



# Multimodal Deep Learning For Eye Disease Recognition From Fundus And OCT Images With Synthetic OCT Generation

Thanh Thuy Nguyen Thi<sup>1</sup>, Xuan Thuy Hoang<sup>2</sup>, Minh Nhut Nguyen<sup>3</sup> and Dinh Thuan Nguyen<sup>4</sup>

## ABSTRACT

Multimodal learning that integrates Color Fundus Photography (CFP) and Optical Coherence Tomography (OCT) can enhance retinal disease recognition by combining surface appearance with depth-resolved structural cues. However, the limited availability of synchronized 2D–3D CFP–OCT datasets hinders the deployment of multimodal AI in real-world screening settings where OCT devices are often unavailable. This study investigates whether synthetic 3D OCT volumes can serve as a practical auxiliary modality to support multimodal classification in fundus-only scenarios. We *synthesize* 3D OCT from fundus images using three generative paradigms—a 3D GAN (DCGAN/WGAN-GP style), a 3D VAE–GAN, and a fundus-conditioned latent diffusion model (3D-DDPM with classifier-free guidance)—and integrate the generated volumes into an EyeMoST-based uncertainty-aware fusion framework. We evaluate the pipeline on RFMiD and ODIR-5K across multiple backbone configurations. Empirically, diffusion-generated OCT yields the most stable fusion performance. In contrast, synthetic OCT does not consistently outperform strong fundus-only models due to domain gap and imperfect biomarker preservation.

## Article information:

**Keywords:** Retinal disease, Multimodal learning, Synthetic OCT, 3D diffusion (DDPM), Uncertainty-aware fusion, EyeMoST

## Article history:

Received: February 13, 2026

Revised: March 26, 2026

Accepted: July 69, 2026

Published: July 31, 2026

(Online)

DOI: 10.37936/ecti-cit.2026203.266190

## 1. INTRODUCTION

Retinal diseases, including age-related macular degeneration (AMD), diabetic retinopathy (DR), and glaucoma, are among the leading causes of visual impairment and irreversible blindness worldwide. Early detection and timely intervention play a critical role in preserving vision and reducing disease burden [1]. In clinical practice, Color Fundus Photography (CFP) is widely adopted for large-scale screening due to its accessibility, low acquisition cost, and ease of operation [2]. However, CFP provides only a two-dimensional view of the retinal surface, which can limit sensitivity to pathologies that manifest primarily in deeper retinal layers.

Optical Coherence Tomography (OCT), by contrast, offers high-resolution cross-sectional and volumetric imaging of retinal microstructure. OCT en-

ables the assessment of depth-related abnormalities that may be subtle or invisible in CFP, such as macular edema, subretinal fluid, and structural changes across retinal layers. Prior studies have demonstrated that multimodal learning can improve diagnostic performance by integrating complementary information from CFP and OCT within computer-aided diagnosis (CAD) systems [3, 4]. Despite its promise, the practical adoption of multimodal AI remains constrained by the scarcity of synchronized paired CFP–OCT datasets. OCT systems are expensive and require more complex acquisition protocols, which makes OCT less available in many screening programs and resource-limited settings [5].

This limitation motivates a practical question: *can generative models be used to bridge the missing-modality gap by synthesizing OCT volumes from widely available fundus images?* Recent advances

<sup>1,2,3,4</sup>The authors are with the Faculty of Information Systems, University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam, Email: [thuynguyen.caroline@gmail.com](mailto:thuynguyen.caroline@gmail.com), [hoangxuanthuy2004@gmail.com](mailto:hoangxuanthuy2004@gmail.com), [nhutnm.17@grad.uit.edu.vn](mailto:nhutnm.17@grad.uit.edu.vn) and [thuannd@uit.edu.vn](mailto:thuannd@uit.edu.vn)

<sup>1,2,3,4</sup>The authors are with the Vietnam National University, Ho Chi Minh City, Vietnam, Email: [thuynguyen.caroline@gmail.com](mailto:thuynguyen.caroline@gmail.com), [hoangxuanthuy2004@gmail.com](mailto:hoangxuanthuy2004@gmail.com), [nhutnm.17@grad.uit.edu.vn](mailto:nhutnm.17@grad.uit.edu.vn) and [thuannd@uit.edu.vn](mailto:thuannd@uit.edu.vn)

<sup>4</sup>Corresponding author: [thuannd@uit.edu.vn](mailto:thuannd@uit.edu.vn)

in generative medical imaging—particularly diffusion probabilistic models—have shown strong ability to generate realistic samples with improved training stability compared with adversarial methods [6, 7]. In principle, if high-quality 3D OCT volumes can be generated from CFP, multimodal classifiers could leverage both surface and depth cues without requiring OCT devices at the initial screening stage.

However, an important issue is that the usefulness of synthetic OCT should not be judged by visual realism alone. For clinical decision support, the central question is whether synthetic OCT *improves downstream disease recognition* under realistic challenges such as class imbalance, cross-dataset domain shift, and potential artifacts introduced by generation. Moreover, in a hybrid setting where one modality is real (fundus) and the other may be synthetic (OCT), robust fusion requires mechanisms that can *adaptively discount unreliable information*. To address this, we adopt EyeMoST [8], an uncertainty-aware multimodal framework. EyeMoST standardizes modality-specific representations and models fusion uncertainty, thereby mitigating modality dominance and reducing sensitivity to noisy auxiliary inputs.

In this work, we propose and systematically evaluate a unified multimodal pipeline that integrates (i) synthetic 3D OCT generation from CFP and (ii) uncertainty-aware fusion for retinal disease recognition. We compare three major generative paradigms—GAN-based, VAE-GAN-based, and diffusion-based synthesis—under an identical EyeMoST-based fusion framework. We conduct experiments on RFMiD [9] and ODIR-5K [10] using multiple CNN and Transformer backbones for both 2D and 3D branches. Through this controlled evaluation, we aim to clarify when synthetic OCT helps, when it fails to improve over fundus-only baselines, and which factors (e.g., synthetic-to-real domain gap and biomarker preservation) most strongly influence performance.

The main contributions of this study are summarized as follows:

- We develop a unified multimodal pipeline that enables Fundus+OCT learning in fundus-only settings by synthesizing 3D OCT volumes from CFP and integrating them into an uncertainty-aware fusion model.
- We provide a controlled comparison of three OCT synthesis families (3D GAN, 3D VAE-GAN, and fundus-conditioned 3D diffusion with CFG) within the same EyeMoST-based fusion framework, isolating the impact of the generator on downstream recognition.
- We benchmark multiple 2D/3D backbone configurations (CNN- and Transformer-based) on RFMiD and ODIR-5K and analyze failure cases where synthetic OCT does not consistently

surpass fundus-only base-lines due to domain mismatch, severe class imbalance, and imperfect pathology preservation.

## 2. RELATED WORKS

The development of computer-aided systems for retinal disease diagnosis has undergone a significant transition from conventional image-processing approaches to modern deep learning architectures. Convolutional neural networks (CNNs), such as Res2Net [11] and EfficientNet [12], have established themselves as foundational models for feature extraction from color fundus photography (CFP). Nevertheless, these CNN-based architectures are inherently constrained in modeling long-range pathological dependencies due to the limited receptive fields of convolutional operations. To address this limitation, attention-based architectures, including the Vision Transformer (ViT) and UNETR [13, 14], have been widely adopted. In particular, UNETR demonstrates notable advantages in three-dimensional data analysis because it captures global contextual information. This capability enables more effective identification of complex structural alterations in OCT volumes [14]. Despite their promise, the deployment of such models in clinical settings remains challenged by the scarcity of high-quality paired 2D–3D datasets.

