Multimodal Retina Image Analysis Survey: Datasets, Tasks and Methods

Hongwei Sheng, **Heming Du**, **Xin Shen**, **Sen Wang** and **Xin Yu**
The University of Queensland
{firstname.lastname}@uq.edu.au

Abstract

Retina images provide a minimally invasive view of the central nervous system and microvasculature, making it essential for clinical applications. Changes in the retina often indicate both ophthalmic and systemic diseases, aiding in diagnosis and early intervention. While deep learning algorithms have advanced retina image analysis, a comprehensive review of related datasets, tasks, and benchmarking is still lacking. In this survey, we systematically categorize existing retina image datasets based on their available data modalities, and review the tasks these datasets support in multimodal retina image analysis. We also explain key evaluation metrics used in various retina image analysis benchmarks. By thoroughly examining current datasets and methods, we highlight the challenges and limitations in existing benchmarks and discuss potential research topics in the field. We hope this work will guide future retina analysis methods and promote the shared use of existing data across different tasks.

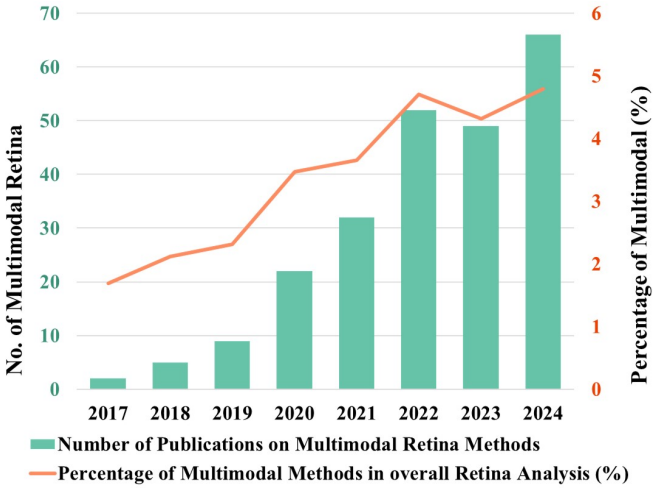

Figure 1: Publication trend on retinal image analysis from PubMed. The histogram indicates the number of total publications on multimodal retina analysis methods, while the curve indicates the percentage of multimodal based methods in overall retina analysis methods.

1 Introduction

The retina is a light-sensitive layer of tissue in the back of the eye that sends visual signals to the brain. As the only human tissue allowing direct, noninvasive visualization of the central nervous system and microvascular circulation, retina images play a vital role in clinical applications. Changes in the retina show signs of not only ophthalmic diseases, such as Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD), and GLaucoma (GL), but also systemic diseases, such as Alzheimer's disease (AD), cardiovascular diseases, and neurological diseases.

In practice, multiple imaging techniques are developed to examine various views of the retina. For example, Color Fundus photography (CF) provides views of the retina surface, Optical Coherence Tomography (OCT) delivers high-resolution cross-sectional images of retina layers, and Optical Coherence Tomography Angiography (OCTA) visualizes microvascular networks without the need for invasive contrast agents. By integrating these techniques, clinicians obtain detailed images of anatomical and vascular structures for precise diagnosis and early disease intervention.

With the increasing availability of retina datasets [Hassan *et al.*, 2022; Khan *et al.*, 2023; Li *et al.*, 2024], deep learning techniques [Liu and Yu, 2021; Wu *et al.*, 2021; Xie *et al.*, 2024; Chen *et al.*, 2024] have proven effective in retina images analysis tasks such as disease classification, cross-modal registration, and lesion segmentation in clinical applications. As demonstrated in Fig. 1, we present the publication trend of multimodal deep learning research in retinal image analysis, along with the increasing proportion of such studies relative to the overall retinal imaging research publications. The adoption of multimodal deep learning algorithms in ophthalmology follows an increasing trajectory, reflecting its growing clinical promise.

In this paper, we revisit the multimodal retina datasets, tasks, and corresponding methods and discuss the limitations and challenges in retina image analysis. To be specific, we first categorize publicly available datasets based on their imaging modalities and analyze their characteristics. We then review the primary multimodal retina image analysis tasks,

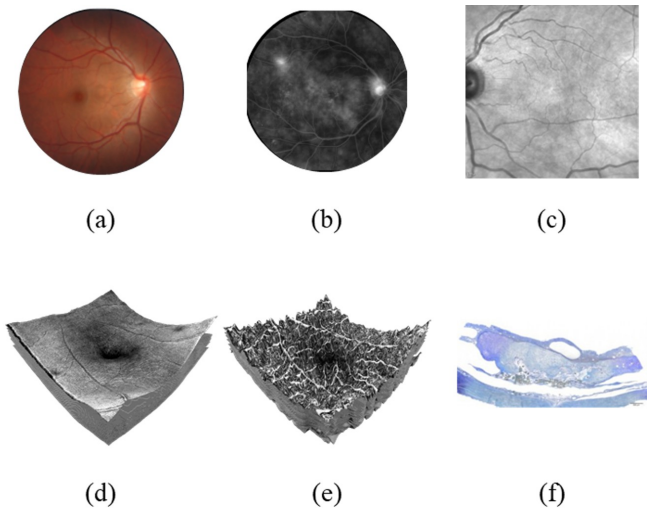

Figure 2: An overview of popular retinal imaging modalities. (a) Color fundus photography (CF); (b) Fundus angiography (FA); (c) Near infrared image (NIR); (d) Optical coherence tomograph (OCT); (e) OCT angiography (OCTA); (f) Histology image (HI).

including classification, grading, registration, segmentation, image generation, and medical report generation, highlighting the associated datasets, evaluation metrics, methods, and the benchmarks established for each task. In addition, we further highlight some critical issues regarding the existing datasets and benchmarks, such as dataset overfitting by existing methods, inconsistent dataset splits, and incompetent evaluation of different modalities.

The remainder of this paper is organized as follows. Section 2 provides a revisit of existing datasets and the corresponding retina imaging modalities. In Section 3, we outline the tasks for retina image analysis and present the used dataset, evaluation metrics, and benchmarks. Section 4 emphasizes the challenges in the current research and explores future research directions. In the end, we conclude the survey with a summary of our findings and a call to action for further research in Section 5.

