

Table 1. Source CFP datasets used for LoRA adapter training. Image counts and descriptive statistics refer to the image-level records included in this study. Age: mean (range); Sex: female/male image counts; Camera: as catalogued in per-image metadata. CFP, colour fundus photograph; LoRA, low-rank adaptation; DR, diabetic retinopathy; AMD, age-related macular degeneration. In the ODIR row, ‘–’ for camera indicates that the camera model was not available.

Dataset	Images included	Camera	Age (mean, range)	Sex (F/M)	Source population	Labels available
ODIR	7,000	–	57.9 (1–91)	3,230/3,770	China	DR, AMD, glaucoma, hypertension, diabetes
BRSET	16,266	Canon CR-2; Nikon NF5050	57.7 (5–97)	10,052/6,214	Brazil	DR, AMD, hypertension, diabetes
mBRSET	5,164	Phelcom Eyer (handheld)	61.4 (3–90)	3,360/1,804	Brazil (community)	DR, hypertension, diabetes

Table 2. Downstream classification datasets, grouped by task. Positive class (%) is the percentage in the held-out 30% partition used for internal validation. When a dataset served as an external target, all images included in the study—not only its internal test partition—were evaluated. Source or acquisition setting reports the available country, region, or dataset provenance; acquisition geography was not available for the aggregated AMDNet23 sources. ONH, optic nerve head; VF, visual field; ICDR, International Clinical Diabetic Retinopathy severity scale; DR, diabetic retinopathy; AMD, age-related macular degeneration.

Dataset	Images analysed	Positive class (%)	Positive-class definition	Source or acquisition setting
Glaucoma				
AITs	502	16.6	Glaucoma (clinical diagnosis)	Hong Kong, China
BEH	634	26.7	Glaucoma (two eye specialists)	Bangladesh
ORIGA	650	25.6	Glaucoma (expert label)	Singapore
REFUGE	800	10.0	Glaucoma (ONH + VF)	China
Diabetic retinopathy				
APTOS	3,662	50.7	Any DR (ICDR grade ≥ 1)	India
IDRiD	413	63.2	Any DR (ICDR grade ≥ 1)	India
MESSIDOR-2	1,744	41.8	Any DR (ICDR grade ≥ 1)	France
DeepDRiD	1,600	55.4	Any DR (ICDR grade ≥ 1)	Shanghai, China
Age-related macular degeneration				
ADAM	400	21.7	AMD (any stage)	China
AMNet23	1,994	24.7	AMD (any stage)	Multiple public sources
HYAMD	1,370	33.3	AMD (any stage)	Israel
Systemic hypertension				
HRDC	712	50.0	Systemic hypertension	China

Table 3. Strategy-specific counted FLOPs for synthetic-image generation and backbone pretraining, with observed unweighted mean AUROC. Internal AUROC is the unweighted arithmetic mean across the 12 downstream datasets; external AUROC is the unweighted arithmetic mean across the eight external dataset pairs for the same disease. For each strategy, the reported FLOP count represents a hypothetical single-strategy run in which the required synthetic images are generated and one backbone is pretrained. These strategy-specific counts should not be added together and differ from the approximately 90 EFLOPs used to generate the shared pool once and run all six study strategies. AUROC, area under the receiver operating characteristic curve; EFLOP, 10^{18} floating-point operations; FLOPs, floating-point operations. N $\times$ denotes a synthetic-to-real image ratio of N:1, with the real corpus held fixed.

Strategy	Counted FLOPs (EFLOPs)	Mean internal AUROC	Mean external AUROC	Empirical Pareto status
real-only	0.017	0.862	0.683	internal Pareto (minimum counted FLOPs); external Pareto (minimum counted FLOPs)
synthetic-only	4.465	0.855	0.695	internally dominated; external Pareto
real-plus-1×synthetic	4.482	0.881	0.734	internal Pareto; external Pareto
real-plus-5×synthetic	22.342	0.885	0.745	internal Pareto; external Pareto
real-plus-10×synthetic	44.667	0.906	0.742	internal Pareto (highest observed mean AUROC); externally dominated
real-plus-20×synthetic	89.316	0.905	0.777	internally dominated; external Pareto (highest observed mean AUROC)