To alleviate data scarcity, generative artificial intelligence (Generative AI) techniques have emerged as a dominant strategy for data augmentation [15]. Early efforts predominantly relied on generative adversarial networks (GANs) [16] to synthesize OCT images from 2D fundus inputs. Although GAN-based methods are capable of producing visually detailed images, they often suffer from training instability and may introduce spurious artifacts that lack clinical relevance. Such artifacts can undermine the reliability of downstream classification models, as GANs tend to prioritize perceptual realism over anatomical and pathological fidelity.

More recently, the transition toward diffusion-based models, particularly diffusion probabilistic models (DDPMs), has provided a more stable paradigm for medical image synthesis. Unlike the adversarial training mechanism of GANs, DDPMs learn the data distribution through a progressive denoising process, resulting in synthetic OCT images with improved structural consistency. Recent studies, such as Gen-3Diffusion [17], have further attempted to preserve subtle pathological biomarkers during the transformation between 2D and 3D modalities. However, despite these improvements, a substantial domain gap persists between synthetic and real-world OCT data. Most existing works primarily emphasize visual synthesis quality, while offering limited rigorous analysis of the actual contribution of synthetic data within multimodal fusion frameworks.

In addition to data generation challenges, effective

feature fusion across heterogeneous imaging modalities remains a critical technical hurdle. Conventional fusion strategies often lead to modality dominance, where the model disproportionately relies on one imaging source while neglecting complementary information from the other. The EyeMoST framework [8] addressed this issue by introducing feature standardization before multimodal integration. This standardization promotes balanced information utilization across modalities. Nevertheless, the interaction between synthetic data generated by hybrid GAN-DDPM approaches and powerful backbones such as ResNet3D has yet to be thoroughly investigated. To bridge this gap, the present study systematically evaluates the roles of both GAN- and DDPM-generated synthetic data. The study also uses an uncertainty-aware fusion mechanism to assess the feasibility and reliability of synthetic data usage in modern clinical diagnostic pipelines.

### 3. METHOD

#### 3.1 Data Overview

This study evaluates the proposed uncertainty-aware multi-modal framework using three publicly available ophthalmic datasets: RFMiD, ODIR-5K, and OLIVES. Fundus images serve as the primary modality. RFMiD and ODIR-5K provide fundus photographs with multi-label disease annotations, while OLIVES provides paired multimodal fundus-OCT data and is used as (i) an upper-bound reference with real OCT and (ii) the source dataset to train OCT synthesis models [2, 18, 6, 3, 19].

**RFMiD.** RFMiD (ISBI 2021) contains 3,200 color fundus images annotated with 46 conditions (45 diseases and 1 normal). In this study, to reduce label sparsity and stabilize multi-label training, the original label space is consolidated into 8 disease groups: *DR*, *ARMD*, *MH*, *DN*, *MYA*, *TSLN*, *ODC*, *OTHER*. The *NORMAL* label is defined by complement: *NORMAL=1* if no disease label is positive otherwise the sample is *Abnormal*. Under this setting, RFMiD includes 2,531 abnormal images (79.09%) and 669 normal images (20.91%). Label frequencies are highly imbalanced (e.g., *OTHER* and *DR* are dominant). RFMiD also exhibits non-trivial co-occurrence: most samples have a single disease, but a considerable fraction contains 2–3 diseases.

**ODIR-5K.** ODIR-5K (ODIR 2019) is a clinical screening dataset collected across multiple hospitals. Each subject has paired left/right fundus images with demographic information and diagnostic keywords. We use the labeled subset only (since the official test split has no labels), consisting of 3,500 subjects (7,000 fundus images). ODIR-5K is multi-label with 8 categories: *N* (*Normal*), *D* (*Diabetes*), *G* (*Glaucoma*), *C* (*Cataract*), *A* (*AMD*), *H* (*Hypertension*), *M* (*Myopia*), *O* (*Other*). The dataset is imbalanced: 2,280 images are normal (32.57%) and 4,720 are abnormal

(67.43%). *N*, *D*, and *O* are the most frequent labels, while *H*, *A*, and *M* are relatively rare. Multi-label co-morbidity is present: 50.69% of images contain exactly one disease, and the remainder contains two or more diseases.

**OLIVES.** OLIVES (NeurIPS 2022) is a multimodal fundus-OCT dataset focused on diabetic retinal diseases (DR/DME) with fine-grained clinical and biomarker annotations. Each visit includes one fundus IR image and one OCT volume of 49 B-scans (a 3D cube). In our experiments, OLIVES comprises 87 subjects and 96 eyes, totaling 9,408 OCT B-scans. Besides disease labels, OLIVES provides 16 OCT biomarkers with strong natural imbalance (some biomarkers are common while others are rare). We use OLIVES to (i) evaluate fusion with real OCT as an upper bound, and (ii) train generative models that synthesize 3D OCT from fundus images for enabling multimodal learning on fundus-only datasets.

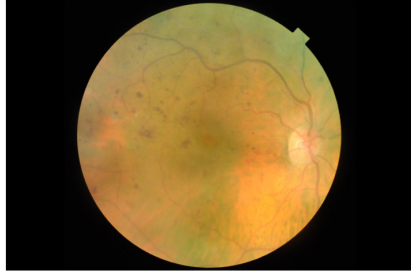
**Synthetic OCT for fundus-only datasets.** Since RFMiD and ODIR-5K do not provide real OCT, we generate synthetic 3D OCT volumes from fundus images using deep generative models (e.g., GAN-based and diffusion-based approaches). Synthetic OCT is treated as an auxiliary modality and is fed into the same pipeline as real OCT during training and inference.

**Data splitting protocol.** All experiments follow stratified 5-fold cross-validation with subject-level splitting to prevent leakage and to maintain label balance across folds. In each fold, approximately 64%/16%/20% of samples are used for train/validation/test. For OLIVES, paired fundus and OCT from the same subject are always assigned to the same split. For RFMiD and ODIR-5K, synthetic OCT strictly follows the split of its corresponding fundus image no synthetic OCT derived from validation/test samples is used to train the models. Class imbalance handling (if applied) is performed only on the training split. The validation and test splits preserve the original label distributions.

#### 3.2 Model

##### 3.2.1 Overall System Architecture

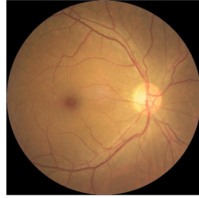
The proposed system primarily focuses on multimodal eye disease recognition (Fundus + OCT) within a unified architecture. It adopts a dual-branch parallel design, where the fundus branch extracts surface-level features using 2D backbones (CNNs or Vision Transformers), while the OCT branch captures depth-aware structural information through 3D architectures such as ResNet3D or UNETR. The two modalities are integrated via an uncertainty-aware fusion mechanism based on the EyeMoST framework, in which modality-specific representations are modeled using Student's t-distributions. This design allows the system to dynamically adjust the contribution of fundus and OCT features according to their es-



(a) RFMiD fundus example.

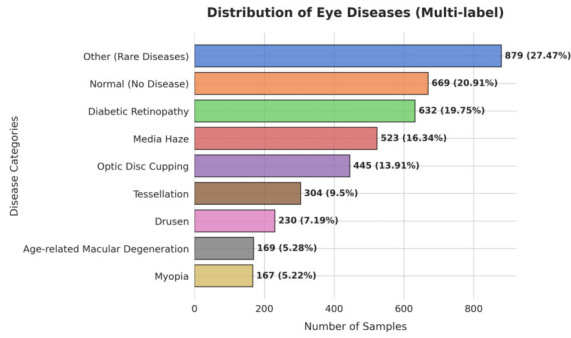


(b) ODIR-5K left eye.

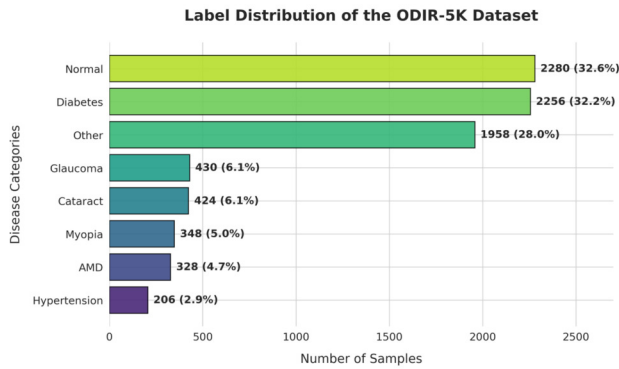


(c) ODIR-5K right eye.