2 Multimodal Retina Image Datasets

Multimodal retina image datasets represent incredibly valuable resources for both the development and evaluation of deep learning models. Through the seamless integration of various imaging modalities, these datasets enable a wide array of fundamental yet diverse computational tasks, including disease classification, lesion segmentation, and vascular analysis. Such datasets exhibit significant diversity in terms of modality combinations, dataset size, annotation quality, and accessibility. In this context, we first present a systematic review of multimodal retina image datasets, focusing on their applications and support for various deep learning tasks.

2.1 Overview of Datasets

During our paper collection, we identified 17 publicly available datasets. These datasets cover a variety of retina imaging modalities, with CF and OCT being the most common.

Dataset Name$_{[Year]}$	Size	Type	Imaging Device
INSPIRE-stereo$_{[2011]}$	0.03k	C,O	Nidek 3Dx
CF-FA$_{[2012]}$	0.06k	C,F	-
CF-OCT-REG-22$_{[2013]}$	0.02k	C,O	Topcon 3D OCT 1000
CF-OCT-OSD-50$_{[2014]}$	0.10k	C,O	Multiple Devices
Project Macula$_{[2015]}$	0.13k	C,O,H	Multiple Devices
GL-CF-OCT$_{[2018]}$	0.05k	C,O	Topcon 3D OCT 2000
Biomisa-ARMD$_{[2018]}$	0.1k/68k	C/O	TopCon TRC 50EX
OCTA-500*$_{[2024]}$	0.5k	O,OA	RTVue-XR
PRIME-FP20$_{[2021]}$	0.01k	C,F	Optos California
DeepEyeNet$_{[2021]}$	18k/14k	F/C	-
GL-Dis$_{[2022]}$	9k/0.18k	O/C	Topcon 3D OCT 2000
OLIVES$_{[2022]}$	1k/78k	IR/O	-
GAMMA$_{[2023]}$	0.3k	C,O	TRC-NW400, Kowa
APTOS2023$_{[2023]}$	55k**		-
FPRM$_{[2024]}$	5k/0.7k	I/V	Discovery Model E
APTOS2024$_{[2024]}$	1k/7k	C/O	Kowa VX-20
FOCTAIR	0.09k/0.2k	F/OA	-

Table 1: Summary of dataset size and recording devices. " / " denotes the inclusion of multiple imaging modalities or sample counts. "-" denotes the information is not available or not found. * The OCTA-500 dataset is first released in the year 2020. ** 55k contains 5.8k FA images, 4.2k ICGA images, and 45k FA&ICGA images. "O" denotes OCT. "C" denotes CF. "F" denotes FA. "H" denotes histology image. "OA" denotes OCTA. "IR" denotes near-infrared image. "I" denotes image. "V" denotes video.

Specifically, 13 datasets include CF images, while 11 utilize OCT. FA presents in 5 datasets, and OCTA appears in 2 datasets. In addition to these widely used modalities, other imaging techniques are represented in individual datasets: histology images (HI) in Project Macula, near-infrared (NIR) images in OLIVES, and indocyanine green angiography (ICGA) images in APTOS2023. Notably, the FPRM dataset stands out for its diverse imaging modalities, including oxygen saturation images, pupillary light reflex videos, and retina blood flow (RBF) videos. However, despite its uniqueness, this newly published dataset has seen limited citations and lacks established benchmarking data or methods for comparison, similar to the GL-Dis dataset. In this section, we categorize these datasets by their imaging modalities and analyze their key characteristics and limitations.

2.2 Datasets with CF and OCT

Eight datasets that include both CF and OCT, namely CF-OCT-REG-22, CF-OCT-OSD-50, Project Macula, Biomisa-ARMD, GL-CF-OCT, GL-Dis, GAMMA, INSPIRE-stereo, and APTOS2024. As the most common combination, the CF and OCT provide complementary structural information about the retina. CF captures a global view of the retina surface, while OCT provides high-resolution cross-sectional imaging of retina layers. This combination is particularly useful for analyzing morphological and structural changes in the retina associated with retina diseases, including DR, AMD, and GL. Several datasets in this category support disease classification and structural analysis tasks. Notable examples include GAMMA, Biomisa-ARMD, GL-CF-OCT, GL-Dis, and

Dataset Name	CF	OCT	FA	OCTA	Others	Supported Tasks
CF-FA	✓		✓			DR Grading, Registration, Generation
PRIME-FP20	✓		✓			Vessel Segmentation
DeepEyeNet	✓		✓			Multi Classification, Medical Report Generation
CF-OCT-REG-22	✓	✓				Registration
CF-OCT-OSD-50	✓	✓				Registration
Project Macula	✓	✓			✓	Multi Classification
Biomisa-ARMD	✓	✓				AMD Classification
GL-CF-OCT	✓	✓				GL Classification
GL-Dis	✓	✓				Multi Classification, OD/C Segmentation, Lesion Segmentation
GAMMA	✓	✓				OD/C Segmentation, GL Grading, Fovea Localization
APTOS2024	✓	✓				Generation
INSPIRE-stereo	✓	✓				Registration, Generation
OLIVES		✓			✓	DR Classification, DME Classification
OCTA-500		✓		✓		Vessel Segmentation, FAZ Segmentation, Multi Classification
APTOS2023			✓		✓	Multi Classification, Eye Impression Assessment
FOCTAIR			✓	✓		Registration
FPRM	✓				✓	Quality Assessment, Multi Classification, Psychological Assessment

Table 2: Distribution of datasets modalities and corresponding computational tasks. "CF-OCT-REG-22" denotes "Database for the purpose of vessel-based registration of Fundus and OCT projection images". "CF-OCT-OSD-50" denotes "OCT data & Color Fundus Images of Left & Right Eyes of 50 healthy persons" dataset. "GL-CF-OCT" denotes "Glaucoma Fundus and OCT Dataset". "GL-Dis" denotes "retina Image Database for Macular and Glaucomatous Disorders" dataset. "Others" denotes other modalities, including histology images (HI), near-infrared (NIR) images, indocyanine green angiography (ICGA), oxygen saturation images, pupillary light reflex videos, and retina blood flow (RBF) videos.

Project Macula dataset. Meanwhile, CF-OCT-REG-22 and CF-OCT-OSD-50 datasets are designed for registration tasks, facilitating multimodal alignment studies. Additionally, the APTOS2024 challenge introduces an image synthesis task, where models generate OCT images from CF inputs.

