DECOMPOSING CROSS-MODALITY FAILURE IN DIABETIC RETINOPATHY GRADING

Anonymous authors
Paper under double-blind review

ABSTRACT

Colour and near-infrared (NIR) fundus imaging depict the same retina through different image representations. A diabetic-retinopathy (DR) grader trained on colour photographs must therefore preserve disease knowledge while accommodating a new channel structure and intensity distribution. Across eight encoder configurations and five paired seeds, zero-shot colour-to-NIR transfer degrades sharply without a corresponding loss of confidence. We factorise the source intervention into channel representation and marginal intensity alignment: each recovers part of the loss, while their combination raises ResNet-50 external QWK from 0.08 to 0.54 and improves every tested backbone. A frozen-model control and layer-wise probes show that much of the failure can be reduced without learning a cross-domain mapping. Predictive uncertainty remains weak for shift detection, whereas Mahalanobis distance reaches OOD AUROC 0.93 and remains informative for residual grading errors. The resulting decomposition separates an input-addressable component from what remains for adaptation, acquisition, or new data.

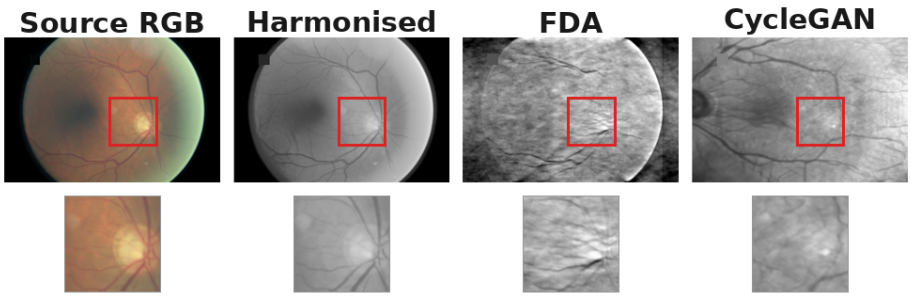

Figure 1: Colour-to-NIR transformations for one EyePACS case. The red box marks a local retinal region and the lower row enlarges that region. Harmonisation changes the source representation while preserving retinal geometry. FDA and CycleGAN provide adaptation and learned-translation comparisons. The CycleGAN panel uses seed 47. Seed variation is shown in Figure 6.

1 INTRODUCTION

Deep learning systems can grade diabetic retinopathy (DR) reliably from colour fundus photographs (Gulshan et al., 2016), yet clinical retinal imaging spans several acquisition modalities. Near-infrared (NIR) reflectance accompanies many OCT examinations and is routinely produced by confocal scanning laser ophthalmoscopy. Successful transfer would extend a colour-trained grader to settings without colour photography; failure at high confidence is more problematic, because a standard abstention rule may not detect it.

A failure on NIR may look like a loss of disease knowledge, but the model is also receiving an input whose channel organisation and intensity statistics differ from those seen during learning. We therefore ask how much DR knowledge survives the modality change, how much failure arises from input representation before the classifier is changed, and whether the model recognises the shift when its predictions fail.

We separate these effects with two low-capacity source operations: replicated green changes channel representation, and quantile alignment matches marginal intensities to an aggregate unlabelled NIR

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

reference. Because neither learns a cross-domain mapping, recovery remains attributable to known input changes. A frozen target-side control, layer-wise probes, adaptation baselines, direct external grading, structural proxies, and shift detection then test how far this explanation extends and what remains after the input mismatch is reduced.

Our contributions are:

- a factorised source intervention that separates channel representation from marginal intensity alignment, together with a frozen-model control that tests the same input-level explanation without retraining.
- evaluation across five paired seeds and eight encoder configurations, spanning ImageNet, self-supervised, and retinal foundation-model pre-training.
- comparison with preprocessing, test-time adaptation, unsupervised domain adaptation, and learned translation, complemented by a blinded lesion-fidelity audit, and
- an explicit comparison between predictive uncertainty and feature-space distance for detecting both the acquisition shift and the errors that remain after harmonisation.

2 Related Work

Cross-modal retinal learning. Cross-modal retinal studies show that useful pathology survives a change of acquisition, but not always in a directly transferable form. Sükei et al. (2024) report lossy interchange of paired fundus/OCT representations on OLIVES; supervised infrared grading is feasible when labels are available (Xue et al., 2023); and lesions can render differently in paired colour and infrared images (Xiao et al., 2024). Infrared representations also support retinal-thickness prediction and downstream DR classification (Holmberg et al., 2020). We instead ask why a colour-trained grader fails when transferred without NIR task labels.

Pre-training and adaptation. Large-scale pre-training offers one route to more transferable retinal representations, motivating our comparison of ImageNet, self-supervised, and retinal foundation models (Zhou et al., 2023). Explicit adaptation instead reduces source–target mismatch through feature alignment (Ganin et al., 2016; Sun & Saenko, 2016), entropy minimisation (Wang et al., 2021; Liang et al., 2020), Fourier alignment (Yang & Soatto, 2020), or learned translation (Zhu et al., 2017). DR-specific work adapts colour fundus to ultra-wide-field imaging (Ran et al., 2024; **?**), while foundation models still degrade across acquisition domains (Li et al., 2026). We ask what part of this failure is removable by controlled input changes before learned adaptation.

Uncertainty and shift detection. Predictive uncertainty need not increase under distribution shift (Ayhan & Berens, 2018; Ovadia et al., 2019; Zadorozhny et al., 2021), including in DR screening (Nandy et al., 2021). Feature-distance methods instead measure how far an input lies from the source representation (Lee et al., 2018; González et al., 2022). We compare these signals before and after harmonisation and test whether they identify the grading errors that remain.

3 Method

3.1 Data and Evaluation Cohorts

EyePACS (Kaggle, 2015) is the labelled source domain, with 35,126 colour fundus images graded 0–4 and patient-aware 70/15/15 splits recovered from the left/right filename stem. OLIVES (Prabhushankar et al., 2022) provides 1,268 unique visit-level NIR images from 96 eyes and 87 patients. Repeated visits remain grouped by eye/patient, duplicated annotation rows and OCT B-scans are excluded, and the aggregate NIR reference histogram is estimated from the OLIVES training partition only.

Direct target evaluation uses the locked NIR-External cohort: 612 NIR images from 306 patients acquired on two cSLO systems absent from OLIVES. Two retinal specialists independently assign International Clinical DR grades 0–4 (pre-adjudication QWK 0.83), with disagreements resolved by a third reader. The cohort contributes neither optimisation data nor harmonisation statistics.

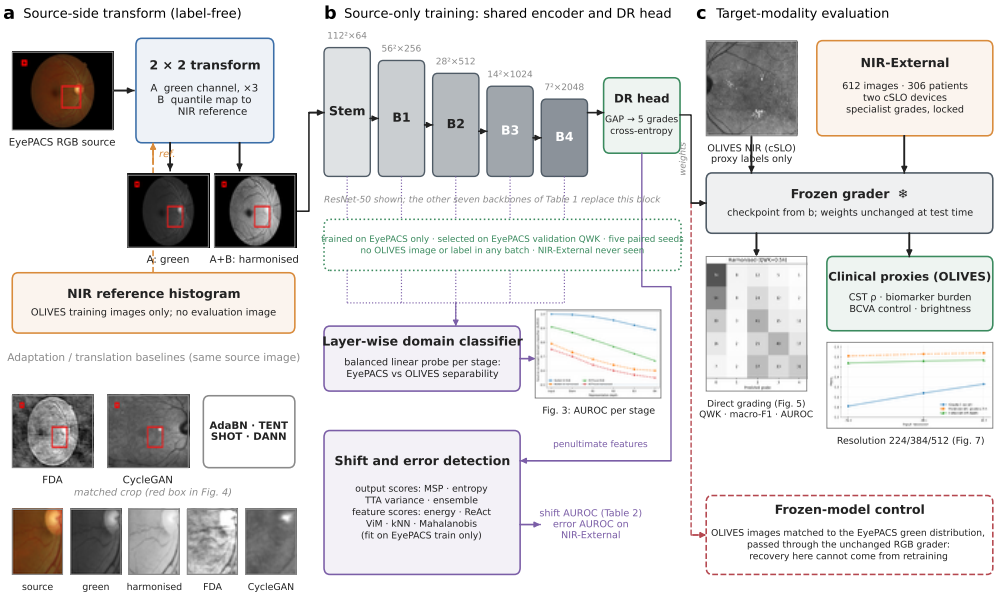

Figure 2: Study design. The DR grader is trained and selected on EyePACS samples only. Aggregate unlabelled OLIVES statistics are used only by explicitly target-informed harmonisation/adaptation conditions. Evaluation uses OLIVES and locked NIR-EXTERNAL; the target-side control keeps the RGB grader fixed while changing target intensity statistics.

