Geometric Correspondence-Based Multimodal Learning for Ophthalmic Image Analysis

Yan Wang, Liangli Zhen, Tien-En Tan, Huazhu Fu, Yangqin Feng, Zizhou Wang, Xinxing Xu, Rick Siow Mong Goh, Yipin Ng, Claire Calhoun, Gavin SW Tan, Jennifer K Sun, Yong Liu, and Daniel SW Ting

***Abstract*— Color fundus photography (CFP) and Optical coherence tomography (OCT) images are two of the most widely used modalities in the clinical diagnosis and management of retinal diseases. Despite the widespread use of multimodal imaging in clinical practice, few methods for automated diagnosis of eye diseases utilize correlated and complementary information from multiple modalities effectively. This paper explores how to leverage the information from CFP and OCT images to improve the automated diagnosis of retinal diseases. We propose a novel multimodal learning method, named geometric correspondence-based multimodal learning network (GeCoM-Net), to achieve the fusion of CFP and OCT images. Specifically, inspired by clinical observations, we consider the geometric correspondence between the OCT slice and the CFP region to learn the correlated features of the two modalities for robust fusion. Furthermore, we design a new feature selection strategy to extract discriminative OCT representations by automatically selecting the important feature maps from OCT slices. Unlike the existing multimodal learning methods, GeCoM-Net is the first method that formulates the geometric relationships between the OCT slice and the corresponding region of the CFP image explicitly for CFP and OCT fusion. Experiments have been conducted on a large-scale private dataset and a publicly available dataset to evaluate the effectiveness of GeCoM-Net for diagnosing diabetic macular edema (DME), impaired visual acuity (VA) and glaucoma. The empirical results show that our method outperforms the current state-of-the-art multimodal learning methods by improving the AUROC score 0.4%, 1.9% and 2.9% for DME, VA and glaucoma detection, respectively.**

***Index Terms*— Multimodal learning, multimodal fusion, multimodal retinal imaging, ophthalmic image analysis**

I. Introduction

The retina contains millions of light-sensitive cells and other nerve cells that receive and organize visual information to enable the capability of visual perception. Retinal diseases such as diabetic retinopathy, diabetic macular edema (DME) and age-related macular degeneration are among the leading causes of severe vision loss and blindness worldwide [1]. Early detection of these retinal diseases allows for prompt treatment, which can often prevent or reverse visual loss. Color fundus photography (CFP) and optical coherence tomography (OCT) images are two of the most widely used modalities in the clinical management of retinal diseases. CFP imaging provides a two-dimensional (2D) image of the posterior aspect of the interior surface of the eye, including the posterior retina, retinal vasculature, optic disc, macula, and posterior pole [2], [3]. OCT is used to obtain non-invasive, high-resolution three-dimensional (3D) volume scans of the retina in vivo, which are often viewed by clinicians as 2D cross-sectional scan slices [4]. These two imaging modalities provide different, but complementary information on the retina in healthy and diseased states, which is used by clinicians to make the diagnosis of retinal disease. Based on such imaging data, many CFP-based and/or OCT-based methods have recently been proposed for automated retinal disease diagnosis [5], [6]. However, most of these methods are based on single modal input, i.e., either using CFP images or OCT slices as the input to diagnose the diseases. For example, using CFP images to classify DR [5] and Glaucoma [7], or using OCT scans to detect DR [6] and DME [8]. The critical component of such an automated retinal disease diagnosis system is the feature extraction module, which extracts discriminative features for image classification. Convolutional neural networks (CNNs) have achieved great success in natural image classification [9], [10]. Additionally, CNNs have been used as feature extractors to extract features from medical images [11]–[13], including CFP and OCT images [5], [8]. CFP imaging has shown promising performance in diagnosing eye diseases, and the captured images lie in a 2D space, which can be directly inputted into CNNs [5], [7], [14]–[16]. While OCT scans are with 3D volumes, they are hard to be handled by classical CNNs (2D CNNs) to extract features from 3D volumes directly. One way to solve this problem is to treat the 3D volume as a multi-instance sample. Taking one slice as one image, 2D CNN can be used to extract features of

This work was supported by the Agency for Science, Technology and Research (A*STAR) through its AME Programmatic Funding Scheme under Project A20H4g2141. (J. Sun, Y. Liu, and D. Ting contributed equally to this work.) (Corresponding author: Liangli Zhen.)