**Fig.1:** Representative fundus images from RFMiD and ODIR-5K.



(a) RFMiD label distribution.



(b) ODIR-5K label distribution.

**Fig.2:** Dataset label distributions (RFMiD and ODIR-5K).

| OLIVES Dataset Summary |           |                 |                  |                                                                                                                                                                                                                                                                                                                                                                        |
|------------------------|-----------|-----------------|------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Modality               | Per Visit | Per Eye         | Total Statistics | Overview                                                                                                                                                                                                                                                                                                                                                               |
| OCT                    | 49        | $N_P \times 49$ | 78189            | <b>General:</b><br>96 Eyes, Visits every 4-16 weeks,<br>Average 16 visits and 7 injections/patient<br><b>Clinical Labels obtained every visit:</b><br>BCVA, CST, Patient ID, Eye ID<br><b>Biomarkers labeled:</b><br>IRHRE, FAVF, IRF, DRT/ME PAVE, VD,<br>Preretinal Tissue, EZ Disruption, IR Hemorrhages,<br>SRF, VMT, Atrophy, SHRM, RPE Disruption,<br>Serous PED |
| Fundus                 | 1         | $N_P$           | 1268             |                                                                                                                                                                                                                                                                                                                                                                        |
| Clinical               | 4         | $N_P \times 4$  | 5072             |                                                                                                                                                                                                                                                                                                                                                                        |
| Biomarker              | 16        | 1568            | 150528           |                                                                                                                                                                                                                                                                                                                                                                        |

**Fig.3:** OLIVES dataset summary.

timated reliability, thereby mitigating modality dominance and reducing the impact of noise.

In scenarios where real OCT data are unavailable, a fundus-to-OCT generation module is activated as an auxiliary component to provide complementary volumetric cues for the OCT branch. Synthetic 3D OCT volumes are generated using multiple paradigms (GAN, VAE-GAN, and diffusion-based models), with the diffusion-based generator selected as the default due to its more stable optimization and more consistent downstream fusion performance. This auxiliary design enables the system to preserve the multi-modal recognition framework and to produce multi-label predictions accompanied by explicit uncertainty estimates, making it suitable for clinical screening in resource-limited settings.

### 3.2.2 Res2Net-50 (2D) and ResNet3D (3D) Default backbones in the proposed model.

In our proposed multimodal framework, we adopt Res2Net-50 as the *primary* 2D backbone for fundus feature extraction and ResNet3D as the *primary* 3D backbone for OCT volumetric representation. This backbone pair constitutes the default configuration of the EyeMoST-based fusion model in our experiments, while other CNN/Transformer backbones are treated as comparative baselines.

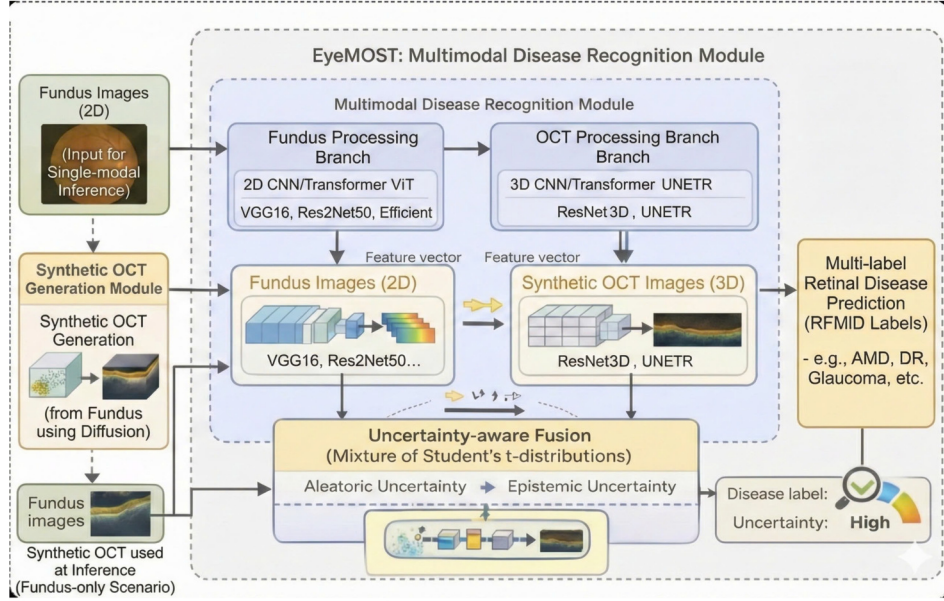
Deep convolutional neural networks (CNNs) have achieved strong performance in medical image analysis however, increasing depth may introduce optimization challenges such as vanishing or exploding gradients, which can slow convergence and degrade accuracy. Residual learning mitigates these issues by introducing identity shortcut connections, enabling stable training of very deep networks [11]. Instead of learning the mapping  $H(\mathbf{x})$  directly, a residual block learns the residual function  $F(\mathbf{x}) = H(\mathbf{x}) - \mathbf{x}$ , and computes the output as:

$$\mathbf{y} = \mathbf{x} + F(\mathbf{x}), \quad (1)$$

where  $\mathbf{x}$  and  $\mathbf{y}$  denote the input and output feature maps, respectively.

**2D fundus backbone (Res2Net-50).** The fundus branch employs Res2Net-50 to encode surface-level retinal appearance from color fundus photographs. Res2Net-50 follows the Res2Net bottleneck design, where each bottleneck block consists of three convolutions: a  $1 \times 1$  layer for channel reduction, a  $3 \times 3$  layer for spatial feature extraction, and a final  $1 \times 1$  layer for channel restoration. This bottleneck structure reduces computation while preserving representational capacity, which is essential for large-scale screening datasets.

Beyond standard ResNet, Res2Net introduces explicit multi-scale representation within a single residual block by splitting channels into several groups (scales) and performing hierarchical feature interactions across these groups. This design is well aligned



**Fig.4:** Overall system architecture of the proposed multimodal eye disease recognition framework.

with fundus disease recognition, where pathological cues appear at multiple spatial scales (e.g., small microaneurysms versus larger hemorrhages/exudates). After the final stage, global average pooling is applied to obtain a compact fundus embedding:

$$\mathbf{f} = E_f(x_f) \in \mathbb{R}^{d_f}, \quad (2)$$

where  $x_f$  is the fundus image and  $E_f(\cdot)$  denotes the ResNet-50 encoder.

**3D OCT backbone (ResNet3D).** To model volumetric OCT data (real or synthetically generated), we employ a 3D residual encoder (ResNet3D) in the OCT branch. ResNet3D extends residual learning to 3D inputs by replacing 2D operators with their 3D counterparts (Conv3D and Pool3D), thereby enabling joint feature learning along depth, height, and width. This design is suitable for OCT volumes, where clinically meaningful patterns are distributed across adjacent B-scans and retinal layers. Given an OCT volume  $x_o \in \mathbb{R}^{D \times H \times W}$ , the OCT branch produces a volumetric embedding:

$$\mathbf{o} = E_o(x_o) \in \mathbb{R}^{d_o}, \quad (3)$$

where  $E_o(\cdot)$  denotes the ResNet3D encoder. By capturing inter-slice dependencies and depth-resolved morphology, ResNet3D provides complementary structural cues that are effective for OCT-based classification [19].