Appendix B gives grade counts and the full protocol; OLIVES instead supplies patient-grouped CST and biomarker proxies.

3.2 MODELS, TRAINING, SEEDS AND UNCERTAINTY

ResNet-50 (He et al., 2016) is the primary grader; we also test ConvNeXt-T (Liu et al., 2022), EfficientNet-B4 (Tan & Le, 2019), ViT-B/16 (Dosovitskiy et al., 2021), DINOv2 (Oquab et al., 2024), RETFound (Zhou et al., 2023) with fundus- and OCT-pretrained weights, and VisionFM (Qiu et al., 2024). Every encoder and five-class head is trained on EyePACS and selected by EyePACS validation QWK; NIR-EXTERNAL is never used for training or selection. Five paired seeds (7, 17, 27, 37, 47) share identical splits and selection rules, and tables report mean ± SD.

Harmonisation is target-informed but label-free: its only target input is the aggregate intensity distribution of unlabelled OLIVES training images. AdaBN, TENT, SHOT, FDA, DANN, and CycleGAN are adaptation or translation baselines, with CORAL (Sun & Saenko, 2016) reported separately. Target-aware methods use only the OLIVES training partition, share the source split/checkpoint rule, and never use NIR-EXTERNAL for adaptation or tuning. Appendix A gives implementation details.

Shift detection uses negative maximum softmax probability, predictive entropy, TTA variance (Ayhan & Berens, 2018), and five-model disagreement (Lakshminarayanan et al., 2017), together with energy (Liu et al., 2020), ReAct (Sun et al., 2021), ViM (Wang et al., 2022), kNN distance (Sun et al., 2022), and class-conditional Mahalanobis distance (Lee et al., 2018). Feature statistics are fitted only on EyePACS training representations. Appendix A specifies OOD settings and confidence strata; cohort intervals use patient- or eye-level cluster bootstrap unless stated otherwise.

3.3 FACTORISED HARMONISATION AND FROZEN-MODEL CONTROL

The source-side probe isolates two differences between colour and NIR: channel representation and marginal intensity distribution. The first replaces RGB with replicated green; the second maps source retinal-field intensities monotonically to the aggregate OLIVES training reference. The circular border is excluded from the estimated distributions and restored afterwards, leaving pixel coordinates and topology unchanged. Because the transform has no learned cross-domain capacity, recovery can be attributed directly to these input factors.

Table 1: Architecture and pre-training sensitivity (mean ± SD over five paired seeds). The first three models are the prespecified primary backbones used in the main experiments. The remaining five are supplementary architecture and pre-training probes. RGB denotes direct evaluation of the colour-trained grader. Harmonised denotes the model retrained on the deterministically harmonised source representation, without any target-feature adaptation term. EyePACS QWK is in-domain held-out test performance. OLIVES G0 is the percentage of target images assigned grade 0. Ext. QWK is QWK on the locked NIR-EXTERNAL cohort. Best value per column in bold, second best underlined.

		EyePACS QWK ↑		OLIVES G0 (%) ↓		Ext. QWK ↑	
Backbone	Pre-training	RGB	Harmonised	RGB	Harmonised	RGB	Harmonised
ResNet-50	ImageNet, supervised	0.641 ± 0.006	0.629 ± 0.007	95.8 ± 0.9	53.0 ± 2.6	0.08 ± 0.03	0.54 ± 0.02
ConvNeXt-T	ImageNet, supervised	0.673 ± 0.005	0.660 ± 0.006	91.2 ± 1.4	46.5 ± 2.4	0.16 ± 0.03	0.57 ± 0.02
RETFound	colour fundus, self-supervised	0.704 ± 0.004	0.694 ± 0.005	73.4 ± 2.1	39.6 ± 2.0	0.35 ± 0.02	0.62 ± 0.02
EfficientNet-B4	ImageNet, supervised	0.681 ± 0.005	0.668 ± 0.006	93.0 ± 1.2	48.9 ± 2.6	0.14 ± 0.02	0.57 ± 0.02
ViT-B/16	ImageNet, supervised	0.688 ± 0.006	0.675 ± 0.006	87.6 ± 1.7	44.2 ± 2.5	0.20 ± 0.03	0.60 ± 0.02
DINOv2 ViT-B/14	natural images, self-supervised	$\underline{0.716 \pm 0.004}$	$\underline{0.705 \pm 0.005}$	81.0 ± 1.9	40.5 ± 2.2	0.28 ± 0.02	0.62 ± 0.02
RETFound	OCT, self-supervised	0.696 ± 0.005	0.688 ± 0.005	$\mathbf{64.9 \pm 2.0}$	$\underline{35.7 \pm 1.9}$	$\mathbf{0.42 \pm 0.02}$	$\underline{0.65 \pm 0.02}$
VisionFM	retinal, self-supervised	$\mathbf{0.724 \pm 0.004}$	$\mathbf{0.714 \pm 0.004}$	$\underline{67.3 \pm 1.8}$	$\mathbf{33.8 \pm 1.8}$	$\underline{0.40 \pm 0.02}$	$\mathbf{0.66 \pm 0.02}$

Crossing the factors gives RGB, colour-preserving intensity alignment, replicated green, and their combination, which we call harmonised. We also remove border masking and add CORAL as sensitivity variants. A separate target-side control matches OLIVES toward EyePACS statistics and passes it through the unchanged RGB grader; recovery here isolates an input-statistics effect under a fixed decision function.

3.4 BASELINES AND VALIDATION

To place this low-capacity intervention against stronger alternatives, we compare it with Graham normalisation (Graham, 2015), green-channel CLAHE (Zuiderveld, 1994), AdaBN (Li et al., 2018), TENT (Wang et al., 2021), SHOT (Liang et al., 2020), FDA (Yang & Soatto, 2020), DANN (Ganin et al., 2016), and CycleGAN (Zhu et al., 2017). Target-aware methods see only the OLIVES training partition and are frozen before external evaluation; Appendix A specifies which components each method updates. NIR-EXTERNAL is not used for optimisation or model selection.

Evaluation separates target-task performance from mechanistic checks. NIR-EXTERNAL supplies QWK, macro-F1, and referable-DR AUROC against specialist grades. OLIVES contributes predicted-grade/CST association and biomarker discrimination through patient-grouped probes on frozen encoder features. Finally, because a transformation can improve grading while altering local pathology, a blinded specialist review evaluates lesion fidelity on transformed EyePACS images; Appendix E gives the audit protocol and confidence intervals.

4 RESULTS

4.1 TRANSFER ACROSS BACKBONES

The failure persists across all eight encoders in Table 1, although pre-training changes its size. Zero-shot external QWK is 0.08 for ResNet-50, 0.28 for DINOv2, and 0.35–0.42 for the retinal foundation-model variants. Harmonisation improves every backbone by 0.23–0.46 QWK while changing EyePACS QWK by at most 0.013. Stronger representations therefore narrow the initial gap without removing the component recoverable through source representation. Appendix D gives OLIVES proxy results.

This pattern suggests two effects acting together: pre-training changes how far the target begins from useful behaviour, while input compatibility still determines a substantial part of the remaining gap. We next examine whether this mismatch is visible inside the encoder itself.

4.2 LAYER-WISE REPRESENTATION MISMATCH

The mismatch remains visible inside the encoder. A balanced linear classifier separates held-out EyePACS and OLIVES representations at successive stages (Figure 3), with repeated OLIVES visits grouped by eye/patient. RGB ResNet-50 still gives domain AUROC 0.889 at the final block; harmonisation reduces it to 0.600. RETFound begins with lower separation and follows the same direction, showing that the input intervention propagates through the learned representation.

