# Enabling Accessible Eye Screening: Privileged Bilateral Learning for Monocular Inference in Individuals with Disabilities

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Machine learning systems for medical imaging are often developed under data-rich conditions but deployed when only incomplete information is available. This challenge is especially relevant for individuals with vision impairment or disabilities that make bilateral imaging difficult or impossible. In colour fundus photography (CFP), both eyes are commonly imaged because they provide complementary information for diagnosing ocular diseases such as age-related macular degeneration (AMD) and diabetic retinopathy (DR). However, bilateral acquisition may not be feasible for people with certain physical or cognitive disabilities, or for those with only one eye. We propose a privileged learning framework in which a teacher model has access to bilateral CFP images during training, while a student model learns from the teacher using only a single-eye image. At inference, the student requires only the available eye and does not need the contralateral image. Experimental results show consistent improvements in AUROC over conventional monocular training, with the largest gains for left-eye images for AMD (0.931 vs. 0.966) and DR (0.789 vs. 0.836). By enabling models to benefit from bilateral information during training while requiring only a single eye at inference, our approach may support more inclusive and accessible screening for individuals with disabilities and in resource-constrained settings.

## 1 Introduction

Machine learning systems are often developed under data-rich conditions but deployed when only incomplete information is available. This mismatch is particularly important in resource-constrained settings, where acquisition failures, limited infrastructure, or operational constraints may prevent models from accessing all observations available during development.

A natural example arises in colour fundus photography (CFP), where both eyes are commonly imaged because the two eyes can provide complementary information for ocular disease assessment [11]. However, bilateral acquisition may not always be feasible, resulting in only a monocular image being available [10, 8]. This creates a distinction between models trained using monocular inputs and those trained using bilateral inputs. The former can operate when only one eye is available but cannot benefit from information contained in the contralateral eye. The latter can exploit information from both eyes but require both images at inference [5]. This motivates the question: *can information from bilateral observations available during training be transferred to a model that requires only a monocular image at inference?*

We study this problem in the context of retinal diseases that can be assessed from CFP. Leveraging publicly available datasets, we propose a teacher-student framework in which a teacher model has access to bilateral image inputs during training and learns representations informed by the complementary information provided by both eyes. A student model is subsequently trained to

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

transfer this knowledge while receiving only monocular inputs. At inference, the student operates exclusively on a single-eye image and does not require paired image inputs. Here, we refer to information available during training but unavailable at inference as privileged information; in our setting, the *contralateral*-eye image constitutes this privileged information.

Prior work has exploited bilateral fundus images for ocular disease classification by weighting and fusing information from both eyes [5, 1]. Such approaches, however, require bilateral images at inference. Related work has also explored learning from privileged information that is available during training but unavailable at deployment. For example, Dual et al. [3] incorporate privileged genetic information into fundus-image-based glaucoma detection, while Korevaar et al. [6] perform macular disease severity classification using optical coherence tomography (OCT) images and incorporate fluid segmentation masks as privileged information to improve domain generalization. In contrast, we investigate the contralateral eye as privileged information, with bilateral observations available during training but only a single-eye image required at inference.

## 2 Methods

We developed a privileged distillation framework built on RETFound-Green, a pretrained retinal foundation model based on the Vision Transformer (ViT) architecture [4]. With 22.2 million parameters, RETFound-Green offered a compact architecture with competitive performance for retinal image analysis, making it well suited for our resource-conscious setting.

Our proposed framework consists of a teacher and a student model. The teacher model was trained on paired left and right images, learning features from both eyes and fusing them via mean pooling to form a privileged bilateral representation that was subsequently passed to a linear classifier. A distillation loss guided the student to learn features that aligned with the teacher's richer bilateral representation. The overall training objective combined the binary cross-entropy (BCE) classification losses of both networks with a feature-level distillation loss, given by the third term:

$$\mathcal{L} = \text{BCE}(\hat{y}_{teacher}, y) + \text{BCE}(\hat{y}_{student}, y) + \alpha \frac{1}{B} \sum_{i=1}^{B} ||\mathbf{Z}^i_{student} - \mathbf{Z}^i_{teacher}||_2^2, \tag{1}$$

where $\mathbf{Z}^i_{student} \in \mathbb{R}^{384}$ and $\mathbf{Z}^i_{teacher} \in \mathbb{R}^{384}$ were the feature representations from the second-to-last layer of each network, extracted from the monocular images and the fused bilateral representation, respectively. The hyperparameter $\alpha$ balanced the classification and distillation objectives.