Figures

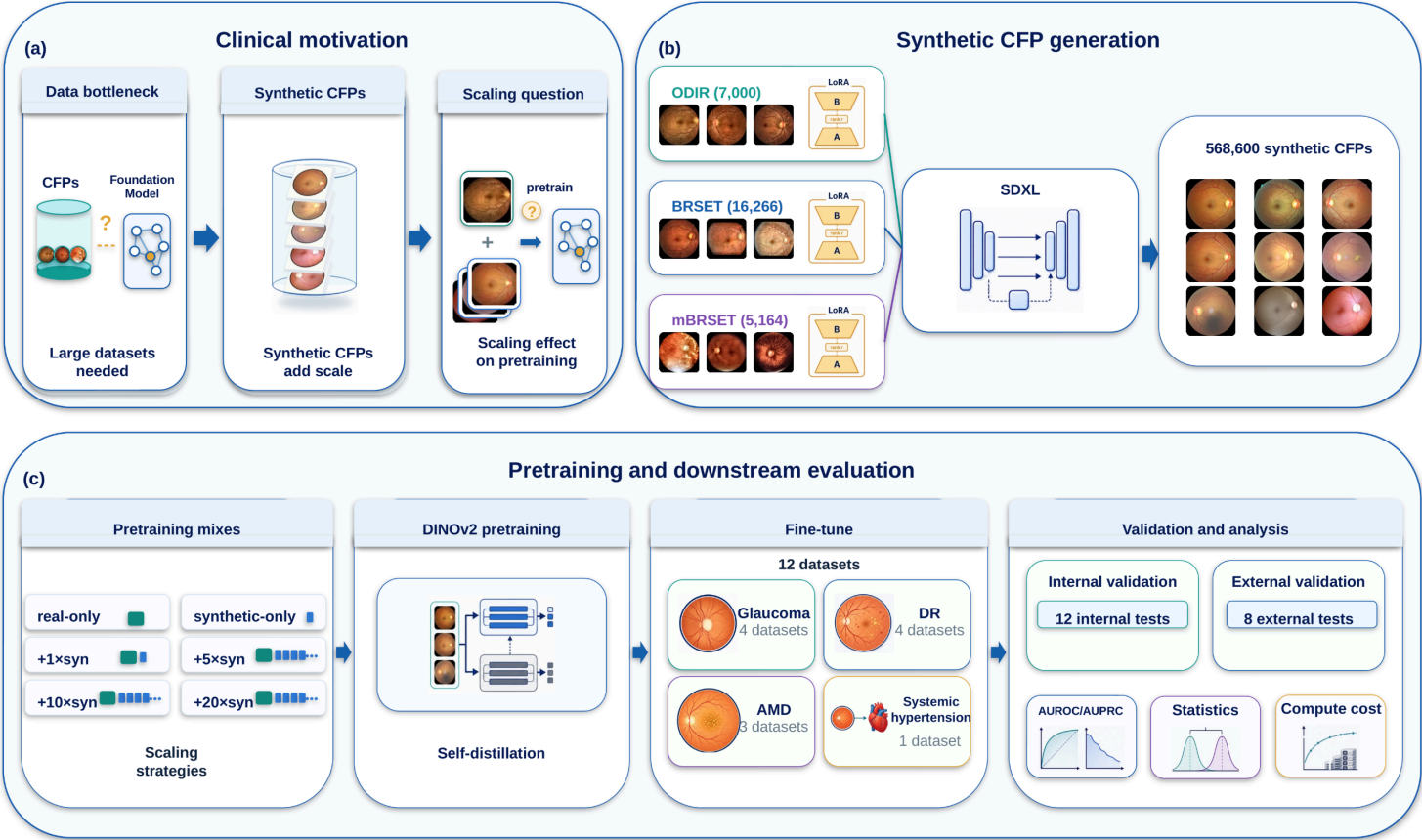


Figure 1

Study design. (a) A fixed corpus of real colour fundus photographs (CFPs) was combined with increasing synthetic CFP volumes to examine the relationship between downstream discriminative performance and counted floating-point operations (FLOPs) for synthetic-image generation and pretraining. (b) Separate low-rank adaptation (LoRA) adapters trained on ODIR, BRSET, and mBRSET were applied to a Stable Diffusion XL (SDXL) base model to generate metadata-conditioned synthetic CFPs. (c) Vision Transformer (ViT) backbones with the ViT-B/14 architecture were pretrained under six real and synthetic data strategies, fine-tuned on internal tasks, and evaluated on internal and external datasets for the same disease.