Yan Wang, Liangli Zhen, Huazhu Fu, Yangqin Feng, Zizhou Wang, Xinxing Xu, Rick Siow Mong Goh, Yipin Ng, and Yong Liu are with the Institute of High-Performance Computing, Agency for Science, Technology and Research, Singapore 138632 (e-mail: {wangyan, zhenll}@ihpc.a-star.edu.sg, hzfu@ieee.org, {fengyq, wang_zizhou, xuxinx, gohsm, ng_yi_pin, liuyong}@ihpc.a-star.edu.sg).

Tien-En Tan, Gavin SW Tan and Daniel SW Ting are with the Singapore Eye Research Institute, Singapore 169856, Singapore National Eye Centre, Singapore 168751 (e-mail: tantienen@gmail.com; gmstansw@nus.edu.sg; daniel.ting45@gmail.com).

Claire Calhoun is with the Jaeb Center for Health Research, Tampa, FL 33647, USA (email: ccalhoun@jaeb.org).

Jennifer K Sun is with the Joslin Diabetes Center, Beetham Eye Institute, Harvard Department of Ophthalmology, Boston, MA 02215, USA (e-mail: jennifer.sun@joslin.harvard.edu).

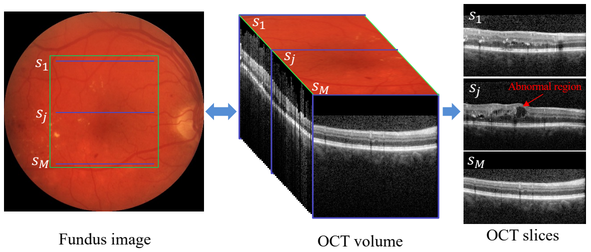


Fig. 1. Illustration of the geometric relationship between CFP image and OCT slices in a DME-positive example. The green box is the general region of the OCT scan. Three blue lines, i.e., s_1, s_j, and s_M, on the CFP image demonstrate the locations of three slices corresponding to OCT volume. The right column shows three specific OCT slices, which are chosen from the OCT volume. This example demonstrates that different slices have different appearances for a DME-positive patient, i.e., the slice of s_j has an obvious abnormal region, but s_M looks normal.

each OCT slice and fuse all slices' feature maps [17]. Another way is using 3D CNN as a feature extractor [18] where one dimension aims to learn the relationships between adjacent slices. Recently, some studies propose to use multimodal inputs to improve performance by utilizing the complementary information of multiple input modalities [19]–[22].

In multimodal learning, there are three multimodal fusion strategies, i.e., early fusion, intermediate fusion, and late fusion, to fuse representations for different modalities. Early fusion combines different modalities' original input data [23] or feature representations [24] before making a prediction, and late fusion fuses each modality's predictions to obtain the final predictions. Intermediate fusion is similar to early fusion, but the model of intermediate fusion strategy has an interaction among modalities during the training process, and the model attempts to discover within-modality correlations [23]. Several recent works of literature have combined multi-modality inputs to classify eye diseases from CFP and OCT images [25]–[28]. However, these works do not take into account the geometric correspondence between the CFP and OCT modalities to learn the correlated features of the two modalities. In contrast, when retinal imaging is evaluated by human clinicians, the anatomic correlation between multimodal imaging inputs is often crucial to making accurate diagnoses. Figure 1 illustrates the geometric correlation between a CFP image and OCT scan, in a sample case with DME. There exists a geometric correspondence; that is, each slice in the OCT volume can find a potential corresponding location on the enface plane of the fundus image [29]. This geometric information may help to localize and confirm potential lesions for a specific disease diagnosis. For example, if a lesion exists on slice s_1, the corresponding rows of the CFP image may have different characteristics from other regions. Besides, the different slices of OCT volume have different appearances, as shown in the right column of Fig 1. Specifically, slice s_j has intraretinal cysts and retinal thickening from DME, slice s_1 has hard exudates (hyper-reflective lesions) related to DME, and s_M does not have any obvious lesions related to DME. Thus, keeping the key features and removing the redundant information for the OCT volume is essential for the feature extraction of OCT modality.