### 2.1 Model Architectures

We considered three model configurations in this study:

**Baseline** ($M_{\text{baseline}}$): This model was trained to classify monocular fundus images, optimized solely via the binary cross-entropy loss.

**Privileged Teacher** ($M_{\text{teacher}}$): This model processed paired left and right fundus images, extracted features from each eye, and fused them via mean pooling to form a bilateral representation $\mathbf{Z}_{\text{teacher}}$. This representation was passed to a linear classifier for disease prediction. It was trained using only the classification loss.

**Privileged Student** ($M_{\text{student}}$): This model used the same architecture as the baseline and was trained jointly with the teacher. Like the baseline, it received monocular images to learn feature representations ($\mathbf{Z}_{\text{student}}$) and was optimized using a combination of classification and feature-level distillation losses that encouraged $\mathbf{Z}_{\text{student}}$ to align with the teacher's bilateral representation $\mathbf{Z}_{\text{teacher}}$. At inference, $M_{\text{student}}$ operated independently, incurring no additional computational cost compared with the baseline.

For a controlled comparison, the baseline and privileged student models used the same architecture (23.4M parameters) and identical training data.The privileged student additionally received a distillation gradient that encouraged alignment with the teacher's bilateral representation. Using a linear classifier minimized architectural differences, ensuring that performance gains could be attributed to privileged information rather than model capacity or data volume. Linear probing (Section ) further assessed whether these gains were reflected in the learned backbone representations.

### 2.2 Experimental Setup

**Dataset.** We evaluated our proposed method using the Ocular Disease Intelligent Recognition (ODIR-5K) dataset [9] for development and validation.

**Data split.** To obtain unbiased performance estimates, we partitioned the ODIR-5K dataset at the patient level, ensuring that images from the same patient never appeared in more than one data subsets. We first performed a 6-fold split to separate a held-out test set (20% of patients). From the remaining patients, we performed a second 5-fold split to create a training set (80%) and a validation set (20%). This hierarchical, patient-aware splitting strategy prevented data leakage and evaluated the model's ability to generalize to new patients.

The validation set was used for (1) early stopping during training and (2) threshold calibration. At the end of each epoch, we computed the binary cross-entropy loss on the validation set and saved the model checkpoint with the lowest validation loss. For threshold calibration, we selected the optimal decision threshold $t^*$ that maximized the macro F1-score:

$$t^* = \arg \max_{t \in [0.01, 0.99]} \mathrm{F1}_{\text{macro}}(y_{\text{val}}, \mathbf{1}[p_{\text{val}} > t]).$$

where $\mathbf{1}$ is the indicator function. This threshold was then applied to the held-out test set predictions. All final metrics reported in this work were computed on the test set, which remained completely unseen during training, model selection, and threshold calibration.

**Performance evaluation.** We computed Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision–Recall Curve (AUPRC), Balanced Accuracy, Precision, and Recall, alongside false positive rate (FPR) to transparently capture task-specific clinical trade-offs.

### 2.4 Mechanistic Analysis of Distillation

In addition to the primary comparative analysis, we conducted complementary analyses to further examine how privileged bilateral information influences the learned representations:

- Exp. 1 Does the student representation become more similar to the bilateral teacher's representation?
- Exp. 2 Does the student's frozen representation contain more linearly accessible disease-relevant information?

**Exp. 1: Representation similarity.** We used Centered Kernel Alignment (CKA) [7] to quantify the structural similarity between feature representations:

$$\mathrm{CKA}(\mathbf{Z}_A, \mathbf{Z}_B) = \frac{\mathrm{HSIC}(\mathbf{Z}_A, \mathbf{Z}_B)}{\sqrt{\mathrm{HSIC}(\mathbf{Z}_A, \mathbf{Z}_A) \cdot \mathrm{HSIC}(\mathbf{Z}_B, \mathbf{Z}_B)}}, \tag{2}$$

where HSIC denotes the Hilbert-Schmidt Independence Criterion. We extracted features from the validation set and computed CKA for two comparisons:

$$\Delta\mathrm{CKA} = \mathrm{CKA}\left(\mathbf{Z}_{\text{student}}, \mathbf{Z}_{\text{teacher}}\right) - \mathrm{CKA}\left(\mathbf{Z}_{\text{baseline}}, \mathbf{Z}_{\text{teacher}}\right). \tag{3}$$

A higher change in CKA for the privileged student indicates that the distillation objective successfully aligned the student's representational geometry with the teacher's bilateral reasoning.