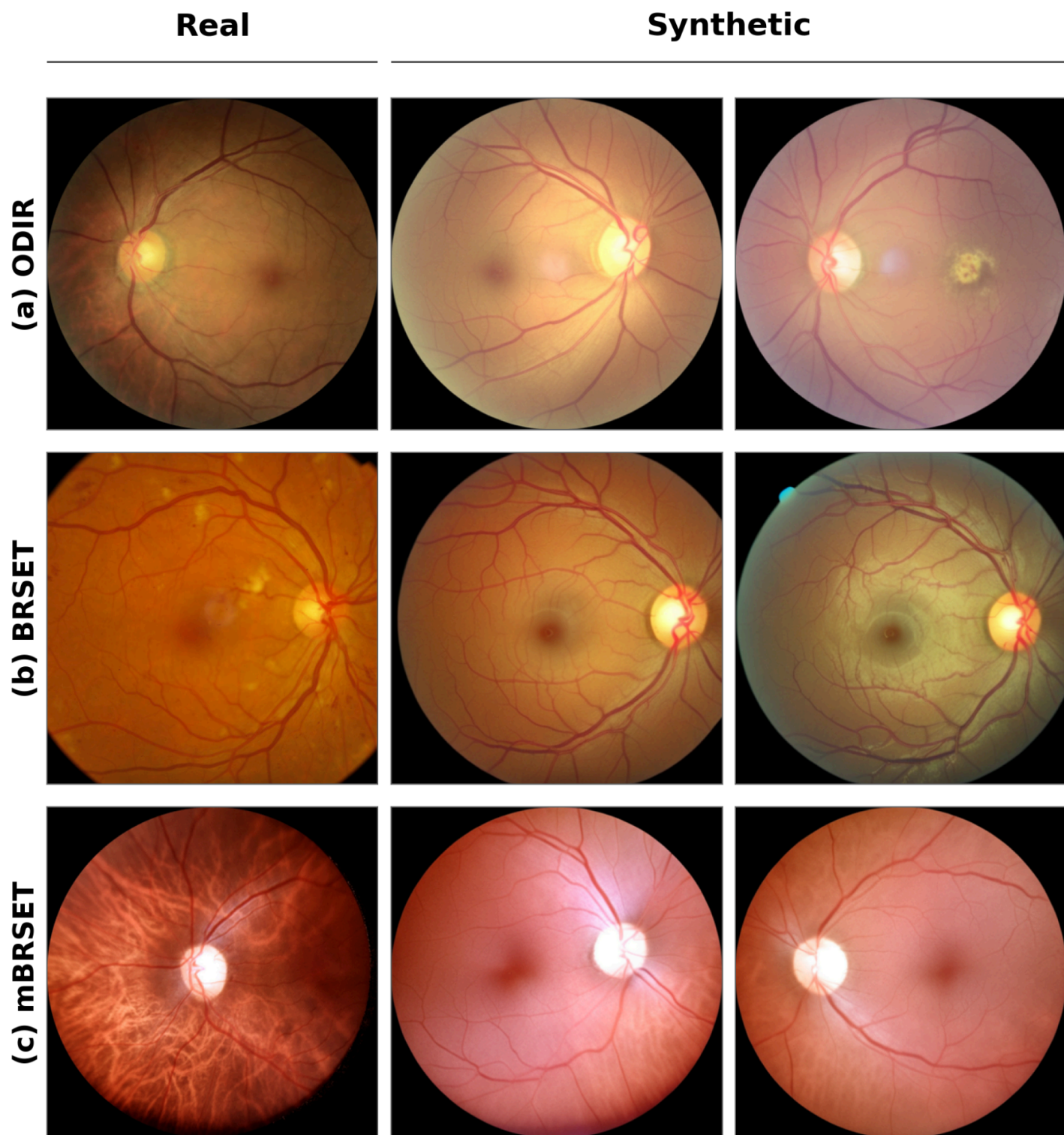


Figure 2

Examples of real and adapter-generated synthetic colour fundus photographs (CFPs). Rows (a), (b) and (c) show ODIR, BRSET and mBRSET, respectively. Each row shows one illustrative real CFP from a source dataset and two independently generated CFPs from that dataset's adapter; the synthetic images are not case-matched to the real example (Methods).