2.3 Datasets with CF and FA

Three datasets contain both CF and FA modalities, namely CF-FA, PRIME-FP20, and DeepEyeNet. FA enables high-contrast visualization of retina vasculature, while CF provides complementary contextual information, capturing retina surface features, including tissue morphology. These datasets are commonly used for tasks involving vascular abnormalities. For example, PRIME-FP20 is derived from a DR study project in clinical settings to segment vasculature. The CF-FA is proposed for DR detection and become an important resource for registration and image generation. Meanwhile, the DeepEyeNet covers 265 different disease conditions along with corresponding textual descriptions, making it a valuable dataset for medical report generation.

2.4 Other Multimodal Datasets

Some datasets incorporate imaging modalities beyond CF, OCT, FA, and OCTA, introducing unique imaging techniques that enable specialized tasks. These datasets provide valuable resources for exploring novel AI applications in retina images. For example, APTOS2023 includes ICGA images, which can provide additional vascular insights beyond FA. FOCTAIR combines FA and OCTA, making it one of the few datasets emphasizing vasculature. FPRM is particularly notable for its diverse range of imaging modalities, including oxygen saturation imaging, videos of pupillary light reflex, videos of retina blood flow (RBF), textual psychological assessment results and image quality reports, enabling research in quality assessment, disease classification, and psychological assessment. These above modalities, though relatively rare, provide unique perspectives for multimodal image analysis. Their diverse imaging techniques contribute to the scope of AI applications in retina research, beyond traditional structural and vascular assessments.

Additionally, the OCTA-500 dataset integrates OCT and OCTA to support not only multi-class disease classification, but also fine-grained segmentation tasks such as capillary segmentation, artery segmentation, vein segmentation, and FAZ segmentation. OLIVES incorporates near-infrared (NIR) imaging, a modality that assists in the analysis of retina tissue properties, particularly in detecting structural alterations and fluid accumulation associated with diabetic retinopathy (DR) and diabetic macular edema (DME). Both OCTA-500 and OLIVES are widely used datasets, supporting various retina AI research applications.

3 Multimodal Retina Image Analysis Tasks

We identify five distinct categories of tasks based on their input and output characteristics. This section discusses the benchmarks, evaluation metrics, and state-of-the-art methods associated with each multimodal retina image analysis task.

3.1 Classification and Grading

Problem Setting. retina classification tasks primarily focus on detecting ocular pathologies by analyzing disease-specific features in retina images, such as GL classification and DME Classification. Grading derives from classification as it follows a predefined ordinal scale, where class labels have a set

Dataset	Method	Year	Acc.	Sen.	Spe.	Pre.	AUROC	F1-score	TV	Kappa
CF-FA	VTGAN	[2021]	85.7	83.3	90.0	-	-	-	-	-
Project Macula	Yoo *et al*.	[2019]	97.3	-	-	-	99.4	-	-	-
	El-Ateif & Idri	[2024]	100	100	100	100	100	100	-	-
GAMMA	SmartDSP	[2023]	-	-	-	-	-	-	-	85.49
	EyeMoSt	[2024]	86.0	-	-	-	-	-	-	76.1
	DGLR	[2023]	-	82.0	94.0	-	98.7	-	0.64	-
	MM-RAF	[2023a]	-	93.33	80.78	89.25	95.84	85.15	-	-

OLIVES	Method	Year	DR/DME Detection				BioMarker		TV	Kappa
			Acc.	Sen.	Spe.	AUROC	Ave. Spe.	Ave. Sen.		
	OLIVES	[2022]	82.33	80.4	74.2	79.0	77.2	67.5	-	-
	EyeMoSt	[2024]	100.0	-	100.0	-	-	-	-	100.0

DEN	Method	Year	Pretrain		Random Init	
			Pre@1	Pre@5	Pre@1	Pre@5
	ResNet	[2016]	37.09	63.36	36.60	62.87
	VGGNet	[2014]	54.23	80.75	35.93	73.73
	Jing *et al*.	[2018]	32.72	63.75	29.11	60.68

Table 3: Classification and grading benchmarks on the CF-FA, Project Macular, GAMMA, and OLIVES datasets. AUROC is tested under 95% confidence interval. "-" denotes the metric that is not available or not mentioned in the original method paper. "Acc." denotes accuracy. "Sen." denotes sensitivity. "Spe." denotes specificity. "Pre." denotes precision. "TV" denotes threshold value. "Ave." denotes average. "Pre@k" denotes top-k precision.

distance between them, rather than being distinct categorical labels. It follows an ordinal classification paradigm based on clinical needs, allowing the assessment of disease progression and severity (*e.g*., DR grading).

Associated Datasets. There are 10 datasets supporting classification and grading tasks. Five of them form available benchmarks, providing standardized evaluation frameworks. Among these, the GAMMA dataset comes with an online challenge. DEN, Biomisa-ARMD, and OLIVES conduct baseline benchmarks.

Evaluation Metrics. As shown in the Table 3, Cohen's kappa (Kappa) is served as an evaluation metric for grading problems to amplify the distance of the error:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{1}$$

where P_o is the observed agreement, and P_e is the expected agreement by chance. The other metrics, such as Accuracy, Sensitivity (Recall, True Positive rate), Specificity (True Negative Rate), Precision, Area Under Receiver Operating Characteristic Curve (AUROC), F1-score are all commonly used metrics in classification tasks. Biomisa-ARMD originally provides a confusion matrix of normal, early, and advanced AMD as a baseline result.

Related Methods. Yoo *et al*. [2019] describe itself as the first to explore the use of both fundus and OCT images in a deep learning-based classification algorithm. Their preliminary findings suggest that multimodal integration of CF and OCT may enhance diagnostic accuracy. Notably, in recent years, the metrics on the Project Macular dataset reach 100%. Similarly, Zou *et al*. [2024] integrate uncertainty estimation into the fusion process, leading to a 100% outcome for the OLIVES dataset in the grading task. However, their models do not achieve top performance on the GAMMA dataset.

3.2 Registration

Problem Setting. multimodal registration aligns corresponding features across images from different modalities. It enables more precise comparisons, downstream multimodal fusion, and enhanced diagnostic decision-making

Problem Setting. The CF-FA dataset is one of the most used public datasets for retina registration. Other multimodal retina datasets offering registration annotations are CF-OCT-REG-22, CF-OCT-OSD-50, INSPIRE-stereo, and FOCTAIR. INSPIRE-stereo dataset is rarely cited in recent registration works. Table 4 demonstrates recent research on datasets CF-FA, CF-OCT-REG-22, CF-OCT-OSD-50, and FOCTAIR.