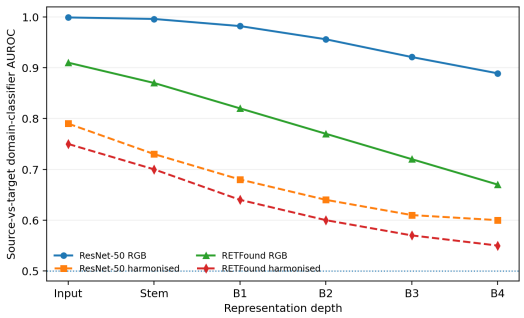

Figure 3: Layer-wise modality separability. Curves give the AUROC of a balanced linear classifier that distinguishes EyePACS from OLIVES at the input, stem, and residual blocks B1–B4. AUROC 0.5 indicates no linear separation. Solid lines show RGB-trained models and dashed lines show harmonised models. Harmonisation reduces modality identifiability at every depth for both backbones.

The confidence signal is nevertheless meaningful within the source domain. For ResNet-50, abstaining on progressively lower-confidence EyePACS predictions raises accuracy from 0.736 ± 0.009 at full coverage to 0.977 ± 0.010 at 10% coverage, while QWK rises from 0.641 ± 0.006 to 0.964 ± 0.015 (Figure 10, Appendix J). The question is therefore not whether confidence is useful at all, but whether it continues to track failure after the modality boundary is crossed.

4.3 Shift and Residual-Error Detection

Crossing into NIR breaks the source-domain relation between confidence and correctness. The zero-shot ResNet-50 assigns $95.8 \pm 0.9\%$ of OLIVES images to grade 0, yet mean TTA confidence rises from 0.621 ± 0.001 on EyePACS to 0.670 ± 0.004 on OLIVES in every seed (Figure 8). Predicted severity in the small non-zero tail still tracks darkness ($\rho = -0.436 \pm 0.032$), while the overall predicted-grade/CST association becomes negative ($\rho = -0.47 \pm 0.04$). On NIR-External, the same model reaches 0.08 ± 0.03 QWK and predicts grade 0 for most images.

Table 2 makes this distinction explicit. Negative maximum-softmax probability has OOD AUROC 0.31, so the confidence signal is not merely weak but inverted relative to the expected uncertainty direction; predictive entropy, TTA variance, and seed disagreement remain near chance at 0.54–0.59. In contrast, logit- and representation-based scores separate the modalities much more clearly, with Mahalanobis distance reaching 0.93. The acquisition shift is therefore strongly expressed in feature space even when the output distribution remains stable or becomes more confident.

Harmonisation reduces but does not erase the representation gap: Mahalanobis shift AUROC falls from 0.93 to 0.67 and kNN from 0.91 to 0.69. On NIR-External, these distances also identify residual errors (AUROC 0.76 and 0.74), whereas output-based scores reach 0.56–0.61. Referable-DR error raises all scores by 0.02–0.03 without changing their order. Feature distance therefore detects both shift and elevated error risk, but is not a calibrated measure of error severity.

4.4 Factorised Ablation and Baselines

The factorial experiment shows that neither source factor alone explains the recovery. Intensity alignment raises external QWK from 0.08 to 0.31, while replacing RGB with replicated green raises it to 0.26; combining the two reaches 0.54 (Table 3). Border masking contributes a smaller additional gain, and adding CORAL changes QWK by only 0.01. Most of the recoverable component in this design is therefore associated with channel representation and marginal intensity statistics rather than with target-feature alignment.

The paired factorial bootstrap gives the same picture quantitatively: the marginal QWK effect is $+0.21$ for channel representation (95% CI $[0.17, 0.24]$) and $+0.26$ for intensity alignment ($[0.22, 0.29]$), whereas the interaction is smaller at $+0.05$ ($[-0.01, 0.11]$). These effects belong to the controlled input interventions themselves; the factorial design does not isolate the underlying acquisition physics.

The frozen target-side experiment provides an important complementary test. Matching OLIVES toward EyePACS statistics improves the unchanged RGB grader, so part of the recovery appears even when the decision function and training trajectory are held fixed. Retraining on a source representation

Table 2: Shift detection and error detection from the frozen ResNet-50. Columns 2–3 give OOD AUROC for distinguishing OLIVES NIR images (positive class) from held-out EyePACS, using scores computed from the source-only RGB grader and from the harmonised grader. Columns 4–5 give AUROC for identifying errors of the harmonised grader on NIR-EXTERNAL, with the positive class defined as an incorrect five-class prediction and, as a sensitivity analysis, as an incorrect referable-DR decision. The first four scores are derived from the output distribution. The remaining five from logits or penultimate features. Best value per column in bold, second best underlined.

	Shift detection (EyePACS vs NIR)		Error detection, harmonised	
Score	RGB grader	Harmonised grader	Exact grade	Referable DR
Negative max. softmax probability	0.31	0.46	0.56	0.58
Predictive entropy	0.54	0.51	0.58	0.60
TTA variance	0.56	0.52	0.59	0.61
Deep-ensemble disagreement	0.59	0.54	0.61	0.63
Energy	0.76	0.62	0.67	0.69
ReAct-based score	0.83	0.64	0.70	0.72
ViM	0.89	**0.71**	0.72	0.75
kNN feature distance	<u>0.91</u>	<u>0.69</u>	<u>0.74</u>	<u>0.77</u>
Mahalanobis distance	**0.93**	0.67	**0.76**	**0.79**

Table 3: Factorised preprocessing conditions and frozen-model control for ResNet-50 (mean $\pm$ SD over five paired seeds). Retrain? indicates whether a model was trained on the transformed source images. Border mask? indicates whether the circular field border was excluded when estimating the intensity distributions. The frozen target-side row transforms OLIVES images toward EyePACS statistics and passes them through the unchanged RGB-trained grader. Columns as in Table 1. Best value per column in bold, second best underlined. The CORAL row uses unlabelled target features during training and is excluded from the ranking.

Condition	Retrain?	Border mask?	OLIVES G0 ↓	CST ρ ↑	Ext. QWK ↑
RGB baseline	yes	–	95.8 ± 0.9	-0.47 ± 0.04	0.08 ± 0.03
RGB + colour-preserving intensity match	yes	yes	74.6 ± 2.7	$+0.09 \pm 0.04$	0.31 ± 0.03
Green replicated only	yes	yes	80.3 ± 2.5	$+0.02 \pm 0.05$	0.26 ± 0.03
Green + intensity match	yes	no	<u>57.4 ± 2.9</u>	<u>$+0.29 \pm 0.04$</u>	<u>0.50 ± 0.03</u>
Green + intensity match (full transform, primary)	yes	yes	$\mathbf{53.0 \pm 2.6}$	$\mathbf{+0.33 \pm 0.03}$	$\mathbf{0.54 \pm 0.02}$
Frozen target-side renormalisation	no	yes	68.7 ± 2.2	$+0.18 \pm 0.04$	0.34 ± 0.03
Full transform + CORAL (target-feature adaptation variant)	yes	yes	52.1 ± 2.8	$+0.34 \pm 0.03$	0.55 ± 0.02

Table 4: Normalisation, adaptation and translation baselines for ResNet-50 (mean $\pm$ SD over five paired seeds). AdaBN, TENT and SHOT all initialise from the same source grader but update different components on unlabelled OLIVES training images; FDA, DANN and CycleGAN use EyePACS and OLIVES training partitions. Every adapted model is frozen before NIR-EXTERNAL evaluation. Columns as in Table 3, with Fluid the AUROC for the macular-fluid composite. Best value per column in bold, second best underlined (ties share the underline).