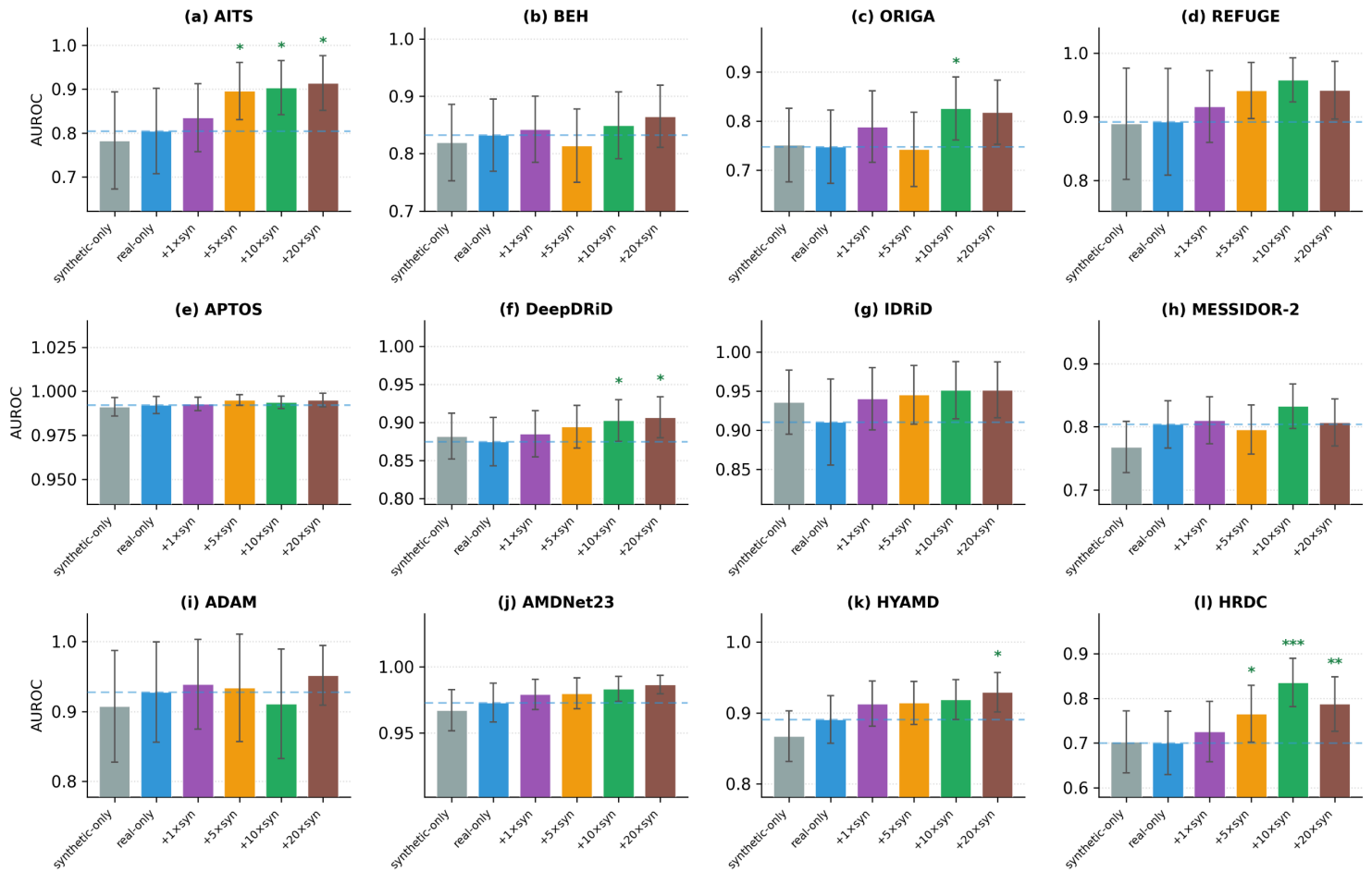


Figure 3

Internal test performance across scaling strategies for downstream evaluation tasks. Per-dataset test-set AUROC across the six scaling strategies is presented as a 12-panel grid covering glaucoma, diabetic retinopathy (DR), age-related macular degeneration (AMD), and systemic hypertension prediction from CFPs. Each panel shows test-set AUROC for the six strategies with DeLong 95% confidence intervals; the dashed blue line marks the real-only baseline. Asterisks indicate adjusted q-value levels after Benjamini-Hochberg correction of DeLong comparisons with real-only (* $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$). Among the evaluated strategies, real-plus-10xsynthetic had the highest observed internal mean AUROC, 0.906, corresponding to $\Delta = +0.043$ relative to the real-only mean of 0.862. AUROC, area under the receiver operating characteristic curve; CFP, colour fundus photograph. Vertical-axis limits vary by panel and do not start at zero; comparisons should use the labelled values. In the artwork, +1xsyn, +5xsyn, +10xsyn, and +20xsyn denote the four real-plus-synthetic strategies. N $\times$ denotes a synthetic-to-real image ratio of N:1, with the real corpus held fixed.