Evaluation Metrics. Success Rate (SR) quantifies the proportion of registration cases where alignment error falls within a predefined threshold, reflecting the reliability of the registration method. SR metrics, such as SR_{ME}(n) and SR_{MAE}(n), measure the proportion of cases where registration error remains below a given threshold. While SR_{ME} relies on mean error, SR_{MAE} uses absolute error, both providing insights into the preservation of anatomical structures. Mean Euclidean Error (MEE) quantifies the average registration error across all points, while Maximum Euclidean Error (MAE) highlights the worst-case misalignment by considering the largest observed discrepancy. Root Mean Squared Error (RMSE) further refines error measurement by emphasizing larger deviations through squaring, making it particularly sensitive to misalignments. Unlike these direct error metrics,

Dataset	CF-FA				CF-OCT-REG-22		CF-OCT-OSD-50	FOCTAIR
Method	[2021]	[2022]	[2022]	[2024a]	[2013]	[2018]	[2018]	[2023]
SR	100	100	-	-	-	-	-	-
$SR_{ME}(\epsilon = 2)$	-	-	92.9	-	-	-	-	-
$SR_{ME}(\epsilon = 3)$	-	-	100.0	-	-	-	-	-
$SR_{MAE}(\epsilon = 3)$	-	-	92.9	-	-	-	-	-
$SR_{MAE}(\epsilon = 5)$	-	-	100.0	-	-	-	-	-
$REP(\epsilon = 5)$	-	-	-	-	-	-	-	-
SDP	-	-	-	-	1.02	3.92	4.21	-
MAE	-	-	2.47	-	-	6.17	5.39	-
MEE	-	-	1.50	2.01	-	3.83	3.93	-
RMSE	-	-	-	-	-	3.12	3.70	-
AUC	-	-	-	0.858	-	-	-	-
Dice	0.679	0.663	0.659	-	-	-	-	0.7166
ZNCC	-	-	-	-	-	-	-	0.8211
$\lvert J_\phi \rvert_{\leq 0}$	-	-	-	-	-	-	-	✓

Table 4: Registration benchmark on datasets. "-" denotes the metric that is not used in the corresponding method paper.

Regional Error Propagation (REP) and Spatial Deformation Penalty (SDP) assess the smoothness and spatial consistency of transformations, ensuring that deformations remain locally coherent. In the context of similarity-based evaluations, Zero-mean Normalized Cross-Correlation (ZNCC) measures the intensity correlation between registered images, making it particularly suitable for multimodal registration tasks. Dice Coefficient (DC) and Area Under the Curve (AUC) provide an overlap-based perspective, with Dice quantifying spatial alignment of segmented structures and AUC evaluating registration performance as a classification task. The Jacobian determinant constraint $|J_\phi|_{\leq 0}$ plays a crucial role in assessing transformation field invertibility. A negative determinant indicates non-physical folding in non-rigid deformations, highlighting structural inconsistencies that could invalidate registration outcomes.

Related Methods. Sindel *et al*. [2022] adopt a keypoint-based vessel structure alignment method to optimize vessel matching, enhancing the alignment accuracy of multimodal retina images. Martínez-Río *et al*. [2023] propose a weakly supervised deep learning approach within a deformable registration framework. This method achieves effective multimodal retina image alignment with minimal annotation requirements. Jahnavi and Sivaswamy [2018] discuss the challenges and optimization directions in cross-modal alignment and provide a generalized multimodal retina registration pipeline. Retina IPA [Wang *et al.*, 2024a] introduces a self-supervised layer enhanced with keypoints as constraints while refining feature extraction across modalities.

3.3 Segmentation

Problem Setting. Segmentation in retina image partitions a raw image at the pixel level to delineate clinically significant anatomical structures or pathological regions, such as lesions, vessel boundaries, or fluid cavities. The output is a labeled mask that highlights these features, aiding in disease diagnosis, staging, and monitoring.

Associated Datasets. PRIME-FP20 dataset supports the vessel segmentation task, and OCTA-500 (two subsets: 6mm and 3mm) supports Capillary Artery Vein Fovea (CAVF) segmentation tasks. GAMMA dataset also provides annotations for fovea localization task but lacks followers.

Evaluation Metrics. AUCPR, Dice Coefficient (DC), Intersection over Union (IoU), and Correctly Classified Area over Lesion (CAL) each provide a unique perspective on segmentation performance. AUCPR measures model effectiveness in imbalanced datasets by evaluating the trade-off between precision and recall, ensuring that high recall is not achieved at the expense of excessive false positives. Dice and IoU both assess spatial overlap between predicted and ground truth segmentations, with Dice emphasizing the harmonic mean of precision and recall, while IoU penalizes over-segmentation more strongly. CAL, in contrast, focuses specifically on segmentation accuracy within lesion regions, decomposing performance into correct lesion segmentation (C), accurate boundary classification (A), and lesion localization (L). These metrics balance global accuracy with lesion-specific performance, making them particularly useful in medical image applications.

Related Methods. The work of [Ding *et al.*, 2021] proposes a weakly-supervised approach for vessel detection in ultra-widefield fundus photography by leveraging iterative multimodal registration and learning. It exploits angiography as an auxiliary modality for supervision without requiring extensive manual annotations. The model refines vessel detection iteratively, demonstrating improved performance over standard weakly-supervised methods in challenging ultra-widefield images. The authors of OCTA-500 [Li *et al.*, 2020; Li *et al.*, 2024] introduce a large-scale OCTA dataset with paired en face and volumetric angiography images. The study benchmarks several deep learning models, including their proposed IPNv2 for segmentation, highlighting the potential of their dataset for advancing OCTA image understanding.

Dataset	Method	Year	Split	AUCPR	DC	CAL(C,A,L)
PRIME-FP20	Self-supervised	[2021]	5-fold cross-valid	0.842	0.772	0.730(0.999,0.849,0.860)
	Unet-based	[2021]	5-fold cross-valid	0.869	0.796	0.755(0.999,0.869,0.870)
OCTA-500				IoU(C)	IoU(A)	IoU(V)	IoU(F)	mIoU
OCTA-500-6mm	IPN-V2	[2024]	pre-defined	84.34	76.74	77.26	88.76	81.77
OCTA-500-3mm	IPN-V2	[2024]	pre-defined	86.16	82.26	81.38	95.15	86.24

Table 5: Segmentation benchmarks on the OCTA-500, PRIME-FP20 dataset. "mIOU" denotes mean IoU.