**Exp. 2: Feature Quality via Linear Probing.** To determine whether the teacher improved the intrinsic quality of the student's features (beyond the final classification head), we performed linear probing [2]. We froze the backbones of both the baseline and privileged student models and replaced their classification heads with a fresh, randomly initialized linear layer. We then trained *only* this new linear layer on the training set for 20 epochs, while keeping the backbones entirely frozen. The performance (AUROC) of this probe was evaluated on the test set. If the privileged backbone achieves a higher AUROC than the baseline backbone under this protocol, it provides direct evidence that the teacher improved the linear separability and quality of the student's learned features.

## 3 Results & Discussion

**Performance & trade-offs.** Table 1 summarizes the results across all experimental trials. We report 95% confidence intervals for all metrics, estimated via bootstrapping on the held-out test set to quantify sampling variability. The privileged-distillation framework improved AUROC and balanced accuracy for the left eye (0.931 vs. 0.966 and 0.834 vs. 0.879, respectively), indicating improved ranking performance and class-balanced accuracy. However, this improvement was accompanied by a higher false-positive rate (0.011 vs. 0.027) and lower precision (0.760 vs. 0.595).

This trade-off reflects a shift toward higher sensitivity (recall: 0.679 vs. 0.786), which may be desirable in screening applications where false negatives can be more costly than false positives [12].

Table 1: Performance metrics with 95% confidence intervals for AMD (top) and DR (bottom). Bold indicates the best result for each eye and task. The privileged framework improves AUROC and balanced accuracy across both tasks, with task-specific trade-offs: increased recall for AMD (catching more positives) and increased precision for DR (reducing false positives).

| Metric | Baseline (Left) | Privileged (Left) | Baseline (Right) | Privileged (Right) |
|---|---|---|---|---|
| **Task: Detection of Age-related Macular Degeneration (AMD)** | | | | |
| **AUROC** | 0.931 (0.846–0.984) | **0.966** (0.936–0.987) | 0.801 (0.679–0.890) | **0.810** (0.710–0.889) |
| **Balanced Accuracy** | 0.834 (0.744–0.924) | **0.879** (0.799–0.957) | 0.676 (0.585–0.774) | **0.689** (0.592–0.784) |
| **AUPRC** | **0.696** (0.520–0.838) | 0.651 (0.473–0.821) | **0.446** (0.259–0.642) | 0.413 (0.221–0.610) |
| **Precision** | **0.760** (0.588–0.917) | 0.595 (0.429–0.744) | **0.769** (0.500–1.000) | 0.579 (0.333–0.833) |
| **Recall** | 0.679 (0.500–0.857) | **0.786** (0.625–0.939) | 0.357 (0.176–0.552) | **0.393** (0.200–0.577) |
| **FPR** (↓ is better) | **0.011** (0.004–0.020) | 0.027 (0.014–0.041) | **0.005** (0.000–0.011) | 0.014 (0.005–0.025) |
| **Task: Detection of Diabetic Retinopathy (DR)** | | | | |
| **AUROC** | 0.789 (0.734–0.843) | **0.836** (0.799–0.872) | 0.825 (0.787–0.860) | **0.837** (0.801–0.871) |
| **Balanced Accuracy** | 0.576 (0.548–0.601) | **0.749** (0.712–0.788) | 0.740 (0.703–0.776) | **0.757** (0.719–0.791) |
| **AUPRC** | 0.511 (0.417–0.608) | **0.772** (0.717–0.824) | 0.759 (0.705–0.811) | **0.778** (0.724–0.827) |
| **Precision** | 0.078 (0.064–0.094) | **0.760** (0.688–0.831) | 0.741 (0.669–0.807) | **0.803** (0.733–0.867) |
| **Recall** | **0.917** (0.865–0.966) | 0.587 (0.516–0.659) | 0.577 (0.508–0.643) | **0.582** (0.511–0.650) |
| **FPR** (↓ is better) | 0.766 (0.744–0.787) | **0.089** (0.061–0.117) | 0.096 (0.069–0.127) | **0.068** (0.044–0.094) |

The increase in FPR remains modest in absolute terms (from 1.1% to 2.7%), while the improvement in balanced accuracy indicates improved aggregate performance across the two classes. Nevertheless, the accompanying reduction in precision highlights a potential cost of this increased sensitivity.