Inspired by the importance of the geometric relationship between a CFP and corresponding OCT slices in clinical practice, this paper proposes a novel multimodal method named geometric correspondence-based multimodal learning network (GeCoM-Net) to perform ophthalmic image analysis. We utilize this correspondence in the feature space for representation learning of two modalities. Specifically, based on the clinical observations, we consider the geometric correspondence between the OCT slice and the CFP image region to learn the correlated features of the two modalities for robust fusion. The correlated information is learned by a new proposed module (named geometric correspondence-based attention, GeCoA). It can be utilized to boost confidence in the learned features for both modalities by encouraging the consistency of the feature vectors from the two modalities. Furthermore, we design a new feature selection strategy to extract discriminative OCT representations by automatically selecting the important feature maps from OCT slices. Our proposed new feature selection (named multi-instance feature selection, MIFS) can extract more discriminative features of 3D OCT volume and reduce redundant information. MIFS is based on the activation of a feature map that represents a certain pattern of the input image. By selecting the activated feature maps from all slices, our method can reduce the redundant feature representations while maintaining the discriminative of the final OCT feature representations. MIFS can also benefit the feature extraction of CFP modality by involving an attention mechanism to the CPF image based on the geometric correspondence information. To be specific, we first compute the responses of OCT slices and then map the responses into the rows of CFP image feature maps. Thus, we can know which area of the CFP image is more likely to be a lesion area.

We summarize the novelty and main contributions of this work as follows:

1) To improve automated ophthalmic image analysis, we propose a novel multimodal geometric correspondence-based multimodal learning network that leverages the geometric relationships between the CFP image and its corresponding OCT slices. To our best knowledge, this work is the first to explicitly consider modeling the geometric information between the CFP and OCT modalities to enhance multimodal eye disease diagnosis.
2) A new feature selection module, MIFS, is designed to select the activated feature maps from OCT slices. It can reduce redundant information and simultaneously preserve discriminative information for the feature extraction of the OCT modality. The MIFS module can also benefit the feature extraction of the CFP modality by working together with the geometry correspondence learning module.
3) A geometry-corresponding attention module, GeCoA, is designed to learn the geometry relationship between CFP image and OCT slices. It utilizes the OCT slice information to guide the attention of CFP feature learning.

The remainder of this paper is organized as follows: We review related studies in Section II. In Section III, we present the details of our proposed method. In Section IV, we report the experimental setup and results. Finally, we conclude this paper in Section V.

II. Related work

The section reviews both the single-modal-based and multimodal-based methods for ophthalmic image analysis. Also, we will highlight how our method differs from the existing ones.

A. Single-Modal-Based Approach for Ophthalmic Image Analysis

CFP and OCT images are among the most commonly used modalities to diagnose retinal and other eye diseases [30]. CFP images are 2D images, and CNNs are primarily used for CFP image analysis. Among the existing studies, Juan *et al.* exploited the application of different CNN models to classify Glaucoma and used transfer learning to improve the performance [31]. Hu *et al.* proposed a neural network that contains two sub-networks. One sub-network is designed to extract discriminative features, and another is designed to predict retinopathy of prematurity [32]. Shankar *et al.* proposed a framework for preprocessing, segmenting, and classifying diabetic retinopathy from CFP images [33]. In the framework, the authors proposed a synergic deep learning model to classify the DR CFP images to various severity levels. Hervella *et al.* proposed a method to do the classification and segmentation tasks simultaneously [14]. In this method, the learning of two tasks is optimized simultaneously to enhance the input image's feature learning. Furthermore, a multi-adaptive optimization strategy is employed to ensure the equal contribution of two tasks to classify glaucoma and segment optic disc and cup from CFP images.