Method	Type	OLIVES G0 ↓	CST ρ ↑	Fluid AUROC ↑	Ext. QWK ↑
RGB baseline	–	95.8 ± 0.9	-0.47 ± 0.04	0.54 ± 0.03	0.08 ± 0.03
Graham normalisation	preprocessing	83.7 ± 2.4	-0.08 ± 0.05	0.61 ± 0.03	0.22 ± 0.03
Green + CLAHE	preprocessing	76.1 ± 2.5	$+0.06 \pm 0.04$	0.65 ± 0.03	0.29 ± 0.03
AdaBN	BN-stat adaptation	72.4 ± 2.3	$+0.12 \pm 0.04$	0.67 ± 0.03	0.30 ± 0.03
Frozen target-side histogram renorm.	preprocessing	68.7 ± 2.2	$+0.18 \pm 0.04$	0.69 ± 0.03	0.34 ± 0.03
TENT	entropy adaptation	60.2 ± 2.8	$+0.25 \pm 0.04$	0.72 ± 0.03	0.42 ± 0.03
FDA	preprocessing	61.5 ± 2.6	$+0.24 \pm 0.04$	0.72 ± 0.03	0.43 ± 0.03
SHOT	source-free DA	58.4 ± 2.7	$+0.27 \pm 0.04$	0.73 ± 0.03	0.45 ± 0.03
DANN	unsupervised DA	55.7 ± 2.5	$+0.30 \pm 0.03$	<u>0.75 ± 0.02</u>	0.49 ± 0.03
CycleGAN	learned translation	<u>54.3 ± 3.1</u>	<u>$+0.31 \pm 0.04$</u>	<u>0.75 ± 0.03</u>	<u>0.52 ± 0.03</u>
Deterministic harmonisation	preprocessing	$\mathbf{53.0 \pm 2.6}$	$\mathbf{+0.33 \pm 0.03}$	$\mathbf{0.76 \pm 0.02}$	$\mathbf{0.54 \pm 0.02}$

that better resembles NIR then produces a larger gain. Together, the two directions of intervention place the effect at the interface between the input representation and the learned source decision function.

The baseline comparison places this effect on a practical scale (Table 4). Graham normalisation, CLAHE, AdaBN, TENT, FDA, SHOT, DANN, and CycleGAN reach external QWK 0.22, 0.29, 0.30, 0.42, 0.43, 0.45, 0.49, and 0.52; deterministic harmonisation reaches 0.54. The paired difference from CycleGAN is $+0.020$ with 95% CI $[-0.011, +0.051]$, compatible with equal grading performance. A simple known input intervention therefore recovers nearly the same target performance as learned cross-domain translation. Appendix G gives the source-channel ablation.

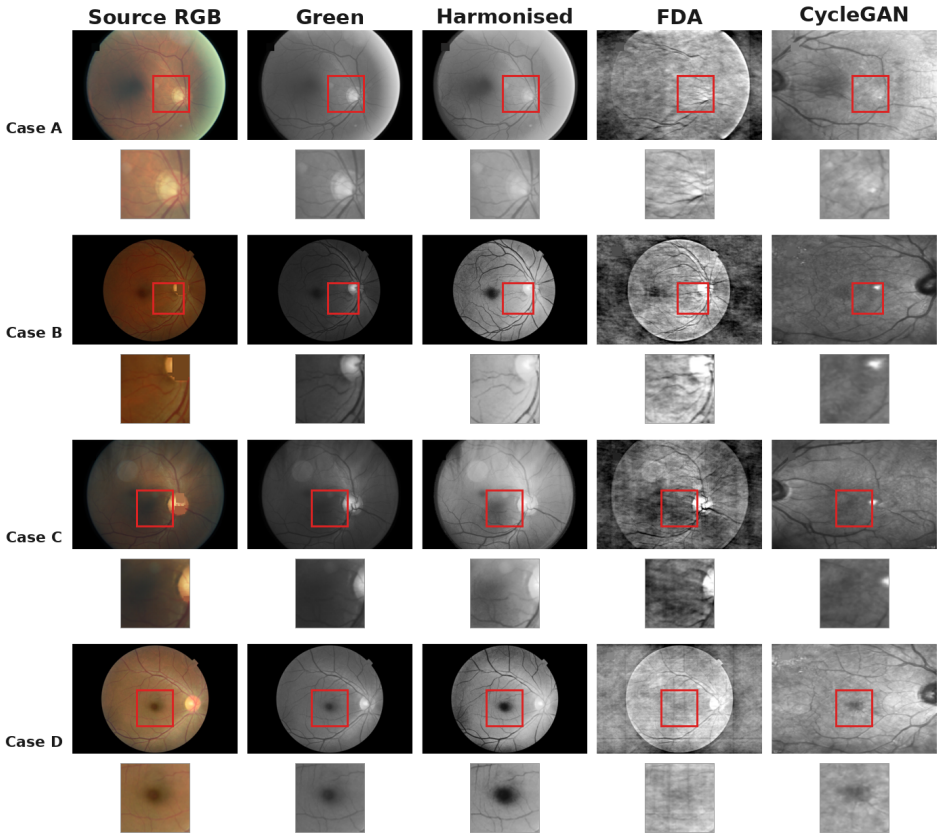

Figure 4: Qualitative comparison on four EyePACS cases. Columns show source RGB, replicated green, deterministic harmonisation, FDA, and CycleGAN using seed 47. Red boxes mark matched retinal regions and the lower strip enlarges the same region. The cases illustrate how each transform changes local appearance while preserving the source coordinate frame.

Target QWK alone does not reveal whether the transformed image still preserves the local pathology on which grading depends. In the blinded 200-image audit (Table 7, Appendix E), every source lesion remains identifiable in 91.0% of CycleGAN outputs, and 10.5% contain at least one new lesion-like focus. Deterministic harmonisation gives corresponding rates of 98.5% and 1.5%. Figure 4 shows matched examples. Because one specialist reviewed 200 source images, these numbers characterise the audited outputs rather than all possible CycleGAN training runs.

4.5 Direct Target Grading and Clinical Proxies

The direct target labels show that the recovery extends beyond OLIVES proxies. On NIR-External, harmonisation raises ResNet-50 QWK from 0.08 ± 0.03 to 0.54 ± 0.02, while referable-DR AUROC rises from 0.58 ± 0.03 to 0.81 ± 0.02 and macro-F1 from 0.16 ± 0.02 to 0.52 ± 0.02. The representative confusion matrices in Figure 5 move toward the reference diagonal for grades 0, 2, 3, and 4, although grade 1 remains largely confused with grades 0 and 2 (Section 4.6). The two external cSLO systems show nearly identical QWK gains, from 0.07 to 0.53 and from 0.09 to 0.55 (Appendix F).

The OLIVES structural proxies move in the same direction using information independent of the external graders. Predicted-grade/CST correlation changes from -0.47 ± 0.04 to $+0.33 \pm 0.03$, and the macular-fluid composite reaches AUROC 0.76 ± 0.02 (Table 4). After harmonisation, the CST association is stronger in the upper confidence tertile ($\rho = 0.472 \pm 0.024$) than in the lower tertile (0.152 ± 0.032; Appendix D). These measurements do not provide DR ground truth, but they offer an independent structural check on the direction of recovery.

The same representation also retains measurable information about individual retinal findings. Patient-grouped five-fold out-of-fold probes on the 192 tier-2 OLIVES images reach macro-AUROC 0.71 [0.67, 0.75], with intraretinal and subretinal fluid strongest and several vitreous or haemorrhagic

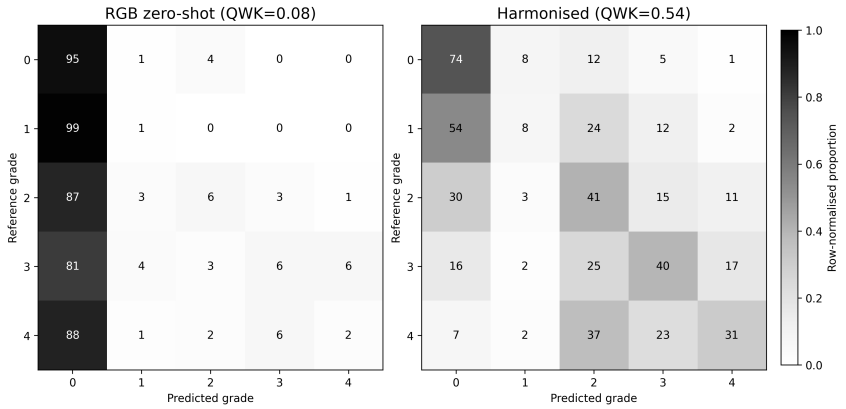

Figure 5: Direct grading on the NIR-EXTERNAL cohort. Row-normalised confusion matrices for the representative ResNet-50 seed (the seed whose external QWK is closest to the five-seed mean). Rows are specialist reference grades 0–4, with final adjudicated supports $n = 146, 92, 179, 111, 84$; columns are predicted grades, and cells show row percentages. Left: the RGB-trained grader applied zero-shot assigns 81–99% of every reference grade to grade 0. Right: after harmonisation the diagonal is populated for grades 0, 2, 3 and 4; grade 1 remains mostly assigned to grades 0 and 2. Panel titles give the QWK for this seed.

labels weaker (Appendix I). These probes are fitted on frozen encoder features and do not update the DR grader.