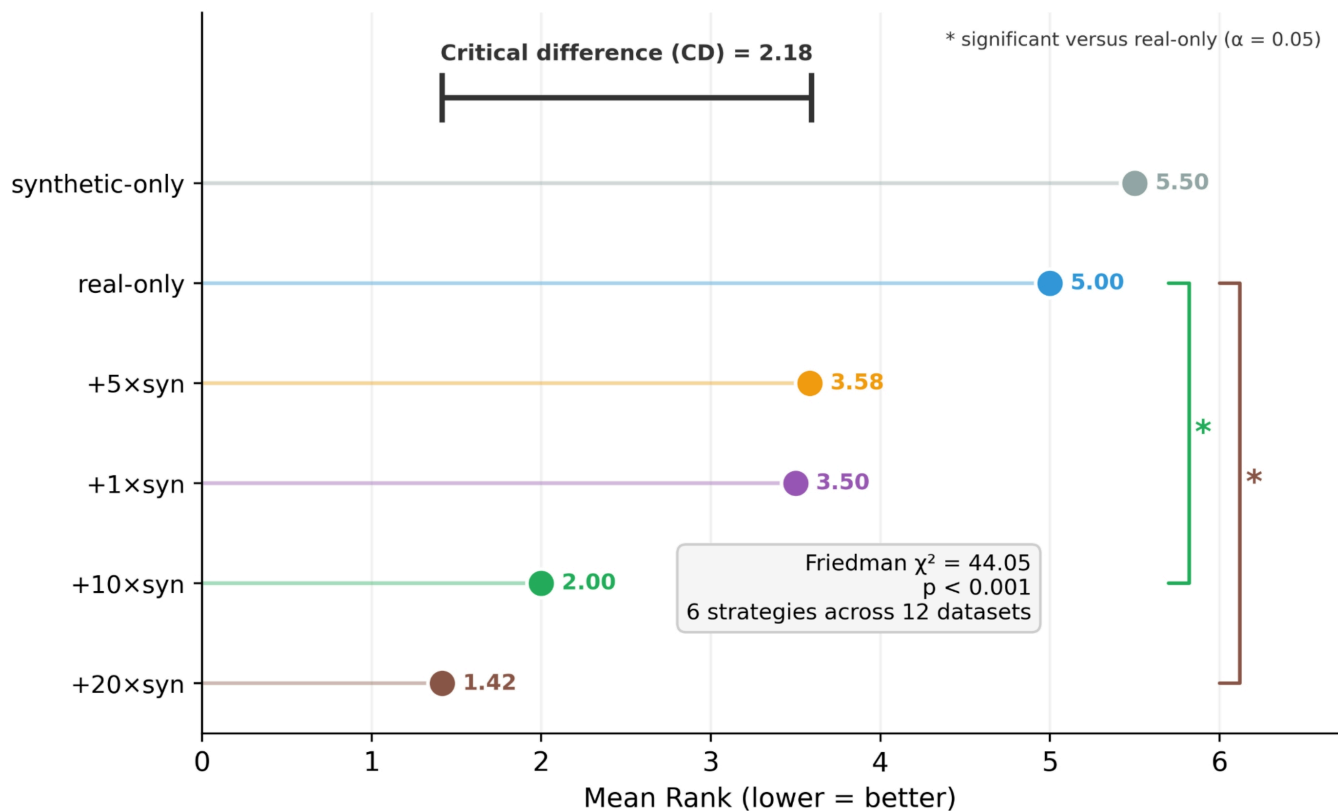


Figure 4

Friedman mean-rank chart with Nemenyi critical difference. Mean ranks of the six scaling strategies across the 12 downstream datasets are shown, with lower rank indicating better AUROC. The Friedman test yielded $\chi^2(5) = 44.05$, $p < 0.001$. For the Nemenyi post hoc test, the critical difference was 2.18 at $\alpha = 0.05$ ($k = 6$ strategies, $N = 12$ datasets); a pairwise mean-rank separation greater than 2.18 indicates a significant difference. The separations of real-plus-10xsynthetic and real-plus-20xsynthetic from real-only exceeded the critical difference. Comparisons that did not exceed the critical difference were not detected as significant and should not be interpreted as equivalent. AUROC, area under the receiver operating characteristic curve. In the artwork, +1xsyn, +5xsyn, +10xsyn, and +20xsyn denote the four real-plus-synthetic strategies. N $\times$ denotes a synthetic-to-real image ratio of N:1, with the real corpus held fixed.