**Role in multimodal fusion.** Within the proposed system, ResNet-50 (2D) and ResNet3D (3D) serve as modality-specific encoders that extract embeddings  $\mathbf{f}$  and  $\mathbf{o}$  from fundus and OCT inputs, respectively. These representations are subsequently integrated by the uncertainty-aware fusion module (EyeMoST) described in the following sections. Importantly,

since the OCT modality may be synthetically generated (DCGAN/VAE-GAN/Diffusion), uncertainty-aware fusion allows the model to adjust the contribution of OCT features when their reliability is lower, enabling controlled multimodal integration rather than naive feature aggregation.

### 3.2.3 VGG16

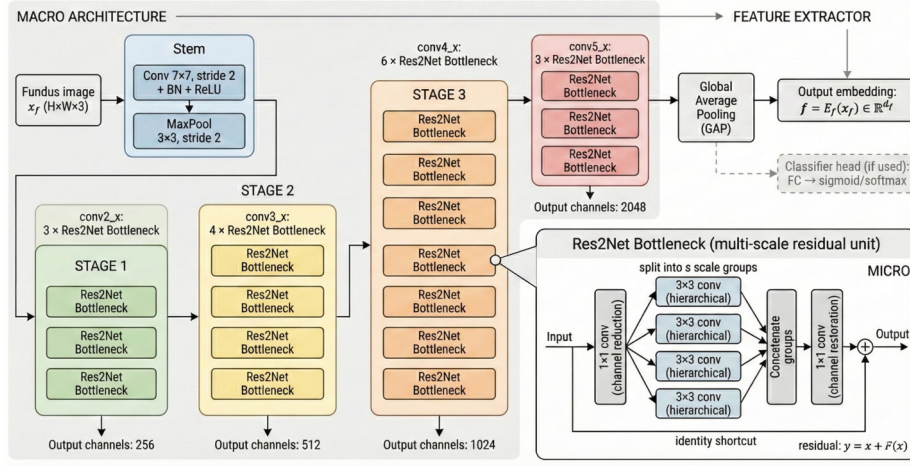
VGG16 is a deep convolutional neural network (CNN) designed to extract hierarchical features through the consistent use of small  $3 \times 3$  convolutional filters. In this study, VGG16 is deployed in the 2D fundus branch to learn fundamental morphological representations of the retina. The operation of a convolutional layer can be formulated as:

$$y_j^l = \sigma \left( \sum_{i \in M_j} w_{ij}^{(l)} * x_i^{(l-1)} + b_j^{(l)} \right) \quad (4)$$

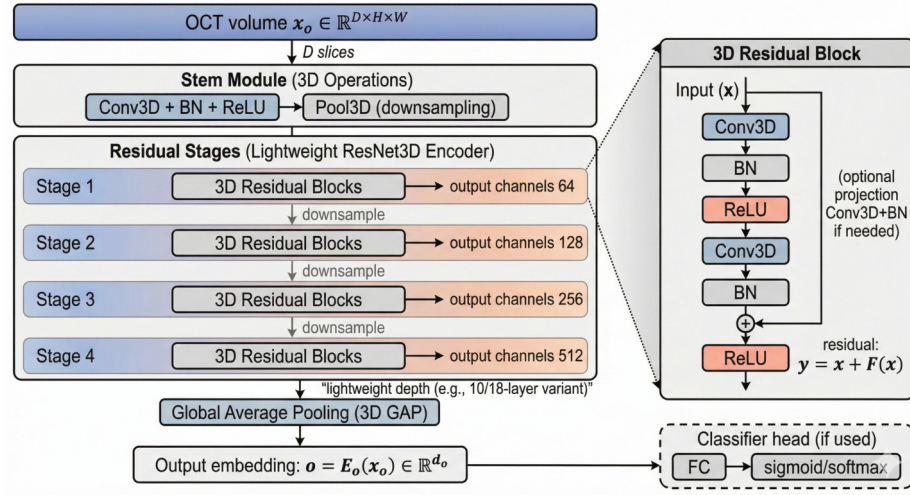
where  $\sigma$  denotes the ReLU activation function,  $w_{ij}^{(l)}$  represents the filter weights at layer  $l$ , and  $*$  indicates the 2D convolution operation. VGG16 serves as an important baseline for evaluating the effectiveness of stacking simple convolutional layers compared to more advanced architectures in retinal disease recognition. VGG16 is included as a conventional CNN baseline to assess the effectiveness of deeper and more recent feature extractors in retinal disease recognition.

### 3.2.4 EfficientNet-B3

EfficientNet-B3 represents a major advancement in optimizing CNN performance through simultaneous scaling along three dimensions: network depth ( $d$ ), width ( $w$ ), and input resolution ( $r$ ). This balanced



**Fig.5:** *ResNet-50 architecture for fundus feature extraction.*



**Fig.6:** *ResNet3D architecture for OCT volumetric feature extraction.*

scaling strategy is governed by a compound coefficient  $\phi$ :

$$\text{s.t. } d = \alpha^\phi, w = \beta^\phi, r = \gamma^\phi, \quad (5)$$

where the constants  $\alpha, \beta$ , and  $\gamma$  are determined via Neural Architecture Search (NAS). By leveraging Mobile Inverted Bottleneck Convolution (MB-Conv) blocks, EfficientNet-B3 enables the extraction of fine-grained pathological features—such as microaneurysms or hemorrhages—from fundus images with high accuracy, while maintaining significantly lower computational cost compared to traditional CNN architectures. EfficientNet-B3 is used as an efficient CNN baseline to evaluate whether compound-scaled convolutional features improve fundus-based disease recognition.

### 3.2.5 Vision Transformer (ViT)

In contrast to convolution-based mechanisms, the Vision Transformer (ViT) processes images as sequences of flattened patches. Its core component is

the Multi-Head Self-Attention (MSA) layer, which allows the model to learn global contextual relationships across all regions of a fundus image. The scaled dot-product attention mechanism is formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (6)$$

where  $Q, K$ , and  $V$  denote the Query, Key, and Value matrices, respectively. ViT overcomes the receptive field limitations inherent to CNNs, making it particularly effective for identifying pathological patterns that are spatially dispersed across the retinal surface. ViT-B/16 is used as a Transformer-based fundus baseline to examine the effect of global self-attention on retinal disease recognition.

### 3.2.6 UNETR (Transformer-based 3D Architecture)

UNETR (U-Net Transformer) is implemented in the 3D OCT processing branch to extract layer-

wise morphological information from volumetric data. UNETR employs a pure Transformer encoder to capture long-range spatial dependencies in three-dimensional space, which is subsequently connected to a convolutional decoder through multi-scale skip connections:

$$z_i = \text{MSA}(\text{LN}(z_{i-1})) + z_{i-1}. \quad (7)$$

This architecture enables UNETR to preserve fine spatial details while simultaneously modeling the global structure of OCT volumes. UNETR serves as an advanced benchmark for direct comparison with the proposed ResNet3D model. This comparison supports an in-depth evaluation of whether pure Transformer-based architectures or residual 3D convolutional networks are more effective for processing synthetic OCT data. UNETR is used as a Transformer-based 3D baseline for comparison with ResNet3D. This baseline evaluates whether long-range volumetric attention provides advantages over residual 3D convolutional encoding for synthetic OCT-based classification.

### 3.3 Synthetic 3D OCT Generation

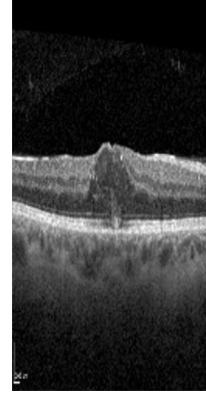
To enable multimodal learning in fundus-only scenarios, we synthesize 3D OCT volumes and use them as an auxiliary modality for downstream disease recognition. We implement and evaluate three generative paradigms in our system: (i) a 3D GAN trained with a Wasserstein objective and gradient penalty (DCGAN/WGAN-GP style), (ii) a 3D VAE-GAN, and (iii) a fundus-conditioned *latent* diffusion model (DDPM) equipped with classifier-free guidance (CFG). These methods represent three major families of generative models and allow a systematic comparison in terms of training stability, anatomical realism, and their contribution to uncertainty-aware multimodal classification.