4.6 RESOLUTION AND MILD DISEASE

One residual failure is concentrated in mild disease. At 224 pixels, grade 1 remains rarely recovered after harmonisation, suggesting that the smallest lesions may be disproportionately affected by the source resolution. Repeating the harmonised ResNet-50 experiment at 384 and 512 pixels raises grade-1 recall from 0.11 to 0.24 and 0.33, while mean recall over grades 2–4 changes only from 0.61 to 0.64 and external QWK from 0.54 to 0.57 (Figure 7, Appendix J). This selective improvement is consistent with a lesion-scale limitation, although the experiment does not exclude other explanations.

5 DISCUSSION

5.1 WHAT THE DECOMPOSITION ESTABLISHES

The results change the interpretation of colour-to-NIR failure. What appears to be lost disease knowledge is, to a substantial extent, sensitive to how the target is presented to the colour-trained model. Channel representation and marginal intensity alignment each recover performance, their combination recovers more, the unchanged target-side control moves a fixed decision function in the same direction, and harmonisation reduces modality separability throughout the encoder. A substantial part of the failure therefore arises before learned cross-domain mapping is introduced.

Stronger baselines recover progressively more of the gap, but CycleGAN and deterministic harmonisation remain statistically compatible in grading. The blinded 200-image audit nevertheless shows lower lesion retention and more new lesion-like foci in the audited CycleGAN outputs. Thus the result is not simply that a simple transform beats a complex one; rather, a large part of the target deficit is already exposed by a transparent input intervention, leaving learned adaptation to address the residual.

The QWK difference between deterministic harmonisation and CycleGAN is small and its 95% confidence interval includes zero, so the study establishes no grading advantage for either method. The lesion-fidelity audit measures a separate property and favours the deterministic transform on the audited images: grading evaluates the downstream decision, whereas fidelity asks whether local pathological evidence is preserved.

Direct and proxy evaluations provide complementary evidence. NIR-EXTERNAL shows ordinal recovery against adjudicated DR grades, while OLIVES shows renewed association with CST and macular-fluid structure. These proxies support but do not replace direct grading. Clinical readiness

also remains untested: the best NIR result is below the in-domain colour grader, and no prospective evaluation has been performed.

Pre-training narrows the initial gap, especially for retinal foundation models, but does not remove the harmonisation gain. The residual is also structured: grade-1 recall improves much more than later-grade recall as resolution increases, consistent with a lesion-scale limitation without uniquely identifying it.

5.2 Shift Detection Is a Separate Problem

Shift detection reveals a second separation. Confidence supports selective prediction on EyePACS but does not reliably signal the modality boundary and can even increase on NIR. Feature distance instead detects the shift, decreases after harmonisation, and remains informative about residual errors. Grading, selective prediction, and acquisition-shift detection should therefore be validated as distinct functions rather than inferred from one confidence score.

5.3 Robustness and Scope

The architecture sweep separates source accuracy from transfer robustness. ImageNet ViT, DINOv2, RETFound, and VisionFM begin closer to useful NIR behaviour than ResNet-50 but still gain substantially from harmonisation. Within RETFound, OCT-pretrained weights transfer better zero-shot than fundus-pretrained weights despite slightly lower EyePACS QWK. Source-domain accuracy therefore does not determine cross-modality robustness; the relevant pre-training factor remains unresolved.

The decomposition is portable as an experimental form, not as a claim about all modalities: it needs a source model, an unlabelled target sample, and interventions that vary one input factor at a time. The present evidence is specific to colour-fundus-to-NIR transfer. Its broader use is diagnostic—test input compatibility first, then determine what residual remains for more complex adaptation.

6 Limitations

We study one disease and one modality transition, and EyePACS and NIR-External differ in cohort and device; the residual after harmonisation therefore cannot be assigned to acquisition physics alone. A paired colour/NIR cohort with common grades would separate these effects more cleanly. The external cohort is retrospective, and the lesion-fidelity audit uses one specialist and 200 images. The transform cannot restore spatial information absent from acquisition. EyePACS is also strongly class-imbalanced, which may contribute to weak grade-1 recall.

Harmonisation also assumes access to an unlabelled sample from the target acquisition because the reference histogram is target-informed. Appendix C shows similar results when this reference is estimated from 50, 200, or all available OLIVES training images, but a very small or drifting deployment sample may behave differently. Prospective referral decisions and post-intervention calibration are not evaluated. These constraints bound the conclusions to the tested setting.

7 Conclusion

Colour-to-NIR DR transfer does not fail as a single phenomenon. Across seeds and encoder families, a substantial part of the gap is recovered by matching source representation to the target, while a residual remains after that correction. The frozen-model control shows that input statistics can move a fixed decision function, and layer-wise probes show that harmonisation also reduces modality separability inside the encoder. Stronger pre-training narrows but does not remove this effect. CycleGAN reaches statistically comparable grading performance, although lesion preservation is weaker in the blinded audit. Predictive uncertainty remains weak at the modality boundary, whereas feature distance detects both shift and residual error. The resulting diagnostic is simple: determine how much failure is addressable at the input before assigning the remainder to adaptation, acquisition changes, or new labelled data.

AI USE STATEMENT

AI assistants were used for wording refinement, LaTeX troubleshooting, and debugging some Bash scripts. They were not used to generate experimental results, formulate the method, or make final scholarly judgments. The authors take full responsibility for all claims, citations, data, analyses, and submitted material.

ETHICS STATEMENT

EyePACS and OLIVES are publicly available, de-identified datasets. The NIR-EXTERNAL cohort consists of retrospectively collected, de-identified images from routine diabetic eye care. Its collection and specialist grading were approved by the authors' institutional research ethics committee. The institution and approval reference are withheld for double-blind review and will be stated in the camera-ready version. The harmonised models are research artefacts and are not intended for clinical use. The best external QWK remains below that of the in-domain colour-fundus grader, and no prospective evaluation has been performed.

REPRODUCIBILITY STATEMENT

EyePACS and OLIVES are public. Splits, seeds, optimisation and the statistical protocol are described in Section 3.2 and Appendix A; the harmonisation transforms and controls are specified in Section 3.3, with reference-histogram sensitivity in Appendix C. Appendix B documents the external grading/adjudication protocol, Appendix E documents the lesion-fidelity audit, and Appendix A reports the OOD and confidence-stratification settings. Code and frozen checkpoints will be publicly released.

REFERENCES

Murat Seçkin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning (MIDL)*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Camila González, Karol Gotkowski, Moritz Fuchs, Andreas Bucher, Armin Dadras, Ricarda Fischbach, Isabel Jasmin Kaltenborn, and Anirban Mukhopadhyay. Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation. *Medical Image Analysis*, 82:102596, 2022. doi: 10.1016/j.media.2022.102596.

Ben Graham. Kaggle diabetic retinopathy detection competition report. Technical report, University of Warwick, 2015.

Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016. doi: 10.1001/jama.2016.17216.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Olle G. Holmberg, Niklas D. Köhler, Thiago Martins, Jakob Siedlecki, Tina Herold, Leonie Keidel, Ben Asani, Johannes Schiefelbein, Siegfried Priglinger, Karsten U. Kortüm, and Fabian J. Theis. Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. *Nature Machine Intelligence*, 2(11):719–726, 2020. doi: 10.1038/s42256-020-00247-1.

Kaggle. Diabetic retinopathy detection competition. `https://www.kaggle.com/c/diabetic-retinopathy-detection`, 2015.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 7167–7177, 2018.

Livie Yumeng Li, Vajira Thambawita, Stine Byberg, and Adam Hulman. Assessing the generalisability of foundation models to ultra-wide field retinal imaging for diabetic retinopathy screening in Denmark and Greenland. *International Journal of Medical Informatics*, 217:106503, 2026. doi: 10.1016/j.ijmedinf.2026.106503.

Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pp. 6028–6039, 2020.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Jay Nandy, Wynne Hsu, and Mong Li Lee. Distributional shifts in automated diabetic retinopathy screening. In *IEEE International Conference on Image Processing (ICIP)*, pp. 255–259, 2021. doi: 10.1109/ICIP42928.2021.9506181.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13991–14002, 2019.

Mohit Prabhushankar, Kiran Kokilepersaud, Yash-yee Logan, Stephanie Trejo Corona, Ghassan AlRegib, and Charles Wykoff. OLIVES dataset: ophthalmic labels for investigating visual eye semantics. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9201–9216, 2022. Dataset: doi:10.5281/zenodo.7105232.

Jianing Qiu, Jian Wu, Hao Wei, Peilun Shi, Minqing Zhang, Yunyun Sun, Lin Li, Hanruo Liu, Hongyi Liu, Simeng Hou, et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI*, 1(12), 2024. doi: 10.1056/AIoa2300221.

Jinye Ran, Guanghua Zhang, Ximei Zhang, Juan Xie, Fan Xia, and Hao Zhang. Source-free active domain adaptation for diabetic retinopathy grading based on ultra-wide-field fundus images. *Computers in Biology and Medicine*, 174:108418, 2024. doi: 10.1016/j.compbiomed.2024.108418.

Emese Sükei, Elisabeth Rumetshofer, Niklas Schmidinger, Andreas Mayr, Ursula Schmidt-Erfurth, Günter Klambauer, and Hrvoje Bogunović. Multi-modal representation learning in retinal imaging using self-supervised learning for enhanced clinical predictions. *Scientific Reports*, 14:26802, 2024. doi: 10.1038/s41598-024-78515-y.

Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In *Computer Vision – ECCV 2016 Workshops*, pp. 443–450, 2016. doi: 10.1007/978-3-319-49409-8_35.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, pp. 20827–20840, 2022.

Mingxing Tan and Quoc V. Le. EfficientNet: rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, pp. 6105–6114, 2019.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4921–4930, 2022.

Fan Xiao, Junlin Hou, Ruiwei Zhao, Rui Feng, Haidong Zou, Lina Lu, Yi Xu, and Juzhao Zhang. Cross-fundus transformer for multi-modal diabetic retinopathy grading with cataract. *arXiv preprint arXiv:2411.00726*, 2024.

Wenwen Xue, Juzhao Zhang, Yingyan Ma, Junlin Hou, Fan Xiao, Rui Feng, Ruiwei Zhao, and Haidong Zou. Deep learning-based analysis of infrared fundus photography for automated diagnosis of diabetic retinopathy with cataracts. *Journal of Cataract & Refractive Surgery*, 49(10): 1043–1048, 2023. doi: 10.1097/j.jcrs.0000000000001269.

Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4085–4095, 2020.

Karina Zadorozhny, Dennis Ulmer, and Giovanni Cinà. Failures of uncertainty estimation on out-of-distribution samples: experimental results from medical applications lead to theoretical insights. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2021.

Yukun Zhou et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622:156–163, 2023. doi: 10.1038/s41586-023-06555-x.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

Karel Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics Gems IV*, pp. 474–485. Academic Press, 1994.

A Training Details and Statistical Protocol

Every optimisation batch for the DR encoder and grading head contains EyePACS training images only. OLIVES images enter no source-training batch and no OLIVES label contributes a gradient. EyePACS is grouped by patient before splitting, preventing fellow-eye leakage by construction. All source-training conditions use the same patient partitions and the paired seed set $\{7, 17, 27, 37, 47\}$. Optimisation uses AdamW (Loshchilov & Hutter, 2019), linear warm-up and cosine annealing, with identical augmentation schedules within an architecture. The grading objective is unweighted five-class cross-entropy. No class-balanced sampler or class weighting is used. The full harmonisation reference histogram is estimated once per split from the OLIVES training patients and is reused unchanged for every evaluation image. Model selection uses EyePACS validation QWK. All in-domain costs reported in the main paper are measured on the held-out EyePACS test set, not on the selection split.

Layer-wise domain probe. Domain-separability analysis is performed on held-out representations rather than on the samples used to fit the DR grader. Source and target examples are balanced for the linear probe. OLIVES visits retain the eye/patient grouping used for the main split so that repeated longitudinal visits do not cross the probe analysis boundary. The same probe construction is used for RGB and harmonised models at every reported encoder stage.

Adaptation baselines. All target-aware baselines start from the corresponding source split/checkpoint rule and use the OLIVES training partition only. AdaBN freezes learned weights and re-estimates BatchNorm running statistics from unlabelled OLIVES training images. TENT starts from the source checkpoint and adapts normalisation parameters by prediction-entropy minimisation, with other model parameters fixed. SHOT keeps the source hypothesis/classifier fixed while adapting target features. FDA performs low-frequency Fourier amplitude replacement using source and OLIVES-training images; DANN uses adversarial source/target feature alignment. Adapted models are frozen before NIR-EXTERNAL evaluation. NIR-EXTERNAL is not used to choose baseline hyperparameters.

CycleGAN implementation. For the unpaired colour-to-NIR translator, each seed is trained with 1,054 EyePACS source images and the 1,054 OLIVES training images, using batch size 1, instance normalisation, 256-pixel inputs, 20 constant-learning-rate epochs followed by 20 decay epochs, and the standard cycle-consistency weights. Because the translation is RGB-to-single-channel NIR, the identity-loss weight is set to zero. The five seeds are trained independently and evaluated on the same held-out EyePACS cases used for the qualitative stability analysis.

For uncertainty and OOD scoring, TTA uses five deterministic views (identity, horizontal flip, vertical flip, and $\pm 90^\circ$ rotations). Ensemble disagreement is computed from the predictions of the five independently seeded source models. Energy uses temperature 1. The ReAct-based score clips penultimate features at the 90th percentile estimated from EyePACS training features before OOD scoring. ViM uses a 256-dimensional principal subspace. kNN uses $k = 20$ on ℓ_2-normalised source-training features. Mahalanobis distance uses class-conditional source-training means with a Ledoit–Wolf shrinkage covariance. All score-fitting statistics are estimated from EyePACS training features only. For confidence-stratified analyses, upper and lower tertiles are defined separately for each seed from that seed's harmonised-model TTA-confidence distribution on OLIVES; the middle tertile is excluded.

For seed-level comparisons we report mean and sample standard deviation. Correlations and AUROCs on clinical cohorts use 2,000 cluster-bootstrap resamples at the patient/eye level, preserving repeated visits within a cluster. Direct grading QWK on NIR-EXTERNAL is bootstrapped by patient. Because all transformations are paired by seed, we additionally inspect the per-seed difference. No reported conclusion depends on a single best run. For the 2×2 decomposition, main effects are the average marginal contrasts across the two levels of the other factor and the interaction is the difference-in-differences. Their confidence intervals use a two-level paired factorial bootstrap: each replicate resamples the five paired seed IDs with replacement and, within each sampled seed, resamples the 306 NIR-EXTERNAL patients with replacement, using the same patient resample for all four factorial cells. The CycleGAN comparison uses a two-level paired bootstrap. Each replicate resamples the five paired seed IDs with replacement and, within each sampled seed, resamples the 306 NIR-EXTERNAL

patients with replacement while keeping both images from a patient together. The same patient sample is used for CycleGAN and deterministic harmonisation. QWK is recomputed for each method within that seed/resample, the within-seed paired difference is calculated, and differences are averaged across sampled seeds. The reported 95% CI is the percentile interval over 2,000 replicates.

B External NIR Grading and Adjudication Protocol

NIR-External is a locked retrospective evaluation cohort of 612 images from 306 patients acquired in routine diabetic eye care on two cSLO systems not represented in OLIVES. Every image was graded independently by two retinal specialists on the five-level International Clinical DR scale (0–4). Agreement between the two primary readers before adjudication was QWK $= 0.83$. Every disagreement was adjudicated by a third reader, and the adjudicated grade is used as the reference label for all reported model evaluations. Final reference-grade counts are 146, 92, 179, 111 and 84 for grades 0–4, respectively. No image from this cohort contributes to model training, checkpoint or hyperparameter selection, adaptation, or estimation of the NIR reference intensity distribution. QWK, macro-F1 and referable-DR AUROC are computed against the adjudicated reference grades; bootstrap resampling is performed at the patient level.