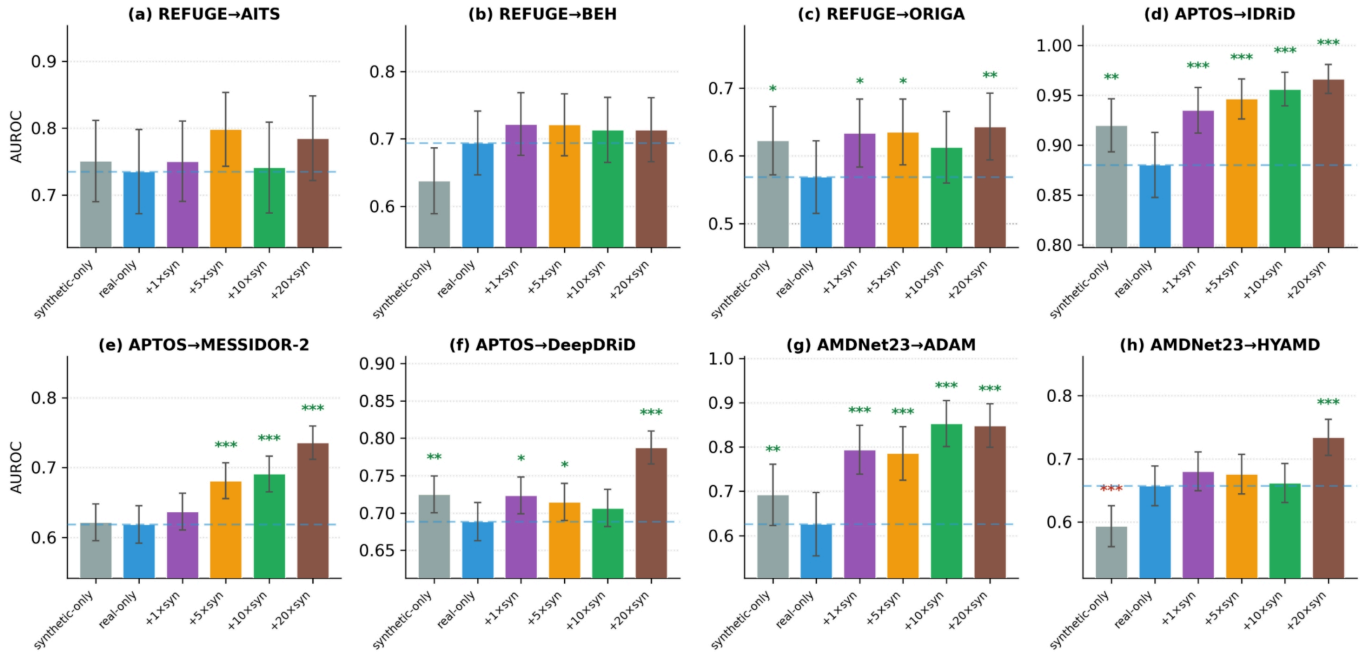


Figure 5

External validation performance across scaling strategies. AUROC is shown for eight external validations: REFUGE→AITS, REFUGE→BEH, and REFUGE→ORIGA for glaucoma; APTOS→IDRiD, APTOS→MESSIDOR-2, and APTOS→DeepDRiD for diabetic retinopathy (DR); and AMDNet23→ADAM and AMDNet23→HYAMD for age-related macular degeneration (AMD). Each panel shows AUROC for the six strategies with DeLong 95% confidence intervals; the dashed blue line marks the real-only baseline. Asterisks indicate adjusted q-value levels after Benjamini–Hochberg correction of DeLong comparisons with real-only (* $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$; green, improvement; red, deterioration). Among the evaluated strategies, real-plus-20×synthetic had the highest observed external mean AUROC of 0.777, corresponding to $\Delta = +0.093$ relative to the real-only mean of 0.683. AUROC, area under the receiver operating characteristic curve. External AUROC was calculated on all images included in each target dataset, not on its internal 30% test partition. Vertical-axis limits vary by panel and do not start at zero; comparisons should use the labelled values. In the artwork, +1×syn, +5×syn, +10×syn, and +20×syn denote the four real-plus-synthetic strategies. N× denotes a synthetic-to-real image ratio of N:1, with the real corpus held fixed.