3.4 Image Generation

Problem Setting. Image generation tasks, sometimes referred to as image translation tasks, are very popular in medical images. Image generation tasks in medical images involve creating new or enhanced images to augment existing datasets or simulate realistic variations. Given a source image or latent representation, these approaches output new medical images with desired attributes. Such generation techniques can aid in data augmentation and improve diagnostic clarity.

Associated Datasets. In the context of multimodal retina images, only the CF-FA dataset, INSPIRE-stereo dataset, and APTOS2024 dataset are accessible. INSPIRE-stereo dataset has rare visibility in the deep learning community. APTOS2024 is a new challenge held by Asia Pacific Tele-Ophthalmology Society (APTOS), aiming to motivate CF to OCT generation.

Evaluation Metrics. Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), Mean Squared Error (MSE), Fréchet Inception Distance (FID), and Kernel Inception Distance (KID) are key metrics for evaluating image generation quality, each capturing different aspects of fidelity and realism. SSIM measures structural similarity between images by considering luminance, contrast, and texture, making it more perceptually relevant than pixel-wise comparisons. PSNR, in contrast, quantifies reconstruction quality based on logarithmic scaling of the pixel-wise error, favoring images with less distortion but often misaligned with human perception. MSE directly computes the average squared pixel difference, serving as a simple but less perceptually aligned metric, where lower values indicate better reconstruction fidelity. For assessing realism in generated images, FID and KID go beyond pixel-level comparisons. FID measures the distance between real and generated image distributions in a deep feature space, assuming a Gaussian distribution, making it effective for capturing global image quality but sensitive to dataset size. KID, as an alternative, leverages Maximum Mean Discrepancy (MMD) to compare distributions without assuming Gaussianity, providing a more robust, unbiased estimate, particularly for smaller datasets. In a nutshell, SSIM, PSNR, and MSE focus on pixel-wise and perceptual fidelity, while FID and KID assess realism in a learned feature space. Kamran *et al*. [2018] use validation loss and a GAN-like module to demonstrate the validity of the pseudo-angiography images. FVD is an extension of the FID metric, specifically designed to evaluate the quality of video-like sequences by measuring their distributional distance from real data in a feature space. In the APTOS2024 challenge, the FVD metric is applied by treating these six-frame sequences as mini-videos and comparing their learned feature distributions with those of real OCT frame sequences

Related Methods. Kamran *et al*. [2020] proposes an adversarial framework, named Fundus2Angio, to improve the realism of generated images, with a generator learning structural and vascular mappings while a discriminator enforces fidelity. Performance evaluation suggests that the synthesized FA images retain key vascular details, supporting potential clinical applications. Their following work [Kamran *et al.*, 2021] proposes a semi-supervised approach by integrating Vision Transformers (ViT) with Generative Adversarial Networks (GANs). The key contribution lies in leveraging ViT for feature modeling to enhance both image generation quality and downstream diagnostic performance.

3.5 Medical Report Generation

Problem Setting. The medical report generation task automates the creation of medical reports from image data, covering descriptions, diagnoses, and recommendations. It can generate structured or free-text reports with disease classification, lesion localization, and clinical guidance. The report generation task enhances diagnostic efficiency, reduces reporting workload, and improves interpretability.

Associated Datasets. The DeepOpht dataset provides a standardized collection of multimodal retina images, enabling the development and evaluation of AI models for medical image captioning and disease classification. It provides a total of 15709 images, including 1811 FA and 13898 CFP. Clinical labels, keywords and clinical descriptions covering 265 different disease conditions are further provided.

Evaluation Metrics. This benchmark primarily focuses on assessing model performance in generating structured medical reports from retina images, ensuring clinical relevance and interoperability. As shown in Table 7, the performance of AI models on the DeepOpht dataset is evaluated using a range of metrics, each serving a distinct role in assessing model effectiveness, BLEU [Papineni *et al.*, 2002], CIDEr [Vedantam *et al.*, 2015] and ROUGE [Lin, 2004]. These are key metrics for evaluating text generation quality, particularly in machine translation, image captioning, and summarization tasks. BLEU measures n-gram precision by comparing generated text to reference text. BLEU-1 to BLEU-4 correspond to unigram to four-gram precision, respectively, with BLEU-avg representing their averaged score. Lower-order BLEU scores

Dataset	Method	Year	SSIM	PSNR	MSE	FID	KID	FVD
CF-FA	Fundus2Angio	[2020]	-	-	-	30.3	-	-
	VTGAN	[2021]	-	-	-	17.3	0.00053	-
	UDLM-FFA	[2024]	0.6237	20.3488	838.9363	-	-	-
APTOS2024	NJUST-EYE	2024	0.1048	13.6502	-	-	-	624.5898

Table 6: Generation benchmarks on the CF-FA and APTOS2024 datasets.

Dataset	Method	Year	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-avg	CIDEr	ROUGE
DEN	[Jing *et al.*, 2018]	2018	0.184	0.114	0.068	0.032	0.100	0.361	0.232
	[Li *et al.*, 2019]	2019	0.181	0.107	0.062	0.032	0.096	0.453	0.230
	[Huang *et al.*, 2022]	2022	0.230	0.150	0.094	0.053	0.132	0.370	0.291
	[Wang *et al.*, 2024c]	2024	0.390	0.267	0.202	0.159	*0.255*	0.762	0.394

Table 7: Medical report generation benchmark on the DeepEyeNet(DEN) dataset. "-" indicates metrics not reported in the respective study. Italic indicates the result is not provided originally and is calculated by us according to provided data, which may vary with the true value.

(*e.g.*, BLEU-1) capture basic word accuracy, while higher-order scores (*e.g.*, BLEU-4) reflect fluency and coherence but can be overly strict due to exact match requirements. CIDEr, in contrast, emphasizes human-like relevance by weighting n-grams based on their importance in a given corpus. Unlike BLEU, which treats all matches equally, CIDEr assigns higher scores to rare yet meaningful phrases, making it particularly effective for tasks like image captioning. ROUGE evaluates recall-oriented overlap, commonly used in summarization. Unlike BLEU, which prioritizes precision, ROUGE measures how much of the reference text is retained in the generated output, making it well-suited for evaluating completeness in extractive and abstractive summarization. For text generation tasks, BLEU and ROUGE appear frequently, reflecting their importance in assessing medical report quality. In the original paper of DEN, the authors also introduce a human-involved visual evaluation named DNN Visual Explanation Module. This module compares the heatmap obtained from CAM [Zhou *et al.*, 2016] with the manually labeled lesion sketch to qualitatively evaluate the effectiveness of the models.