C Reference-Histogram and CORAL Sensitivity

The train-only target reference is stable to its sample size. Using 50, 200 or all available OLIVES-training images to estimate the reference produces external ResNet-50 QWK of 0.51, 0.53 and 0.54 respectively, with corresponding CST correlations of 0.30, 0.32 and 0.33. The similar values across sample sizes support the stability of the train-only reference histogram. Adding a CORAL term to the primary model changes external QWK by 0.01 and CST correlation by 0.01 (Table 3), so target-feature alignment is not required for the main effect.

D OLIVES Proxy Metrics for the Primary Backbones

Table 5: OLIVES proxy metrics for the three primary backbones (mean ± SD over five paired seeds). Columns as in Tables 1 and 4. Best value per column in bold, second best underlined.

Backbone	Source representation	EyePACS QWK	OLIVES G0	CST ρ	Fluid AUROC	Ext. QWK
ResNet-50	RGB	0.641 ± 0.006	95.8 ± 0.9	-0.47 ± 0.04	0.54 ± 0.03	0.08 ± 0.03
ResNet-50	Harmonised	0.629 ± 0.007	53.0 ± 2.6	$+0.33 \pm 0.03$	0.76 ± 0.02	0.54 ± 0.02
ConvNeXt-T	RGB	0.673 ± 0.005	91.2 ± 1.4	-0.31 ± 0.05	0.59 ± 0.03	0.16 ± 0.03
ConvNeXt-T	Harmonised	0.660 ± 0.006	$\underline{46.5 \pm 2.4}$	$\underline{+0.39 \pm 0.03}$	$\underline{0.79 \pm 0.02}$	$\underline{0.57 \pm 0.02}$
RETFound (fundus)	RGB	$\mathbf{0.704 \pm 0.004}$	73.4 ± 2.1	$+0.14 \pm 0.04$	0.67 ± 0.03	0.35 ± 0.02
RETFound (fundus)	Harmonised	$\underline{0.694 \pm 0.005}$	$\mathbf{39.6 \pm 2.0}$	$\mathbf{+0.44 \pm 0.03}$	$\mathbf{0.81 \pm 0.02}$	$\mathbf{0.62 \pm 0.02}$

Table 6: Per-seed OLIVES associations for ResNet-50. OLIVES G0 is the zero-shot grade-0 assignment. The non-zero prediction tail is its complement. Darkness ρ is the Spearman correlation between predicted grade and mean image intensity within that tail, computed separately for each seed. The tail comprises 41–68 images per seed. High- and low-confidence CST ρ are the Spearman correlations between harmonised predicted grade and central subfield thickness within the upper and lower confidence tertiles defined in Section 3.2. All summary values are mean ± SD across seeds.

	Zero-shot RGB model			Harmonised model	
Seed	OLIVES G0 (%)	Tail n	Darkness ρ	High-conf. CST ρ	Low-conf. CST ρ
7	94.6	68	-0.39	0.44	0.11
17	95.4	58	-0.47	0.50	0.19
27	95.9	52	-0.42	0.46	0.13
37	96.3	47	-0.46	0.49	0.17
47	96.8	41	-0.44	0.47	0.16
Mean ± SD	95.8 ± 0.9	–	-0.436 ± 0.032	0.472 ± 0.024	0.152 ± 0.032

E LESION-FIDELITY AUDIT

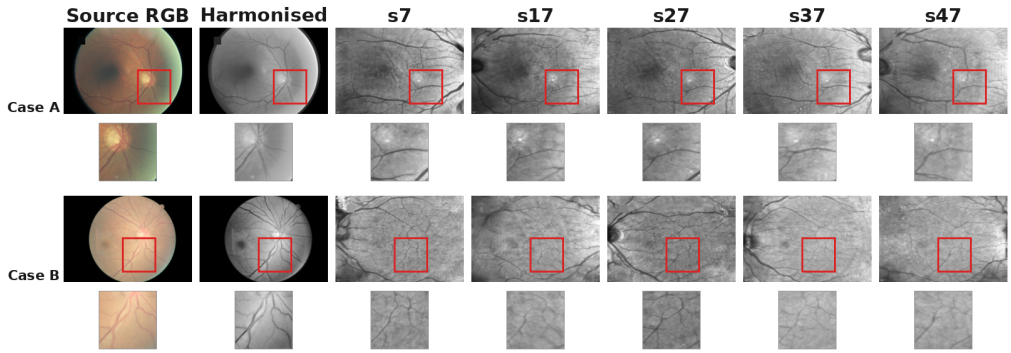

Figure 6: CycleGAN seed stability for two representative EyePACS cases. Source RGB and deterministic harmonisation are fixed. The remaining columns show independently trained CycleGAN translators for seeds 7, 17, 27, 37 and 47. The red box marks a matched local region, enlarged beneath each full image, making seed-to-seed variation in translated appearance directly visible.

The two-level paired-bootstrap QWK difference (deterministic minus CycleGAN; see Appendix A) is $+0.020$ with a 95% CI of $[-0.011, +0.051]$. Because the interval includes zero, these data do not establish superiority of either method. A retinal specialist reviewed the transformed image pairs for lesion preservation. The specialist did not participate in the primary grading or the adjudication of the NIR-EXTERNAL cohort and had no access to its DR grades. Transformation identity and downstream model predictions were concealed during the audit.

Table 7: Blinded specialist audit of pathology preservation on 200 EyePACS source images with visible haemorrhages, exudates or microaneurysm-like foci. All-lesions-retained is the image-level proportion for which every source lesion remained identifiable after transformation. New focus is the image-level proportion containing at least one lesion-like focus absent from the source. Brackets give Wilson 95% binomial confidence intervals. Best value per column in bold, second best underlined.

Transform	Ext. QWK ↑	Images retaining all source lesions ↑	Images with ≥1 new lesion-like focus ↓
FDA	0.43 ± 0.03	192/200 (96.0%) [92.3, 98.0]	7/200 (3.5%) [1.7, 7.0]
CycleGAN	$\underline{0.52 \pm 0.03}$	182/200 (91.0%) [86.2, 94.2]	21/200 (10.5%) [7.0, 15.5]
Deterministic harmonisation	$\mathbf{0.54 \pm 0.02}$	**197/200 (98.5%)** [95.7, 99.5]	**3/200 (1.5%)** [0.5, 4.3]

F PER-DEVICE RESULTS ON NIR-EXTERNAL

Table 8: Direct grading QWK on NIR-EXTERNAL by acquisition device for ResNet-50 (mean ± SD over five paired seeds). Neither device is represented in OLIVES.

Device	Images	Patients	RGB QWK	Harmonised QWK	ΔQWK
cSLO-A	318	159	0.07 ± 0.04	0.53 ± 0.03	$+0.46$
cSLO-B	294	147	0.09 ± 0.04	0.55 ± 0.03	$+0.46$
Overall	612	306	0.08 ± 0.03	0.54 ± 0.02	$+0.46$

G SOURCE-CHANNEL ABLATION

The channel ablation is an exploratory single-seed experiment using seed 27 and otherwise the same preprocessing, border masking, resolution and training protocol as Table 3. Differences of a few hundredths should not be read as meaningful. Each condition replicates one single-channel source representation across three channels and retrains the grader without intensity alignment. The final row adds marginal intensity alignment to the green channel. The green-only row is one realisation of the five-seed condition reported in Table 3 and not a separate experimental setting.

Table 9: Single-seed (seed 27) source-channel ablation for ResNet-50. Columns as in Table 3. Green gives the highest transfer among the replicated single channels in this exploratory run, but channel selection alone remains well below green with intensity alignment.