Let  $x_f$  denote a fundus image and  $x_o \in \mathbb{R}^{D \times H \times W}$  denote the corresponding OCT volume (in our implementation,  $D=49$ ,  $H=256$ ,  $W=128$ ). In our pipeline, all generators are trained on OLIVES to learn a realistic OCT manifold. They are then used to synthesize OCT volumes for datasets where only fundus images are available (e.g., ODIR-5K and RFMiD). The synthesized OCT is used *only* as auxiliary input to the multimodal classifier when real OCT is available, synthesis is bypassed and  $x_o$  is directly utilized.

**3D GAN-based OCT generation (DCGAN/WGAN-GP style).** We first implement a 3D GAN that maps latent noise  $z \sim \mathcal{N}(0, I)$  to an OCT volume:

$$\hat{x}_o = G(z). \quad (8)$$

A 3D discriminator  $D$  is trained to distinguish real OCT volumes from generated ones. To improve stability for high-dimensional volumetric synthesis, we adopt a Wasserstein objective with gradient penalty



**Fig.7:** Representative real OCT slice from OLIVES used as reference for synthetic OCT generation.

(WGAN-GP):

$$\mathcal{L}_D = \mathbb{E}_z[D(G(z))] - \mathbb{E}_{x_o}[D(x_o)] + \lambda_{gp} \mathbb{E}_{\tilde{x}}[(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2]. \quad (9)$$

where  $\tilde{x}$  is sampled along straight lines between real and generated volumes. The generator is trained with the adversarial loss

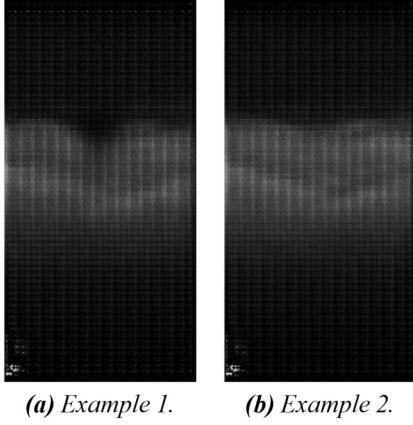
$$\mathcal{L}_{adv} = -\mathbb{E}_z[D(G(z))], \quad (10)$$

augmented with structure-preserving regularizers to encourage anatomically plausible retinal layers, e.g.,

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_1 \|x_o - \tilde{x}_o\|_1 + \lambda_{ssim} \text{SSIM}_{3D}(x_o, \tilde{x}_o). \quad (11)$$

While GAN-based synthesis can produce sharp samples, it is typically sensitive to hyperparameters and may exhibit instability or mode collapse when modeling full 3D retinal structure. In our implementation, the generator progressively upsamples a low-dimensional latent representation into a 3D OCT volume using a coarse-to-fine decoding pipeline. The discriminator is designed as a 3D critic that evaluates both real and generated OCT volumes and outputs patch-level realism scores, which are aggregated into a scalar Wasserstein score.

To stabilize training, the gradient penalty is applied to interpolated samples between real and generated volumes, thereby enforcing the Lipschitz constraint required by the WGAN-GP formulation. This design enables adversarial learning of volumetric retinal structures while reducing training instability. Nevertheless, adversarial supervision alone may not preserve clinically relevant retinal-layer morphology or disease-specific biomarkers therefore, reconstruction and structural consistency losses are included to further constrain the generated OCT volumes.



**Fig.8:** Representative synthesized 3D OCT samples generated by the GAN-based model.

**3D VAE-GAN-based OCT generation.** To improve training stability and learn a regularized latent representation, we further explore a 3D VAE-GAN formulation. An encoder  $E$  maps an OCT volume to a latent distribution  $q(z|x_o)$ , and a generator (decoder) reconstructs the volume from sampled latent codes:

$$(\mu, \log \sigma^2) = E(x_o), \quad Z = \mu + \sigma \odot \epsilon, \epsilon \sim \mathcal{N}(0, I),$$

$$\tilde{x}_o = G(z).$$
(12)

Training combines (i) reconstruction fidelity, (ii) KL regularization toward a standard Gaussian prior, and (iii) adversarial supervision via a 3D discriminator to preserve fine structural details:

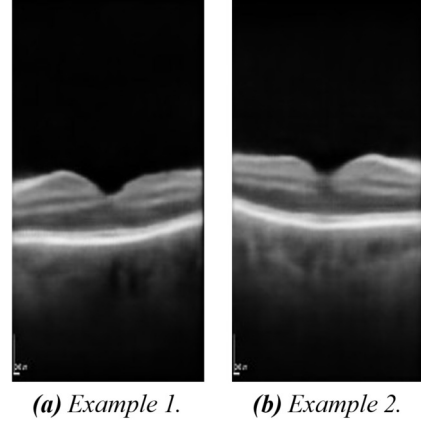
$$\mathcal{L}_{\text{VAE-GAN}} = \lambda_1 \|x_o - \tilde{x}_o\|_1 + \lambda_{\text{ssim}} \text{SSIM}_{3D}(x_o - \tilde{x}_o) + \beta_{\text{kl}} KL(q(z|x_o) \|\mathcal{N}(0, I)) + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}.$$
(13)

where  $\mathcal{L}_{\text{adv}}$  is implemented using a WGAN-style critic. Compared to a plain GAN, the explicit reconstruction constraint generally stabilizes training and improves coverage of anatomical variations, while the adversarial term helps mitigate oversmoothing and enhances retinal-layer sharpness.

**Fundus-conditioned latent diffusion (DDPM with CFG).** We employ a fundus-conditioned latent diffusion model as the third generative paradigm and as the primary generator in our final system. This design decouples (a) learning a compact 3D OCT latent space and (b) learning fundus-guided generation in that latent space. Concretely, a lightweight 3D VAE encodes the OCT volume into a latent tensor  $z$ ,

$$z = \text{Enc}(x_o),$$
(14)

and a 3D U-Net  $\epsilon_\theta$  is trained to predict the injected



**Fig.9:** Representative synthesized 3D OCT samples generated by the 3D VAE-GAN.

noise at diffusion step  $t$  under conditioning from a fundus embedding  $c(x_f)$ :

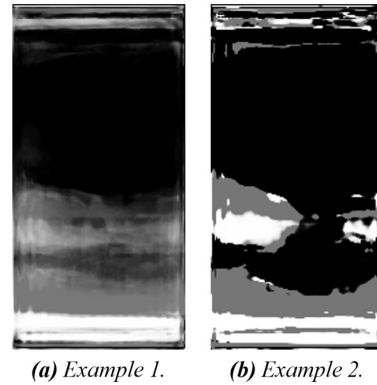
$$z_t = \sqrt{\alpha_t} z + \sqrt{1 - \alpha_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$
(15)

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z, t, \epsilon [\|\epsilon - \epsilon_\theta(z_t, t, c(x_f))\|_2^2].}$$
(16)

We adopt classifier-free guidance (CFG) by randomly dropping the conditioning vector during training (conditioning dropout) to learn both conditional and unconditional score estimates within a single network. At inference, conditional and unconditional predictions are combined to strengthen fundus-consistent retinal structures, and the predicted latent is decoded back to the OCT space:

$$\hat{x}_o = \text{Dec}(\hat{z}_0).$$
(17)

Compared with GAN and VAE-GAN baselines, the diffusion model provides markedly more stable optimization and typically yields more coherent volumetric layer organization. However, this improvement comes with higher sampling cost.



**Fig.10:** Representative synthesized 3D OCT samples generated by the diffusion model.

**Usage in the downstream multimodal classifier.** Synthetic OCT volumes are used exclusively as auxiliary inputs to the multimodal disease recognition model. For datasets without OCT, we generate a pseudo-OCT volume  $\hat{x}_o$  for each fundus image and feed  $\{x_f, \hat{x}_o\}$  into the fusion network. When real OCT is available, we bypass synthesis and directly use  $\{x_f, x_o\}$ . This unified setup enables a controlled comparison of 3D generative strategies and their effects on uncertainty-aware multimodal eye disease classification.