Related Methods. DeepOpht [Jing *et al.*, 2018] proposes the dataset DeepEyeNet (DEN) and explores the framework and central challenges of automated medical report generation. Li *et al*. [2019] integrate prior knowledge into encoding and retrieval processes, employing multi-stage paraphrasing to enhance report quality. Huang *et al*. [2022] introduce a non-local attention mechanism to capture more comprehensive image features, thereby improving the description generation for retina images. EyeGraphGPT [Wang *et al.*, 2024c] is built on a multimodal large language model incorporating knowledge graphs, aiming to enrich medical knowledge integration and expression in ophthalmic report generation.

4 Discussion

4.1 Challenge of Current Multimodal Datasets

Dataset Constraints. Some datasets, such as OLIVES and Project Macula, have reached full accuracy, making it difficult to measure further improvements. Recent studies, such as EyeMoSt [Zou *et al.*, 2024], achieve full accuracy even after adding Gaussian noise to OLIVES [2022], suggesting that some benchmarking datasets may no longer differentiate model performance effectively. These limitations reflect not only the saturation of current benchmarks, but also deeper issues in dataset design and availability. Many datasets remain small or contain biases, which limit the generalizability of models trained on them. This scarcity is often rooted in the challenges of clinical data governance, where privacy concerns, patient consent requirements, and institutional regulations restrict the ability to share data across centers. These challenges highlight the need for larger, more diverse, and well-standardized multimodal retina image analysis datasets.

Modality Misalignment. A common response to these constraints is to merge multiple datasets in order to construct larger and more diverse evaluation pools [Stankevièius *et al.*, 2018; Pratap and Kokil, 2019]. While this strategy can improve sample diversity by introducing a larger sample pool, it also introduces practical challenges. In particular, these efforts often encounter inconsistent labeling standards and divergent annotation protocols, making it difficult to establish unified benchmarks and conduct reliable cross-dataset evaluations. These issues become more pronounced in multimodal settings. Beyond label inconsistencies, this complexity arises not only from device heterogeneity, but also from inconsistencies in modality composition. First, multimodal retina image datasets exhibit significant variation in acquisition devices, as summarized in Tables 1 and 2. Different datasets often use different imaging devices for the same modality. Such variability in imaging devices, including differences in hardware, resolution, contrast, and preprocessing pipelines, introduces domain shifts. The cross-domain shifts in the cross-modal scenarios may raise challenges to generalization. Second, imaging modalities vary across datasets. While some include OCT and CFP, others contain OCTA or FA, resulting in limited modality overlap and poor cross-dataset compatibility. In multimodal settings, each sample refers to one subject. Different subjects may have different sets of available

modalities, particularly when data from multiple sources are combined. This makes it difficult to build data pipelines that rely on consistent modality combinations, and constrains the scalability of multimodal fusion strategies across large and heterogeneous data sources.

4.2 Issues in Benchmarks

Evaluation Protocols. The current benchmarks exhibit significant ambiguities regarding data splits for training and testing. Many datasets do not specify clear partitions for training versus testing, leading some studies to adopt 5-fold cross-validation to compare their work against others. However, this method can introduce randomness [Varma and Simon, 2006], while other researchers opt for purely random splits, further complicating direct comparisons. Additionally, some work (4.1) mix heterogeneous public datasets and even private collections in their training or testing, often without clear documentation of the split strategies. Such inconsistencies limit the transparency, reproducibility, and comparability of downstream evaluations. A robust benchmark should feature a well-designed testing set that guarantees the training and testing data originate from the same distribution, ensuring more consistent and reliable evaluations.

Evaluation Metrics. This survey exposes 28 distinct evaluation metrics and 14 variations. Although assessing performance from various perspectives can be advantageous, an overabundance of metrics often results in redundancy due to overlapping characteristics. Moreover, the use of different evaluation metrics across studies directly hinders the ability to compare results effectively. Researchers in the retina registration field frequently reimplement existing methods using their own chosen metrics and datasets to ensure fair comparisons. This practice imposes additional computational and implementation burdens, and often results in incompatible results across studies. Benchmarks should therefore select metrics that are best suited to the dataset and task, providing a comprehensive reflection of model performance and covering all clinical-relevant aspects.

4.3 Evaluation of Individual Modality

Fusion Ambiguity. Multimodal fusion is a central objective in retina AI research, yet its practical benefits are not well-characterized. While some works (e.g., Uni4Eye [Cai *et al.*, 2022]) report per-modality performance, they rarely quantify the incremental gains achieved through fusion. In some cases, such as IPNv2 [Li *et al.*, 2024], models trained solely on OCTA even outperform their fused counterparts, challenging the common assumption that combining modalities always improves performance. These findings suggest that fusion may introduce redundancy, suboptimal architectural interactions, or additional noise. This reveals a lack of robust tools to isolate and evaluate individual modality contributions, limiting the ability to systematically optimize the fusion strategies.

Modality Variability. Beyond these methodological limitations, the modality contribution itself is inherently dynamic and context-dependent. First, the modality importance may vary by task, as different objectives rely on different imaging inputs. Second, the contributions may fluctuate across subjects, depending on disease characteristics or image quality. Third, the utility of a modality during training may not align with its value during inference, as some inputs enhance feature learning but contribute little to final predictions. These nuances call for dropout-based or ablation-based evaluations to assess modality-level robustness. Altogether, these challenges emphasize the need for principled evaluation frameworks that can disentangle, interpret, and benchmark modality-specific contributions in multimodal retina systems.