For DR, the privileged model similarly improved AUROC and balanced accuracy across both eyes, with the largest improvement observed for the left eye (AUROC: 0.789 vs. 0.836; balanced accuracy: 0.576 vs. 0.749).Unlike AMD, the left-eye DR model achieved substantially higher precision and lower false-positive rate at the expense of recall.

**Centered Kernel Alignment.** CKA analysis on the DR task revealed a modest increase in feature similarity (+0.015) between the privileged student and the bilateral teacher relative to the baseline. This finding is consistent with the hypothesis that the distillation objective encouraged the student to learn representations that more closely resemble those of the bilateral teacher, although the modest magnitude suggests that the effect on the overall representational geometry was limited. For AMD, the CKA analysis was technically inconclusive.

**Linear probing.** Linear probing revealed that the privileged student's frozen backbone achieved a +2.0 percentage-point AUROC improvement over the baseline backbone for AMD when trained with a new linear classifier (0.815 vs. 0.836). A similar +2.9 percentage-point improvement was observed for DR. These results indicate that disease-relevant information was more linearly accessible from the privileged student's frozen representations than from the baseline representations. This suggests that the benefits of privileged-information distillation extend beyond optimization of the original classification head and are reflected in the learned backbone representations.

**Limitations.** First, the relatively small test set limits the statistical power and precision of our estimates and therefore the strength of our generalizability claims. External validation is needed to assess robustness across diverse populations, acquisition conditions, and imaging systems. Second, while the proposed method improves classification performance, clinical deployment would require careful integration with existing screening workflows, including appropriate human oversight and evaluation of model interpretability. Finally, our experiments were conducted using a single dataset split. Evaluation across multiple random seeds and, where feasible, cross-validation would provide a more robust assessment of performance variability.

## 4 Conclusions

We introduced a privileged distillation framework for retinopathy detection using monocular images, where a teacher network trained on paired fundus images guides a student network to process only monocular images. The proposed method was evaluated against a baseline that treats each eye independently. Our framework offers a practical, deployable solution for DR screening, particularly in settings where only one fundus image is available. The observed trade-offs and asymmetries reveal insights for future research into multimodal and asymmetric medical imaging.

# References


[1] Shumoos Al-Fahdawi, Alaa S Al-Waisy, Diyar Qader Zeebaree, Rami Qahwaji, Hayder Natiq, Mazin Abed Mohammed, Jan Nedoma, Radek Martinek, and Muhammet Deveci. Fundus-deepnet: Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images. *Information Fusion*, 102:102059, 2024.

[2] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR) Workshop*, 2016.

[3] Lixin Duan, Yuan Yuan, Qiang Yang, et al. Incorporating privileged genetic information for fundus image based glaucoma detection. 2014. doi: 10.1007/978-3-319-10470-6_26. URL `https://link.springer.com/chapter/10.1007/978-3-319-10470-6_26`.

[4] Justin Engelmann and Miguel O Bernabeu. Training a high-performance retinal foundation model with half-the-data and 400 times less compute. *Nature Communications*, 16(1):6862, 2025.

[5] Junjun He, Cheng Li, Jin Ye, Shanshan Wang, Yu Qiao, and Lixu Gu. Classification of ocular diseases employing attention-based unilateral and bilateral feature weighting and fusion. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1258–1261. IEEE, 2020.

[6] Steven Korevaar, Ruwan Tennakoon, Ricky O'Brien, Dwarikanath Mahapatra, and Alireza Bab-Hadiashar. Domain generalization by learning from privileged medical imaging information. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

[7] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

[8] Fernando K Malerbi, Luis Filipe Nakayama, Gustavo Barreto Melo, José A Stuchi, Diego Lencione, Paulo V Prado, Lucas Z Ribeiro, Sergio A Dib, and Caio V Regatieri. Automated identification of different severity levels of diabetic retinopathy using a handheld fundus camera and single-image protocol. *Ophthalmology science*, 4(4):100481, 2024.

[9] Peking University. Ocular disease intelligent recognition (ODIR-2019). `https://odir2019.grand-challenge.org/`, 2019. Accessed: 2026-08-01].

[10] M Sangeetha, V Vikashini, G Sanmathi, P Anbumani, S Prabakaran, and S Geetha. Eye as a health window: an ai-driven ocular disease screening solution for Rural Karur, Tamil Nadu. In *2026 Third International Conference on Innovations in Cybersecurity and Data Science (ICICDS)*, pp. 1039–1045. IEEE, 2026.