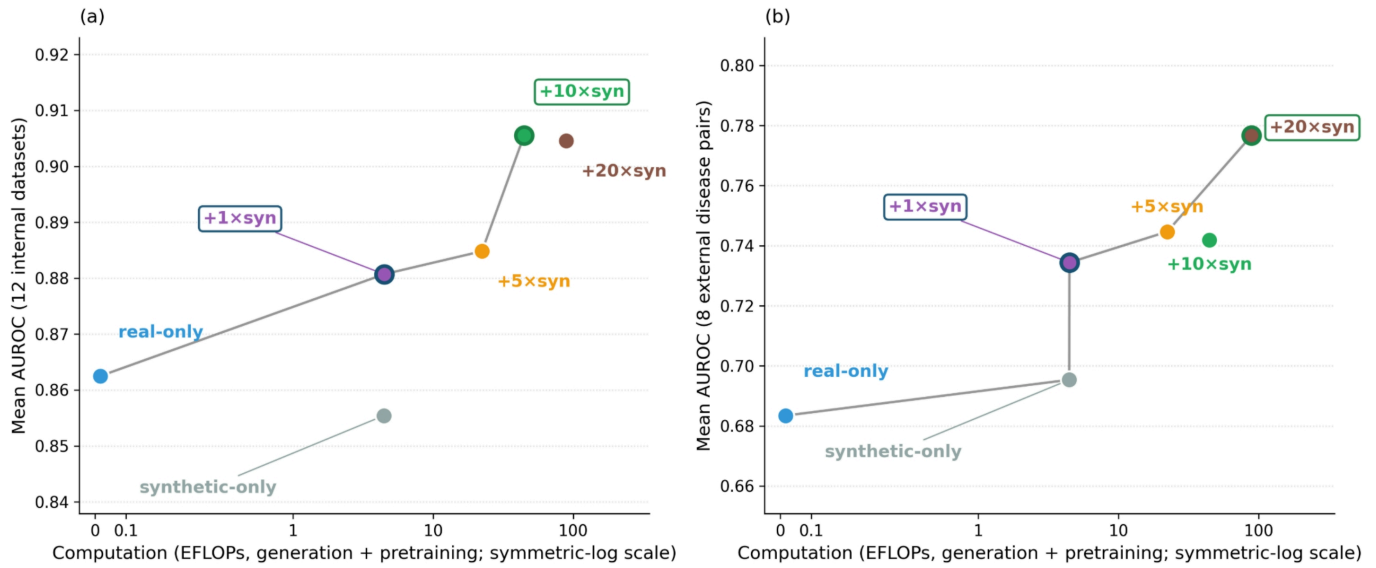


Figure 6

Empirical Pareto frontiers for mean AUROC versus counted FLOPs. Each point shows a strategy's counted FLOPs for generation and pretraining and its observed unweighted mean AUROC. Panel (a) shows internal mean AUROC across the 12 downstream datasets; panel (b) shows external mean AUROC across the eight external dataset pairs for the same disease. The horizontal axis shows counted FLOPs for synthetic-image generation and backbone pretraining on a symmetric-log scale. Real-plus-1×synthetic (blue ring) is the strategy designated by the study-defined rule in both panels. The green ring marks the strategy with the highest observed mean AUROC in each panel: real-plus-10×synthetic internally and real-plus-20×synthetic externally. In the internal summary, synthetic-only and real-plus-20×synthetic are empirically dominated. In the external summary, real-plus-10×synthetic is empirically dominated by real-plus-5×synthetic. Dominance refers only to the two plotted summary values and does not imply statistically significant, task-by-task, or clinical inferiority. The vertical axis is truncated to the observed data range. AUROC, area under the receiver operating characteristic curve; FLOPs, floating-point operations. In the artwork, +1×syn, +5×syn, +10×syn, and +20×syn denote the four real-plus-synthetic strategies. N× denotes a synthetic-to-real image ratio of N:1, with the real corpus held fixed.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SyntheticFundusScalingSupplementaryInformation.pdf](#)