The 3D OCT volume contains multiple 2D slices, enabling ophthalmologists to cross-sectionally examine distinct retinal layers for more accurate diagnosis and evaluation. The methods for OCT image analysis can be grouped into two main categories. The first group of strategies is to treat the classification of samples with multiple slices as a multi-instance problem. This category of methods usually employs a 2D-CNN for feature extraction and then applies a feature fusion strategy to fuse the representations from multiple slices. For instance, EVT-MIL proposes an iterative sampling framework to classify 3D OCT volumes as a multi-instance problem [34]. It utilizes an iterative algorithm to infer slice labels that increase the algorithm's complexity. Li *et al.* proposed a multi-instance multi-scale (MIMS)-CNN to fuse the representations from multiple slices with multi-scale operation [17]. In MIMS-CNN, a top-k pooling strategy is presented to select the most active representations from different scale feature vectors and fuse the selected representations as the final feature vector. Wang *et al.* proposed an uncertainty-driven multi-instance scheme that uses a recurrent neural network to generate the bag-level representations for the final classification [35]. Alternatively, another group of methods is to classify 3D OCT volumes using 3D-CNNs by treating the dimension associated with indices of the slides as the third dimension. For example, Thakoor *et al.* proposed Hybrid 3D-2D-CNN to classify 3D OCT images [36]. It contains two 3D-CNNs that are designed to extract representations from 3D OCTA and OCT structural inputs. Then, the 3D representations from the two 3D-CNNs are concatenated with the 2D-CNN representations extracted from 2D B-scan images to detect AMD disease. George *et al.* proposed an attention-guided 3D-CNN framework that combines with Grad-CAM to locate the possible lesions to classify Glaucoma from 3D OCT volumes [37]. Zhang *et al.* proposed LamNet, a lesion attention maps-guided model, to predict the Choroidal Neovascularization from SD-OCT images by leveraging a multi-scale 3D CNN [38]. Unlike 2D-CNNs, the third dimension of 3D-CNN is used to learn the correlation information between several adjacent slices [18]. It aims to discover the potential features among slices, but it will contain more trainable parameters, which may be more challenging to train [39].

All methods mentioned above are single-modal-based methods. In contrast, our method is a multimodal-based method that utilizes complementary information and correlated information from both CFP and OCT images simultaneously.

B. Multimodal Learning Approach for Ophthalmic Image Analysis

CFP image and OCT volume are typically 2D and 3D data. Zhao *et al.* give a comprehensive review of deep learning-based 2D and 3D fusion methods [40]. Although several applications utilize the 2D and 3D data to improve performance, such as segmentation [41], [42] and detection [43], [44], few methods have considered the geometric correspondence of the two modalities among these applications. For ophthalmic image registration, the key step is to find the strict correlation between CFP image and OCT volume. Miri *et al.* proposed a feature-based method that used the histograms of the oriented gradients to improve the registration performance without additional blood vessel segmentation [45]. Mokhtari *et al.* studied the symmetry between two eyes based on the fusion feature of CFP images and OCT volumes. They used the vessel information of both modalities to register the CFP images and OCT volumes [46].

For utilizing the information from multiple modalities in the ophthalmic image classification task, the most straightforward idea is concatenating the representations from both modalities. Based on this idea, Yoo *et al.* first attempted to combine the CFP and OCT representations from both modalities as the final representations. Then they used the final representations to detect AMD by training a random forest classifier [25]. This method needs to concatenate the pair of CFP and OCT representations from each sample. Jin *et al.* proposed a feature-level fusion (FLF) method to fuse the features to diagnose typical neovascular AMD [47]. Differently, FLF used OCT and OCTA images as the input data. Different from this method, Wang *et al.* proposed a two-stream CNN, which uses a loose pairing strategy to classify AMD [20], [48]. This strategy concatenates the CFP representations with OCT representations from the samples with the same label instead of strictly from one sample. To learn more discriminative features