4.4 Future Direction

We believe that a dataset and the corresponding benchmark are essential to ensure a fair comparison and facilitate reproducibility by providing uniform evaluation protocols. To build such reliable and transparent evaluation protocols, several key improvements are necessary. First, the community must adopt shared protocols for dataset preprocessing, uniform training-validation-test splits, and explicit ablation studies to determine the impact of individual imaging modalities. Second, expanding dataset diversity, both in terms of pathologies and imaging devices, will mitigate the risk of overfitting specific data distributions. Third, federated learning offers a promising pathway for privacy-preserving data sharing across institutions, allowing for large-scale validation without compromising patient confidentiality. Finally, the development of standardized benchmarking frameworks, including robust cross-dataset validation, clinically interpretable evaluation criteria, and open-source reproducibility, is critical for transforming multimodal retina AI from promising research prototypes into clinically deployable tools. Emerging foundation models offer new potential for cross-modal reasoning and automated reporting in retinal analysis [2023b; 2024; 2024b; 2025]. Most existing foundation models are either limited to single-modality settings or trained on multimodal data without subject-level pairing. Foundation models aligned at the patient level hold promise for meaningful clinical integration and reasoning, a potential that is contingent upon the availability and quality of multimodal datasets.

5 Conclusion

This survey highlights the critical role of multimodal retina images in advancing ophthalmic diagnostics and AI-driven analysis. By integrating complementary structural and vascular modalities, multimodal approaches enhance disease detection and monitoring. However, the field lacks standardization in dataset usage, fusion methodologies, and evaluation frameworks, limiting reproducibility and comparability across studies. Our review systematically examines existing datasets, benchmarking methodologies, and deep learning strategies, emphasizing the need for standardized evaluation frameworks to ensure robust and clinically relevant AI models. Future research should prioritize the development of large-scale, well-annotated datasets and clinically interpretable benchmarks to bridge the gap between research advancements and real-world applications. Addressing these challenges will be pivotal in realizing the full potential of AI-driven multimodal retina analysis in clinical practice.

Acknowledgments

This research is funded in part by ARC-Discovery grant (DP220100800), ARC-DECRA grant (DE230100477). We gratefully thank all the reviewers and chairs for their constructive comments.

References

[An *et al.*, 2022] Cheolhong An, Yiqian Wang, Junkang Zhang, and Truong Q Nguyen. Self-supervised rigid registration for multimodal retinal images. *IEEE TIP*, 2022.

[Cai *et al.*, 2022] Zhiyuan Cai, Li Lin, Huaqing He, and Xiaoying Tang. Uni4eye: Unified 2d and 3d self-supervised pre-training via masked image modeling transformer for ophthalmic image classification. In *MICCAI*, 2022.

[Chen *et al.*, 2024] Yiwei Chen, Yi He, Hong Ye, Lina Xing, Xin Zhang, and Guohua Shi. Unified deep learning model for predicting fundus fluorescein angiography image from fundus structure image. *JIOHS*, 2024.

[Ding *et al.*, 2021] Li Ding, Ajay E. Kuriyan, Rajeev S. Ramchandran, Charles C. Wykoff, and Gaurav Sharma. Weakly-supervised vessel detection in ultra-widefield fundus photography via iterative multi-modal registration and learning. *TMI*, 2021.

[El-Ateif and Idri, 2024] Sara El-Ateif and Ali Idri. Multimodality fusion strategies in eye disease diagnosis. *JIIM*, 2024.

[Golabbakhsh and Rabbani, 2013] Marzieh Golabbakhsh and Hossein Rabbani. Vessel-based registration of fundus and optical coherence tomography projection images of retina using a quadratic registration model. *IET Image Process.*, 2013.

[Hajeb Mohammad Alipour *et al.*, 2012] Shirin Hajeb Mohammad Alipour, Hossein Rabbani, and Mohammad Reza Akhlaghi. Diabetic retinopathy grading by digital curvelet transform. *Comput. Math. Methods Med.*, 2012.

[Hassan *et al.*, 2022] Taimur Hassan, Hina Raja, Bilal Hassan, Muhammad Usman Akram, Jorge Dias, and Naoufel Werghi. A composite retinal fundus and oct dataset to grade macular and glaucomatous disorders. In *ICoDT2*, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Hemelings *et al.*, 2023] Ruben Hemelings, Bart Elen, Alexander K Schuster, Matthew B Blaschko, João Barbosa-Breda, Pekko Hujanen, Annika Junglas, Stefan Nickels, Andrew White, et al. A generalizable deep learning regression model for automated glaucoma screening from fundus images. *NPJ digital medicine*, 2023.

[Hervella *et al.*, 2018] Álvaro S Hervella, José Rouco, Jorge Novo, and Marcos Ortega. Retinal image understanding emerges from self-supervised multimodal reconstruction. In *MICCAI*, 2018.

[Huang *et al.*, 2021] Jia-Hong Huang, C-H Huck Yang, Fangyu Liu, Meng Tian, Yi-Chieh Liu, Ting-Wei Wu, et al. Medical report generation for retinal images via deep models and visual explanation. In *WACV*, 2021.

[Huang *et al.*, 2022] Jia-Hong Huang, Ting-Wei Wu, C-H Huck Yang, Zenglin Shi, I Lin, Jesper Tegner, Marcel Worring, et al. Non-local attention improves description generation for retinal images. In *WACV*, 2022.

[Jahnavi and Sivaswamy, 2018] Gamalapati S Jahnavi and Jayanthi Sivaswamy. Multimodal registration of retinal images. In *NCVPRIPG*, 2018.

[Jing *et al.*, 2018] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *ACL*, 2018.

[Kamran *et al.*, 2020] Sharif Amit Kamran, Khondker Fariha Hossain, Alireza Tavakkoli, Stewart Lee Zuckerbrod, et al. A conditional gan architecture for generating fluorescein angiography images from retinal fundus photography. *arXiv*, 2020.

[Kamran *et al.*, 2021] Sharif Amit Kamran, Khondker Fariha Hossain, Alireza Tavakkoli, Stewart Lee Zuckerbrod, and Salah A. Baker. VTGAN: Semi-supervised retinal image synthesis and disease prediction using vision transformers. In *ICCV Workshops*, 2021.

[Khalid *et al.*, 2018] Samina Khalid, M Usman Akram, Taimur Hassan, Amina Jameel, and Tehmina Khalil. Automated segmentation and quantification of drusen in fundus and optical coherence tomography images for detection of armd. *JDI*, 2018.

[Khalil *et al.*, 2018] Tehmina Khalil, M Usman Akram, Hina Raja, Amina Jameel, and Imran Basit. Detection of glaucoma using cup to disc ratio from spectral domain optical coherence tomography images. *IEEE Access*, 2018.

[Khan *et al.*, 2023] MD Wahiduzzaman Khan, Hongwei Sheng, Hu Zhang, Heming Du, Sen Wang, et al. Rvd: a handheld device-based fundus video dataset for retinal vessel segmentation. *Neurips*, 2023.