[11] Lehan Wang, Chongchong Qi, Chubin Ou, Lin An, Mei Jin, Xiangbin Kong, and Xiaomeng Li. Multieye: Dataset and benchmark for oct-enhanced retinal disease recognition from fundus images. *IEEE Transactions on Medical Imaging*, 44(4):1711–1722, 2024.

[12] Yueye Wang, Chi Liu, Wenyi Hu, Lixia Luo, Danli Shi, Jian Zhang, Qiuxia Yin, Lei Zhang, Xiaotong Han, and Mingguang He. Economic evaluation for medical artificial intelligence: accuracy vs. cost-effectiveness in a diabetic retinopathy screening case. *NPJ Digital Medicine*, 7(1):43, 2024.

# A Definitions

## A.1 Performance Metrics

Let $TP$, $TN$, $FP$, and $FN$ denote the number of true positives, true negatives, false positives, and false negatives, respectively, at a given decision threshold.

**Precision and Recall (Sensitivity).**

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{4}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN}. \tag{5}$$

**Specificity and Balanced Accuracy.**

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{6}$$

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\text{Recall} + \text{Specificity}\right). \tag{7}$$

**Area Under the Receiver Operating Characteristic Curve (AUROC).** The AUROC is defined as the integral of the true positive rate (recall) against the false positive rate (FPR $= \frac{FP}{FP+TN}$) over all possible classification thresholds $t$:

$$\text{AUROC} = \int_0^1 \text{Recall}(t)\, d\text{FPR}(t). \tag{8}$$

AUROC ranges from 0 to 1, with 0.5 indicating random chance. It is insensitive to class imbalance and is the standard ranking metric.

**Area Under the Precision-Recall Curve (AUPRC).** The AUPRC is defined as the integral of precision against recall over all possible thresholds $t$:

$$\text{AUPRC} = \int_0^1 \text{Precision}(t)\, d\text{Recall}(t). \tag{9}$$

AUPRC is particularly informative for imbalanced datasets, as it focuses on the positive (minority) class and is more sensitive to false positives than AUROC. In a random classifier, AUPRC equals the prevalence of the positive class.

**Threshold Optimisation.** For Balanced Accuracy, Precision, and Recall, the decision threshold $t^*$ was optimised on the validation set to maximise the macro F1-score:

$$t^* = \arg\max_{t \in [0.01, 0.99]} \text{F1}_{\text{macro}}(y_{\text{val}}, \mathbb{1}[p_{\text{val}} > t]). \tag{10}$$

All metrics are reported with 95% confidence intervals estimated via bootstrapping (1,000 resamples) on the held-out test set.

## A.2 HSIC

Centered Kernel Alignment (CKA) relies on the Hilbert-Schmidt Independence Criterion (HSIC) to measure the dependence between two sets of features. Given two feature matrices $\mathbf{X} \in \mathbb{R}^{n \times d_1}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d_2}$ for $n$ samples, HSIC is defined as:

$$\text{HSIC}(\mathbf{X}, \mathbf{Y}) = \frac{1}{(n-1)^2} \operatorname{tr}\left(\tilde{\mathbf{K}}\tilde{\mathbf{L}}\right),$$

where:

- $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$ and $\mathbf{L} = \mathbf{Y}\mathbf{Y}^\top$ are the Gram matrices of $\mathbf{X}$ and $\mathbf{Y}$ (using the linear kernel, which measures cosine similarities between samples),

- $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ is the centering matrix, with $\mathbf{I}_n$ the identity matrix and $\mathbf{1}_n$ a vector of ones,
- $\tilde{\mathbf{K}} = \mathbf{HKH}$ and $\tilde{\mathbf{L}} = \mathbf{HLH}$ are the centered Gram matrices.

Intuitively, HSIC measures how much information is shared between two feature representations. A high HSIC value means that the representations vary together and therefore contain statistically related information, while a value close to zero suggests that they are largely independent. This makes HSIC useful for quantifying whether two learned representations capture overlapping or complementary information.

In practice, since the centering operation $\mathbf{HKH}$ is equivalent to first centering the features $\mathbf{X}$ and $\mathbf{Y}$ (i.e., subtracting the column-wise mean) and then computing their Gram matrices, we implement HSIC as:

$$\text{HSIC}(\mathbf{X}, \mathbf{Y}) = \frac{1}{(n-1)^2} \operatorname{tr}\left(\mathbf{K}_c \mathbf{L}_c\right),$$

where $\mathbf{K}_c$ and $\mathbf{L}_c$ are the Gram matrices computed from the centered features. This formulation matches the implementation used in our experiments and avoids numerical instabilities associated with explicit centering matrices.