Source input representation	OLIVES G0 (%) ↓	CST ρ ↑	Ext. QWK ↑
RGB	95.9	-0.46	0.08
Blue replicated	89.2	-0.16	0.14
Red replicated	84.6	-0.08	0.20
Luminance replicated	82.1	-0.01	0.23
Green replicated	<u>79.6</u>	<u>$+0.03$</u>	<u>0.27</u>
Green + intensity alignment	**52.8**	$\mathbf{+0.33}$	**0.54**

H Full OOD Detection Results

For ReAct, the 95% patient/eye-level cluster-bootstrap intervals for the three quantities in Table 2 are $[0.79, 0.86]$ (shift detection, RGB grader), $[0.60, 0.68]$ (shift detection, harmonised grader) and $[0.66, 0.74]$ (exact-grade error detection).

Table 10: OOD detection under colour-to-NIR shift. AUROC treats OLIVES NIR images as the positive class against held-out EyePACS. All scores are computed from the same frozen RGB-trained ResNet-50. Ensemble disagreement uses the five seed models. 95% confidence intervals are from 2,000 patient/eye-level cluster-bootstrap resamples. The first four scores are derived from the output distribution. The remaining five from logits or penultimate features. Best value in bold, second best underlined.

Type	OOD score	AUROC ↑	95% CI
Output	Negative max. softmax probability	0.31	[0.27, 0.35]
Output	Predictive entropy	0.54	[0.50, 0.58]
Output	TTA variance	0.56	[0.52, 0.60]
Output	Deep-ensemble disagreement	0.59	[0.55, 0.63]
Logit	Energy	0.76	[0.72, 0.79]
Representation	ReAct-based score	0.83	[0.79, 0.86]
Representation	ViM	0.89	[0.86, 0.92]
Representation	kNN distance	<u>0.91</u>	[0.88, 0.93]
Representation	Mahalanobis distance	**0.93**	[0.90, 0.95]

I Per-Biomarker Out-of-Fold Results

Table 11: Patient-grouped five-fold out-of-fold biomarker detection on all 192 tier-2 OLIVES images. Positives gives the number of images annotated positive for each biomarker. AUROC is computed on the pooled out-of-fold predictions. 95% confidence intervals are from 2,000 patient/eye-level cluster-bootstrap resamples. Abbreviations: IRF, intraretinal fluid. SRF, subretinal fluid. EZ, ellipsoid zone. DRT/ME, diffuse retinal thickening or macular oedema. IR, intraretinal.

Biomarker	Positives	AUROC	95% CI
Fluid (IRF)	126 / 192	0.89	0.83–0.94
Fluid (SRF)	48 / 192	0.84	0.76–0.91
Vitreous debris	110 / 192	0.78	0.70–0.85
Disruption of EZ	62 / 192	0.73	0.65–0.80
DRT/ME	115 / 192	0.69	0.61–0.76
Partially attached vitreous face	86 / 192	0.65	0.57–0.72
Fully attached vitreous face	142 / 192	0.64	0.56–0.71
Preretinal tissue / haemorrhage	38 / 192	0.61	0.52–0.70
IR haemorrhages	93 / 192	0.58	0.50–0.66
Macro-average		0.71	0.67–0.75

J ADDITIONAL FIGURES

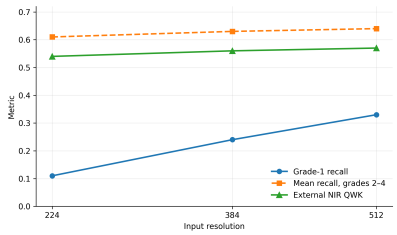

Figure 7: Effect of input resolution on the harmonised ResNet-50. The horizontal axis gives the input resolution in pixels (224, 384, 512) and the vertical axis gives the metric value. Grade-1 recall on NIR-EXTERNAL (circles) rises from 0.11 to 0.33. Mean recall across grades 2–4 (squares) and external QWK (triangles) change by 0.03 each. The improvement is concentrated in the grade whose defining lesions are smallest.

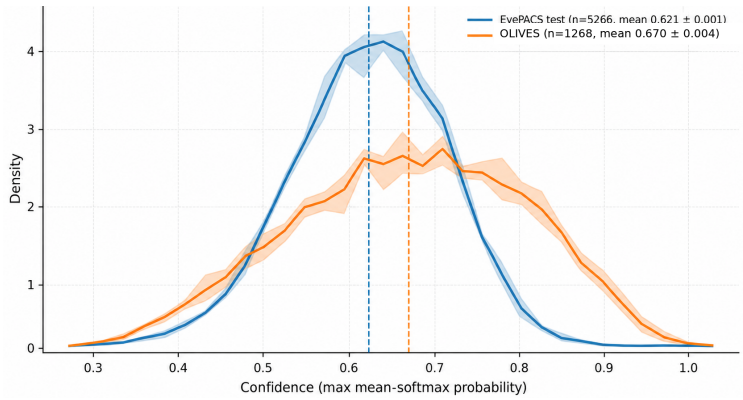

Figure 8: Distributions of TTA confidence (mean maximum softmax probability) on held-out EyePACS (blue) and on OLIVES (orange) for the source-only ResNet-50. Curves are the mean density across the five seeds, shaded bands the per-seed range, and dashed lines the five-seed mean confidence (means 0.621 and 0.670; SD over seeds 0.001 and 0.004). The colour grader is more confident on NIR input than in domain.

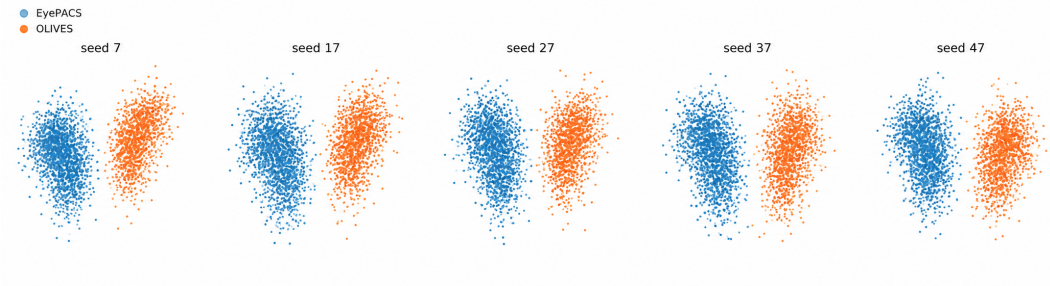

Figure 9: Two-dimensional UMAP projections of the RGB-trained ResNet-50 penultimate feature space, one panel per seed (7, 17, 27, 37, 47), coloured by dataset (blue for held-out EyePACS, subsampled; orange for all 1,268 OLIVES images). Each panel is a separate embedding, so positions are not comparable across panels, and the axes are UMAP components with no physical units. In every seed the two modalities form two separate clusters. This is the qualitative counterpart of the layer-wise domain-classifier result in Figure 3.

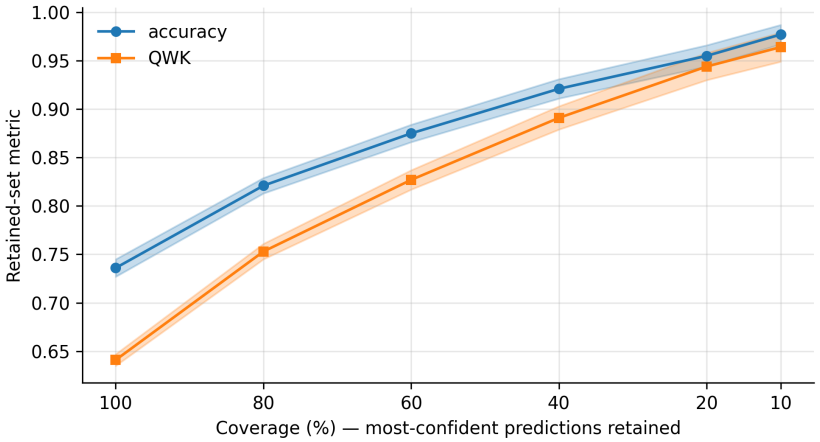

Figure 10: Selective prediction on the held-out EyePACS test set for the RGB-trained ResNet-50. The horizontal axis gives coverage, the percentage of most-confident predictions retained, from 100% on the left to 10% on the right. The vertical axis gives accuracy (blue circles) and QWK (orange squares) computed on the retained set. Lines are the mean over the five seeds and shaded bands are $\pm$ one SD. Both metrics rise as low-confidence predictions are abstained, so confidence tracks correctness in domain. Section 4.3 reports near-chance modality-shift detection from the same signal.