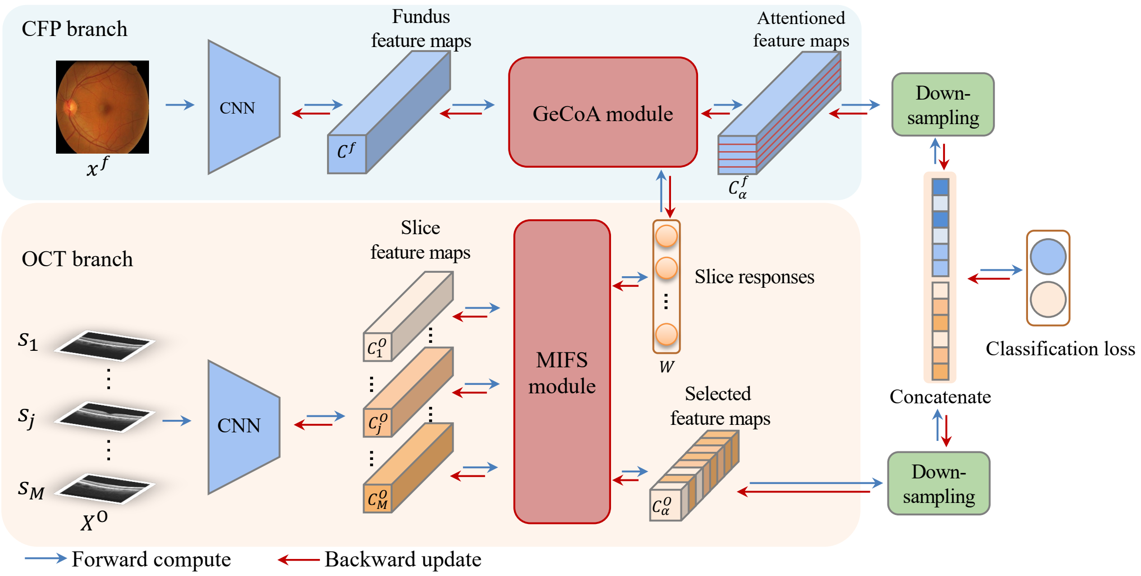


Fig. 2. The architecture of GeCoM-Net. The upper branch is the CFP branch, which extracts the feature vector from the i-th input CFP image, and the bottom branch is the OCT branch, which extracts the feature vector from the i-th input OCT volume. There is a multi-instance feature selection (MIFS) module in the OCT branch to compute the response of each slice and select feature maps. The geometric correspondence-based attention (GeCoA) module is designed to model the geometric information between the CFP image and OCT slices.

of each modality, He *et al.* proposed a modality-specific attention network (MSAN) to classify multiple diseases [49]. The main components of MSAN are two attention subnets, i.e., a multi-scale attention subnet for extracting multi-scale features from CFP images and a region-guided attention subnet for extracting features from OCT images using OCT regions-of-interest (ROIs).

As the experimental results show, the above methods have achieved better performance than those using a single modality. However, it is important to note that these methods only utilized one well-labeled OCT slice for each sample, whereas each OCT scan is a 3D volume scan that contains many more slices, and much more information. Selecting and labeling one slice from each OCT volume is exceptionally laborious, but more importantly is prone to sampling error, with the classification results being highly dependent on the OCT slice selection strategy [19]. To address this issue, Li *et al.* proposed a multimodal multi-instance learning (MM-MIL) model to detect retinal diseases [19]. The main idea of MM-MIL is selectively fusing CFP and OCT modalities. It divides the CFP image into a certain number of patches to match the number of OCT slices. It fuses the representations of multiple patches and slices using a multimodal multi-instance subnet and a mean pooling layer. Finally, the fused representations are used to detect the seven diseases.

As mentioned above, most of the existing multimodal-based methods, including our method, are based on the early fusion strategy. The key difference between the existing multimodal methods and our method is we introduce a new geometric correspondence-based multimodal learning strategy to utilize the geometric relationship information of the two modalities to enhance representation learning, which can improve classification accuracy significantly. Besides, a new multi-instance feature selection module is designed to extract discriminative features for 3D OCT images.

III. The Proposed Method

This section introduces the details of our proposed method - GeCoM-Net. The inputs of GeCoM-Net are the CFP image and corresponding OCT slices, and its output is the probability of a particular disease being present. Let $\mathcal{D} = \{(x_i^f, X_i^O, y_i)\}_{i=1}^N$ be the set of samples in the multimodal dataset, where x_i^f denotes a CFP image, $X_i^O = \{s_{i1}, \ldots, s_{iM}\}$ denotes a bag of OCT slices that include M slices in total, y_i is the corresponding label of the i-th sample, and N is the total number of samples in the dataset. The goal of our proposed method is to train a neural network model as a mapping function $p = \xi\left(x^f, X^O, \boldsymbol{\theta}\right)$ to map the input sample from the multimodal image space to its label space, where $\boldsymbol{\theta}$ represents the trainable parameters of the neural network model and p is the output probability of suffering a certain disease for the input sample. The overall architecture of our method is shown in Fig. 2, from which we can see that our proposed method contains two branches, one for CFP image feature extraction and another for OCT image feature extraction. The feature extraction procedures of both modalities are based on 2D CNNs. The CFP image will be forwarded to a CNN in the CFP branch to obtain the feature maps. In the OCT branch, since the input is a bag of OCT slices, we will get M sets of feature maps of one OCT volume input after using a 2D CNN for feature extraction. Then, the M sets feature maps will be input into our MIFS component for feature selection and computing the response of the slice. After calculating the responses of OCT slices, a geometric correspondence-based attention module is employed to transform the slice responses into the row-spatial importance of CFP feature maps based on the geometric correspondence. An attention mechanism computes the enhanced CFP feature maps according to the row-spatial importance. Then, the selected OCT feature maps and enhanced CFP feature maps will go through a pooling