**Primary generator in our system.** Although we implement and evaluate GAN- and VAE-GAN-based generators as competitive baselines, the diffusion-based model is selected as the primary OCT synthesis module in our final system due to its superior training stability and its ability to generate more anatomically coherent volumetric structures. Therefore, unless otherwise stated, all subsequent fundus-only multimodal results in this work adopt diffusion-generated OCT as the default synthetic modality.

## 4. RESULTS AND DISCUSSION

Unless otherwise stated, all reported results are averaged over 5-fold cross-validation. We report three metrics: F1-macro, Recall, and AUPRC.

### 4.1 Evaluation on the OLIVES Dataset with Real OCT

We first evaluate multimodal models on OLIVES using real OCT volumes to provide an upper-bound reference for fundus-OCT fusion.

**Table 1:** Performance on OLIVES with real OCT (5-fold average). Metrics: F1-macro, Recall, AUPRC.

| Model                    | F1-macro | Recall | AUPRC |
|--------------------------|----------|--------|-------|
| Res2Net-50 + ResNet3D-10 | 0.971    | 0.974  | 0.995 |
| EffNet-B3 + UNETR        | 0.960    | 0.979  | 0.995 |
| ViT-B/16 + UNETR         | 0.955    | 0.975  | 0.995 |
| VGG16 + ResNet3D-10      | 0.953    | 0.943  | 0.990 |

Table 1 shows consistently strong performance across multimodal architectures when real OCT is available. Res2Net-50 + ResNet3D-10 provides the best overall balance with the highest F1-macro, while EffNet-B3 + UNETR attains the highest Recall. The high AUPRC across methods suggests that real OCT provides informative complementary signals, resulting in stable fusion performance.

### 4.2 Fundus-Only Baseline on OLIVES

To quantify the contribution of OCT, we evaluate fundus-only baselines on OLIVES by removing the OCT branch under the same 5-fold protocol.

As shown in Table 2, fundus-only models maintain high Recall and competitive AUPRC on OLIVES, indicating that fundus images alone provide strong

**Table 2:** Fundus-only performance on OLIVES (5-fold average). Metrics: F1-macro, Recall, AUPRC.

| Model      | F1-macro | Recall | AUPRC |
|------------|----------|--------|-------|
| Res2Net-50 | 0.896    | 0.948  | 0.960 |
| ViT-B/16   | 0.885    | 0.937  | 0.955 |
| EffNet-B3  | 0.876    | 0.935  | 0.943 |
| VGG16      | 0.816    | 0.930  | 0.911 |

diagnostic cues. However, compared with multimodal learning using real OCT (Table 1), fundus-only settings show lower F1-macro and AUPRC, suggesting that OCT helps improve label balance and robustness under class imbalance. Among fundus-only backbones, Res2Net-50 provides the strongest overall balance.

### 4.3 EyeMoST on OLIVES with Synthetic OCT

We evaluate EyeMoST (Res2Net-50 + ResNet3D-10) on OLIVES using synthetic OCT generated by diffusion, VAE-GAN, and DCGAN. All settings are fixed only the OCT source varies.

**Table 3:** EyeMoST on OLIVES with synthetic OCT (5-fold average). Metrics: F1-macro, Recall, AUPRC.

| OCT source | F1-macro | Recall | AUPRC |
|------------|----------|--------|-------|
| Diffusion  | 0.912    | 0.958  | 0.972 |
| VAE-GAN    | 0.899    | 0.950  | 0.963 |
| DCGAN      | 0.885    | 0.942  | 0.950 |

Table 3 shows that diffusion yields the strongest synthetic-OCT performance on OLIVES across F1-macro, Recall, and AUPRC. VAE-GAN performs below diffusion DCGAN shows the lowest results. Overall, the generator ranking on OLIVES is Diffusion > VAE-GAN > DCGAN.

### 4.4 Overall Comparison on OLIVES by OCT Source

We summarize OLIVES results across settings: fundus-only vs EyeMoST with synthetic OCT vs EyeMoST with real OCT. Res2Net-50 is used as the fundus-only reference, and EyeMoST is fixed for multimodal settings.

**Table 4:** Summary on OLIVES by OCT source (5-fold average). Metrics: F1-macro, Recall, AUPRC.

| Setting                  | F1-macro | Recall | AUPRC |
|--------------------------|----------|--------|-------|
| Fundus-only (Res2Net-50) | 0.896    | 0.948  | 0.960 |
| EyeMoST + Diffusion      | 0.912    | 0.958  | 0.972 |
| EyeMoST + VAE-GAN        | 0.899    | 0.950  | 0.963 |
| EyeMoST + DCGAN          | 0.885    | 0.942  | 0.950 |
| EyeMoST + Real OCT       | 0.971    | 0.974  | 0.995 |

Table 4 supports three points: real OCT provides the strongest upper bound diffusion-based synthetic

OCT improves over the best fundus-only baseline on OLIVES and the ranking Diffusion > VAE-GAN > DCGAN remains consistent among synthetic generators.

#### 4.5 Fundus-Only Results on RFMiD and ODIR-5K

We establish fundus-only baselines on RFMiD and ODIR-5K as references for subsequent fusion experiments.

**Table 5:** Fundus-only performance on RFMiD (5-fold average). Metrics: F1-macro, Recall, AUPRC.

| Model      | F1-macro | Recall | AUPRC  |
|------------|----------|--------|--------|
| Res2Net-50 | 0.7407   | 0.7594 | 0.8090 |
| VGG16      | 0.7259   | 0.7158 | 0.7983 |
| EffNet-B3  | 0.7083   | 0.7114 | 0.7920 |
| ViT-B/16   | 0.7109   | 0.6896 | 0.7771 |

On RFMiD (Table 5), Res2Net-50 yields the strongest fundus-only baseline across all three metrics.

**Table 6:** Fundus-only performance on ODIR-5K (5-fold average). Metrics: F1-macro, Recall, AUPRC.

| Model      | F1-macro | Recall | AUPRC  |
|------------|----------|--------|--------|
| Res2Net-50 | 0.5386   | 0.5265 | 0.5782 |
| VGG16      | 0.5938   | 0.6369 | 0.6170 |
| EffNet-B3  | 0.5614   | 0.6214 | 0.5856 |
| ViT-B/16   | 0.5532   | 0.5412 | 0.5682 |

On ODIR-5K (Table 6), VGG16 provides the strongest fundus-only baseline across F1-macro, Recall, and AUPRC.

Notably, a performance discrepancy is observed in the Fundus-only baselines across the three benchmarks, with F1-macro scores of 0.896, 0.7407, and 0.5386 for OLIVES, RFMiD, and ODIR-5K, respectively. Although the core architecture (Res2Net-50) remains identical, these variances are inherently driven by the distinct task complexity and dataset-specific characteristics.

First, the classification objective differs significantly: OLIVES focuses on a binary detection task (DME vs. non-DME), which represents a relatively lower hypothesis space. In contrast, RFMiD and ODIR-5K require multi-label classification across eight categories. The necessity of identifying co-occurring pathologies and overlapping clinical signs in these datasets expands the label cardinality, thereby increasing the difficulty of maintaining a high macro-averaged F1 score.

Second, class imbalance and image variability play critical roles. ODIR-5K, as a real-world multi-center dataset, exhibits extreme class imbalance and high intra-class heterogeneity due to varying acquisition protocols across clinical sites. Such factors, coupled

with the long-tailed distribution of rare pathologies, disproportionately degrade macro-averaged metrics compared with controlled, single-modality environments like OLIVES.