[Li *et al.*, 2019] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *AAAI*, 2019.

[Li *et al.*, 2020] Mingchao Li, Yerui Chen, Zexuan Ji, Keren Xie, Songtao Yuan, Qiang Chen, and Shuo Li. Image projection network: 3d to 2d image segmentation in octa images. *IEEE TMI*, 2020.

[Li *et al.*, 2024] Mingchao Li, Kun Huang, Qiuzhuo Xu, Jiadong Yang, Yuhan Zhang, et al. Octa-500: a retinal dataset for optical coherence tomography angiography study. *Medical image analysis*, 2024.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL*, 2004.

[Liu and Yu, 2021] Jinhui Liu and Xin Yu. Few-shot weighted style matching for glaucoma detection. In *CICAI*, 2021.

[Mahmudi *et al.*, 2014] Tahereh Mahmudi, Rahele Kafieh, et al. Comparison of macular octs in right and left eyes of normal people. In *Medical Imaging*, 2014.

[Martínez-Río *et al.*, 2023] Javier Martínez-Río, Enrique J Carmona, Daniel Cancelas, Jorge Novo, and Marcos Ortega. Deformable registration of multimodal retinal images using a weakly supervised deep learning approach. *Neural Comput. Appl.*, 2023.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[Prabhushankar *et al.*, 2022] Mohit Prabhushankar, Kiran Kokilepersaud, Yash-yee Logan, Stephanie Trejo Corona, et al. Olives dataset: Ophthalmic labels for investigating visual eye semantics. *NeurIPS*, 2022.

[Pratap and Kokil, 2019] Turimerla Pratap and Priyanka Kokil. Computer-aided diagnosis of cataract using deep transfer learning. *Biomed. Signal Process. Control.*, 2019.

[Silva-Rodriguez *et al.*, 2025] Julio Silva-Rodriguez, Hadi Chakor, Riadh Kobbi, Jose Dolz, and Ismail Ben Ayed. A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Medical Image Analysis*, 2025.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014.

[Sindel *et al.*, 2022] Aline Sindel, Bettina Hohberger, Andreas Maier, and Vincent Christlein. Multi-modal retinal image registration using a keypoint-based vessel structure aligning network. In *MICCAI*, 2022.

[Stankevièius *et al.*, 2018] Gediminas Stankevièius, Dalius Matuzevièius, et al. Deep neural network-based feature descriptor for retinal image registration. In *AIEEE*, 2018.

[Tang *et al.*, 2011] Li Tang, Mona K Garvin, Kyungmoo Lee, Wallace LW Alward, Young H Kwon, and Michael D Abramoff. Robust multiscale stereo matching from fundus images with radiometric differences. *TPAMI*, 2011.

[Varma and Simon, 2006] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 2006.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[Wang *et al.*, 2021] Yiqian Wang, Junkang Zhang, Melina Cavichini, Dirk-Uwe G Bartsch, William R Freeman, et al. Robust content-adaptive global registration for multimodal retinal images using weakly supervised deep-learning framework. *IEEE TIP*, 2021.

[Wang *et al.*, 2024a] Jiacheng Wang, Hao Li, Dewei Hu, Rui Xu, Xing Yao, et al. Retinal ipa: I terative key p oints a lignment for multimodal retinal imaging. In *MOVI*, 2024.

[Wang *et al.*, 2024b] Xiaosong Wang, Xiaofan Zhang, Guotai Wang, Junjun He, Zhongyu Li, et al. Openmedlab: An open-source platform for multi-modality foundation models in medicine. *arXiv preprint arXiv:2402.18028*, 2024.

[Wang *et al.*, 2024c] Zhirui Wang, Xinlong Jiang, Chenlong Gao, Fan Dong, et al. Eyegraphgpt: Knowledge graph enhanced multimodal large language model for ophthalmic report generation. In *BIBM*, 2024.

[Wu *et al.*, 2021] Zhuojie Wu, Zijian Wang, Wenxuan Zou, Fan Ji, et al. A progressive attention-enhanced network for 3d to 2d retinal vessel segmentation. In *BIBM*, 2021.

[Wu *et al.*, 2023] Junde Wu, Huihui Fang, Fei Li, Huazhu Fu, Fengbin Lin, Jiongcheng Li, et al. Gamma challenge: glaucoma grading from multi-modality images. *Medical image analysis*, 2023.

[Xie *et al.*, 2024] Xinyu Xie, Yawen Cui, Tao Tan, Xubin Zheng, and Zitong Yu. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *VI*, 2024.

[Yoo *et al.*, 2019] Tae Keun Yoo, Joon Yul Choi, Jeong Gi Seo, Bhoopalan Ramasubramanian, et al. The possibility of the combination of oct and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment. *MBEC*, 2019.

[Yu *et al.*, 2024] Kai Yu, Yang Zhou, Yang Bai, Zhi Da Soh, Xinxing Xu, et al. Urfound: Towards universal retinal foundation models via knowledge-guided masked modeling. In *MICCAI*. Springer, 2024.

[Zanzottera *et al.*, 2015] Emma C Zanzottera, Jeffrey D Messinger, Thomas Ach, R Theodore Smith, K Bailey Freund, and Christine A Curcio. The project macula retinal pigment epithelium grading system for histology and optical coherence tomography in age-related macular degeneration. *IOVS*, 2015.

[Zhang *et al.*, 2023] Weiyi Zhang, Peranut Chotcomwongse, Xiaolan Chen, Florence HT Chung, Fan Song, et al. Angiographic report generation for the 3rd aptos's competition: Dataset and baseline methods. *medRxiv*, 2023.

[Zhang *et al.*, 2024] Guanran Zhang, Yanlin Qu, Yanping Zhang, Jiayi Tang, Chunyan Wang, et al. Multimodal eye imaging, retina characteristics, and psychological assessment dataset. *Scientific Data*, 2024.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[Zhou *et al.*, 2023a] You Zhou, Gang Yang, Yang Zhou, Dayong Ding, and Jianchun Zhao. Representation, alignment, fusion: A generic transformer-based framework for multi-modal glaucoma recognition. In *MICCAI*, 2023.

[Zhou *et al.*, 2023b] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 2023.

[Zou *et al.*, 2024] Ke Zou, Tian Lin, Zongbo Han, Meng Wang, Xuedong Yuan, et al. Confidence-aware multi-modality learning for eye disease screening. *Medical image analysis*, 2024.