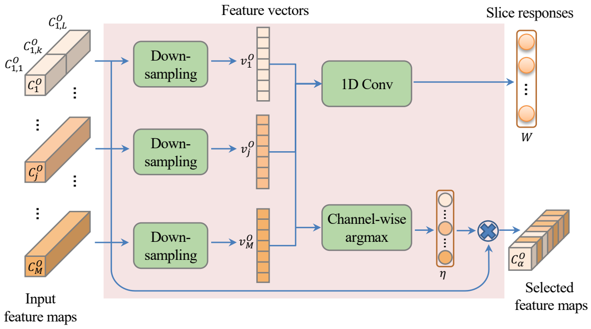


Fig. 3. The detailed architecture of our proposed MIFS module. The inputs of this module are the M sets of feature maps from the OCT volume, i.e., $\boldsymbol{C^O_1}, \ldots, \boldsymbol{C^O_j}, \ldots, \boldsymbol{C^O_M}$. The outputs contain the weights $\boldsymbol{W}$ for the slices and the feature vector $\boldsymbol{v^O_\alpha}$ of OCT volume. The multiplication mark stands for the feature channel selection according to the channel index $\boldsymbol{\eta}$.

layer to transform the feature maps into feature vectors. The feature vectors of the two modalities will be concatenated as the final feature vector. Finally, a classifier is connected to classify concatenated feature vectors. Generally, our method contains four main components: 1) multimodal input feature extraction, 2) multi-instance feature selection for the OCT branch, 3) geometric correspondence-based attention module, and 4) feature fusion and classification.

A. Multimodal Feature Extraction

Since the CFP image and OCT slices have distinctly different characteristics, for example, the CFP image is a color RGB image, but OCT slices are grey-scale images. Thus, we utilize two independent 2D CNN backbones but the same architecture, i.e., ϕ and ψ, to extract features for CFP images and OCT slices, respectively. For each input CFP image x^f and OCT slice s_j, the feature extraction procedure can be computed as:

$$\begin{aligned} C^f &= \phi(x^f), \\ C^O_j &= \psi(s_j), \end{aligned} \tag{1}$$

where j denotes the j-th slice of the input OCT volume. In our method, the CFP image and OCT slice are in different resolutions, so the sizes of the feature map for them are different sizes.

B. Multi-Instance Feature Selection

Multi-instance feature selection is designed to discover more critical ones from a bag of input OCT slices for each channel. The detailed architecture of the MIFS component is shown in Fig. 3. The input of MIFS is a set of OCT feature maps $C^O = \{C^O_1, \ldots, C^O_M\}$, which are extracted by the OCT CNN backbone ψ. Firstly, the feature maps of each slice should go through a down-sampling module to convert the feature maps into a feature vector using a pooling operation as:

$$v^O_j = \text{Pooling}(C^O_j), \tag{2}$$

where $j = 1, \ldots, M$ is the slice index. Each feature map is transformed into an element in the feature vector v^O_j. Here, we utilize a pooling layer to compute the feature vectors. Then, we use the feature vectors to select the feature maps from the OCT volume. We regard each element $v^O_{j,k}$ of the feature vector v^O_j as the degree of activation of a certain pattern. For all M slices, we select the maximum value from the feature vectors on the k-th element of the feature vectors as the selected feature map index for the k-th channel. Equation (3) shows the detailed computation of the index.