#### 4.6 Fusion with Synthetic OCT on RFMiD and ODIR-5K

We evaluate multimodal fusion (fundus + synthetic OCT) on RFMiD and ODIR-5K, where OCT is generated by diffusion, VAE-GAN, or DCGAN.

**Table 7:** Fusion on RFMiD with synthetic OCT (5-fold average). Metrics: F1-macro, Recall, AUPRC.

| Model                    | Metric   | Diffusion | VAE-GAN | DCGAN  |
|--------------------------|----------|-----------|---------|--------|
| Res2Net-50 + ResNet3D-10 | F1-macro | 0.6852    | 0.6615  | 0.6569 |
|                          | Recall   | 0.6425    | 0.6086  | 0.5937 |
|                          | AUPRC    | 0.7250    | 0.7243  | 0.7185 |
| VGG16 + ResNet3D-10      | F1-macro | 0.6715    | 0.6550  | 0.6480 |
|                          | Recall   | 0.6510    | 0.6150  | 0.6050 |
|                          | AUPRC    | 0.7180    | 0.7150  | 0.7120 |
| EffNet-B3 + UNETR        | F1-macro | 0.6650    | 0.6480  | 0.6410 |
|                          | Recall   | 0.6350    | 0.6020  | 0.5850 |
|                          | AUPRC    | 0.7100    | 0.7080  | 0.7050 |
| ViT-B/16 + UNETR         | F1-macro | 0.6580    | 0.6420  | 0.6350 |
|                          | Recall   | 0.6280    | 0.5950  | 0.5820 |
|                          | AUPRC    | 0.7050    | 0.7020  | 0.7000 |

On RFMiD (Table 7), diffusion-based fusion yields the strongest results across the three metrics for all backbone pairs. The best configuration (Res2Net-50 + ResNet3D-10) achieves F1-macro = 0.6852, Recall = 0.6425, and AUPRC = 0.7250. Overall, the generator ranking on RFMiD is consistent: Diffusion > VAE-GAN > DCGAN.

**Table 8:** Fusion on ODIR-5K with synthetic OCT (5-fold average). Metrics: F1-macro, Recall, AUPRC.

| Model                    | Metric   | Diffusion | VAE-GAN | DCGAN  |
|--------------------------|----------|-----------|---------|--------|
| Res2Net-50 + ResNet3D-10 | F1-macro | 0.5276    | 0.5150  | 0.5085 |
|                          | Recall   | 0.4572    | 0.4450  | 0.4380 |
|                          | AUPRC    | 0.5852    | 0.5780  | 0.5720 |
| VGG16 + ResNet3D-10      | F1-macro | 0.5185    | 0.5080  | 0.5015 |
|                          | Recall   | 0.4480    | 0.4320  | 0.4250 |
|                          | AUPRC    | 0.5750    | 0.5650  | 0.5600 |
| EffNet-B3 + UNETR        | F1-macro | 0.5120    | 0.5010  | 0.4950 |
|                          | Recall   | 0.4350    | 0.4250  | 0.4180 |
|                          | AUPRC    | 0.5680    | 0.5580  | 0.5550 |
| ViT-B/16 + UNETR         | F1-macro | 0.5050    | 0.4950  | 0.4880 |
|                          | Recall   | 0.4280    | 0.4180  | 0.4100 |
|                          | AUPRC    | 0.5620    | 0.5520  | 0.5480 |

On ODIR-5K (Table 8), diffusion again provides the strongest fusion results among synthetic sources, but overall performance remains limited. In most configurations, F1-macro and Recall are lower than fundus-only baselines, and AUPRC does not improve consistently, indicating that synthetic OCT does not reliably add discriminative signal under ODIR-5K imbalance and heterogeneity. Nevertheless, the generator ranking remains consistent on ODIR-5K: Diffusion > VAE-GAN > DCGAN.

#### 4.7 Analysis of the Impact of Synthetic OCT

A direct comparison between fundus-only baselines and fusion models with synthetic OCT shows that generated OCT does not consistently improve F1-macro, Recall, or AUPRC on fundus-only datasets (RFMiD and ODIR-5K). On RFMiD, the best fundus-only baseline (Res2Net-50) achieves F1-macro = 0.7407, Recall = 0.7594, and AUPRC = 0.8090 (Table 5), whereas the best fusion configuration with diffusion reaches F1-macro = 0.6852, Recall = 0.6425, and AUPRC = 0.7250 (Table 7). On ODIR-5K, the best fundus-only model (VGG16) also attains higher F1-macro, Recall, and AUPRC than most fusion configurations (Table 6 vs. Table 8).

A clearer interpretation can be obtained by analyzing RFMiD and ODIR-5K separately, as well as comparing the two table pairs.

A comparison across the two table pairs further reveals that the impact of synthetic OCT differs between RFMiD and ODIR-5K. For RFMiD (Table 5 vs. Table 7), the best fusion result with diffusion-generated OCT is lower than the best fundus-only baseline, but the gap remains moderate. In contrast, for ODIR-5K (Table 6 vs. Table 8), the performance gap becomes larger, with more pronounced degradation in F1-macro and Recall. This indicates that the negative impact of synthetic OCT is dataset-dependent and becomes stronger when the target dataset is more heterogeneous and less aligned with the synthesis training domain.

For RFMiD (Table 5 vs. Table 7), the fusion model with synthetic OCT does not outperform the fundus-only baseline. This result suggests that fundus images already contain strong discriminative information for this dataset, allowing Res2Net-50 to achieve competitive performance without additional modalities. In this setting, synthetic OCT does not provide sufficiently reliable complementary cues. Although the generated OCT images may appear visually plausible, they may fail to preserve critical disease-specific biomarkers, such as subtle retinal layer abnormalities or fluid-related features. As a result, the auxiliary modality may introduce ambiguous or noisy signals, which negatively affect the fusion process and lead to reduced performance.

For ODIR-5K (Table 6 vs. Table 8), the performance drop is more pronounced. The best fundus-only model (VGG16) surpasses all fusion configurations with synthetic OCT across multiple evaluation metrics. This can be attributed to the higher heterogeneity and class imbalance of ODIR-5K, which make cross-dataset generalization more challenging. Since the OCT synthesis model is trained on OLIVES, a dataset with different disease distributions and imaging characteristics, the generated pseudo-OCT may not align well with the target domain of ODIR-5K. Consequently, the synthetic modality may introduce unstable or misleading structural information, which

further limits the effectiveness of multimodal fusion.

In addition, the fusion mechanism inherently assumes that all modalities provide complementary and reliable information. However, when the auxiliary modality (synthetic OCT) contains inaccuracies or domain inconsistencies, it can degrade the overall representation rather than enhance it. This effect is particularly evident when the baseline fundus model already achieves strong performance, leaving limited room for improvement unless the additional modality is of sufficiently high quality.

Despite these limitations, diffusion-based synthesis consistently produces the strongest results among all generative approaches. For both RFMiD and ODIR-5K, diffusion achieves higher F1-macro, Recall, and AUPRC than VAE-GAN and DCGAN in fusion settings. This suggests that improved synthesis quality leads to more reliable auxiliary signals and partially mitigates the negative impact of domain mismatch.

Overall, the results in Tables 5–8 indicate that fusion with synthetic OCT does not consistently improve performance on RFMiD and ODIR-5K, and in some cases leads to performance degradation. However, this does not imply that multimodal fusion is inherently ineffective. Instead, it highlights that the success of multimodal learning critically depends on the fidelity, biomarker preservation, and domain alignment of the synthetic modality. Therefore, future improvements in synthesis quality and cross-dataset generalization are essential to fully realize the potential of synthetic OCT in multimodal eye disease recognition.

## 5. CONCLUSION

In this study, we proposed and systematically evaluated a multimodal deep learning framework for retinal disease classification that integrates fundus images and 3D OCT data. To address the lack of real OCT data in commonly used public datasets, synthetic OCT volumes were generated and incorporated as an auxiliary modality. The overall system was designed in a modular manner, consisting of a 3D OCT generation module and a multimodal disease recognition module with uncertainty-aware fusion.

Regarding synthetic OCT generation, three representative generative paradigms—DCGAN, VAE-GAN, and conditional diffusion models (DDPM)—were investigated. Experimental results demonstrate that diffusion-based OCT generation produces more stable and consistent volumetric structures, leading to improved downstream fusion performance on the RFMiD dataset. In contrast, DCGAN-based generation suffers from mode collapse and limited biological diversity, while VAE-GAN improves training stability but remains inferior to diffusion-based approaches in most evaluation metrics.

From the recognition perspective, experiments on the RFMiD dataset show that multimodal Fun-

dus+OCT fusion models can achieve strong performance under certain configurations, particularly when powerful backbones such as Res2Net-50 + ResNet3D or EfficientNet + UNETR are combined. However, on the ODIR-5K dataset, which is characterized by severe class imbalance and domain discrepancies, multimodal fusion does not yield consistent improvements over fundus-only models.

The comparison between fundus-only models and multimodal fusion models reveals that multimodal learning is not inherently superior when the auxiliary modality is synthetically generated. In some cases, synthetic OCT data fail to capture disease-specific microstructural biomarkers and may act as a source of noise, leading to marginal or even degraded performance. These findings show that the effectiveness of multimodal fusion strongly depends on the quality and reliability of each modality.

Overall, this work emphasizes that the benefit of multimodal learning is conditional rather than guaranteed. Data quality, generative fidelity, and fusion strategies must therefore be considered carefully. The integration of uncertainty-aware fusion mechanisms, such as EyeMoST, shows promise in mitigating the negative impact of unreliable modalities and improving robustness in real-world clinical screening scenarios.

Future work may focus on enhancing the fidelity of synthetic OCT generation, reducing domain mismatch across datasets, and leveraging self-supervised or transfer learning techniques to improve generalization. In addition, incorporating uncertainty estimation into clinical decision-making pipelines remains an important direction for developing trustworthy ophthalmic AI systems.

## ACKNOWLEDGEMENT

This research was supported by The Vietnam National University Ho Chi Minh City - University of Information Technology's Scientific Research Support Fund.

## AUTHOR CONTRIBUTIONS

Conceptualization, Dinh Thuan Nguyen, Minh Nhut Nguyen, Xuan Thuy Hoang, and Thanh Thuy Nguyen Thi methodology, Dinh Thuan Nguyen, Minh Nhut Nguyen, and Xuan Thuy Hoang software, Xuan Thuy Hoang and Thanh Thuy Nguyen Thi; validation, Xuan Thuy Hoang, Thanh Thuy Nguyen Thi, and Minh Nhut Nguyen formal analysis, Xuan Thuy Hoang and Thanh Thuy Nguyen Thi investigation, Xuan Thuy Hoang and Thanh Thuy Nguyen Thi data curation, Xuan Thuy Hoang and Thanh Thuy Nguyen Thi writing—original draft preparation, Xuan Thuy Hoang and Thanh Thuy Nguyen Thi writing—review and editing, Dinh Thuan Nguyen and Minh Nhut Nguyen visualization, Xuan Thuy

Hoang and Thanh Thuy Nguyen Thi supervision, Dinh Thuan Nguyen and Minh Nhut Nguyen project administration and funding acquisition, Dinh Thuan Nguyen. All authors have read and agreed to the published version of the manuscript.

## References

- [1] World Health Organization, "World report on vision," WHO, Geneva, Switzerland, 2019.
- [2] M. Gandor *et al.*, "Diagnostics of diabetic retinopathy based on fundus photos using machine learning methods with advanced feature engineering algorithms," *Scientific Reports*, vol. 15, p. 34486, 2025.
- [3] M. Elkholy and M. A. Marzouk, "Deep learning-based classification of eye diseases using convolutional neural network for OCT images," *Frontiers in Computer Science*, vol. 5, p. 1252295, 2024.
- [4] Y. Li *et al.*, "Multimodal information fusion for glaucoma and diabetic retinopathy classification," in *Ophthalmic Medical Image Analysis*, pp. 53–62, 2022.
- [5] E. Sükei *et al.*, "Multi-modal representation learning in retinal imaging using self-supervised learning for enhanced clinical predictions," *Scientific Reports*, vol. 14, p. 26802, 2024.
- [6] M. A. Rodríguez, H. AlMarzouqi and P. Liatsis, "Multi-Label Retinal Disease Classification Using Transformers," in *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 6, pp. 2739–2750, June 2023.
- [7] T. K. Yoo *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," *Medical & Biological engineering & Computing*, vol. 57, no. 3, pp. 677–687, 2019.
- [8] K. Zou *et al.*, "Reliable multimodality eye disease screening via mixture of Student's t distributions," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 596–606, 2023.
- [9] S. Pachade *et al.*, "Retinal fundus multi-disease image dataset (RFMiD): A dataset for multi-disease detection research," *Data*, vol. 6, no. 2, p. 14, 2021.
- [10] N. Li *et al.*, "A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection (ODIR-5K)," in *Proc. Bench*, pp. 177–193, 2021.
- [11] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.
- [12] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural net-

- works,” in *Proceedings of the 36th International Conference on Machine Learning*, pp. 6105–6114, 2019.
- [13] A. Dosovitskiy *et al.*, “An image is worth 16×16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv: 2010.11929*, 2021.
- [14] A. Hatamizadeh *et al.*, “UNETR: Transformers for 3D Medical Image Segmentation,” *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, pp. 1748–1758, 2022.
- [15] B. Liu *et al.*, “APTOS-2024 challenge report: Generation of synthetic 3D OCT images from fundus photographs,” *arXiv preprint arXiv: 2506.07542*, 2025.
- [16] Z. Wang *et al.*, “Synthetic artificial intelligence using generative adversarial network for retinal imaging in detection of age-related macular degeneration,” *Frontiers in Medicine*, vol. 10, p. 1184892, 2023.
- [17] Y. Xue *et al.*, “Gen-3Diffusion: Realistic image-to-3D generation via 2D and 3D diffusion synergy,” *arXiv preprint arXiv: 2412.06698*, 2024.
- [18] U. Bhimavarapu, N. Chintalapudi and G. Battineni, “Automatic detection and classification of diabetic retinopathy using the improved pooling function in the convolution neural network,” *Diagnostics*, vol. 13, no. 15, p. 2606, 2023.
- [19] R. K. Rasel *et al.*, “Assessing the efficacy of 2D and 3D CNN algorithms in OCT-based glaucoma detection,” *Scientific Reports*, vol. 14, p. 11758, 2024.



**Thanh Thuy Nguyen Thi** was born in 2004. She received her Bachelor’s degree in Information Systems with Distinction from the University of Information Technology, Vietnam National University Ho Chi Minh City, in June 2026. She is currently working as a Business Analyst. Her professional and research interests include business analysis, requirements engineering, information systems, artificial intelligence, mul-

timodal deep learning, and medical image analysis. She is particularly interested in applying technology to develop reliable systems that support healthcare screening and clinical decision-making.



**Minh Nhut Nguyen** is currently a master’s student in Information Systems, at the Faculty of Information Systems, University of Information Technology, Vietnam National University Ho Chi Minh City, Vietnam.



**Xuan Thuy Hoang** received the B.S. degree in Information Systems from the University of Information Technology, Vietnam National University, Ho Chi Minh City, Vietnam, in 2026. He is currently working as a Software Developer. His research interests include machine learning, computer vision, image processing, generative models, and biomedical data analysis, with a particular focus on synthetic OCT generation and mul-

timodal retinal image analysis.



**Dinh Thuan Nguyen** has been the Associate Professor at the University of Information Technology (UIT), Vietnam National University Ho Chi Minh city (VNU-HCM). He has published more than 110 research papers in the area of database and data science at international/national level conferences and journals. He is currently guiding 4 PhD students and 7 Msc students in the area of data science.