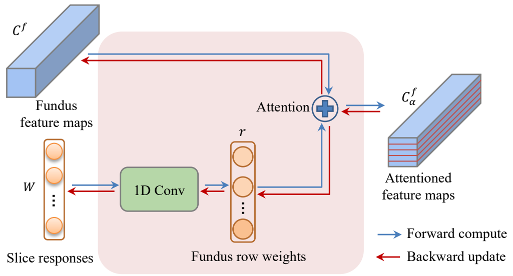


Fig. 4. The architecture of geometric correspondence-based attention module. The inputs of this module are the feature maps $\boldsymbol{C^f}$ of the CFP image and the slice responses $\boldsymbol{W}$ which are from the MIFS module. The output is the feature vector $\boldsymbol{v^f_\alpha}$ of the CFP image. The addition mark stands for row attention according to the fundus row weights $\boldsymbol{r}$.

$$\eta_k = \text{argmax}_j([v^O_{1,k}, v^O_{2,k}, \ldots, v^O_{j,k}, \ldots, v^O_{M,k}]). \tag{3}$$

After obtaining the indexes, we select the feature maps from all slices to obtain the final selected feature maps as:

$$C^O_{\alpha,k} = C^O_{\eta_k,k}. \tag{4}$$

Finally, we use a down-sampling module, i.e., a global average pooling layer, to calculate the feature vector of the input OCT volume:

$$v^O_\alpha = \text{GAP}(C^O_\alpha), \tag{5}$$

where $\text{GAP}(\cdot)$ denotes the global average pooling operation.

To ensure that the feature selection process can extract sufficient information, we incorporate both global average pooling and global max pooling in the pooling layer, as described in Eq. (2). Afterward, we concatenate the feature vectors, which are calculated using Eqs. (3)−(5) under two different pooling operations, to form the final feature vector for the OCT branch.

At the same time, we compute the response of each slice based on the obtained feature vectors v^O_j, where $j = 1, \ldots, M$. Since there may be more than one slice containing lesion information, we need to compute the probability of each slice independently. In this work, we leverage a 1D convolutional layer to calculate it for each feature vector. Lastly, the response of each OCT slice is computed as:

$$w_j = \sigma(\text{Conv1D}(v^O_j))), \tag{6}$$

where $\sigma(\cdot)$ is a sigmoid activate function and $\text{Conv1D}(\cdot)$ denotes the 1D convolutional layer with the padding size be 0 and the kernel size is the length of the input feature vector. The output of the sigmoid function indicates the probability of the slice having a potential lesion area. These weight values of the slices will be used to guide the correlated feature learning illustrated in the following.

C. Geometric Correspondence-Based Attention

This module aims to learn the correlated features between the two modalities. Specifically, we leverage the slice responses from OCT slices to guide the model's attention to some rows of CFP feature maps and calculate the final CFP feature maps using the attention mechanism. The end-to-end learning process will enable the model to learn the correlated features for the final disease diagnosis. The specific architecture of this component is shown in Fig. 4. The outputs of MIFS are the response values $W = [w_1, \ldots, w_M]$ for the input slices. This component first leverages a 1D convolutional layer to transform the weights into the size of the row number of the CFP feature map. Using the 1D convolutional layer considers the several adjacent slices that may contribute to one row of the CFP image feature maps. The computational equation is:

$$r = \sigma(\text{Conv1D}(W)), \tag{7}$$

where Conv1D$(\cdot)$ is a 1D convolutional layer whose padding size is 2, kernel size is 6, and stride is 4, and $r \in \mathbb{R}^{h \times 1}$ indicates the row weights of the rows for each CFP feature map.

Then, the enhanced CFP feature maps are obtained by applying the row weights to the CFP feature maps, which are calculated as:

$$C^f_\alpha = C^f + r \otimes C^f, \tag{8}$$

where $\otimes$ denotes the multiplication between each row of feature maps and the weights.

Finally, a global average pooling layer is utilized to convert the enhanced feature maps into a feature vector as:

$$v^f_\alpha = \text{GAP}(C^f_\alpha). \tag{9}$$

D. Feature Fusion and Classification

After obtaining the feature vectors v^O_α and v^f_α for input OCT volumes X^O and CFP images x^f, GeCoM-Net utilizes a simple but effective way [50], i.e., concatenating the feature vectors from two modalities, to fuse the multimodal feature vectors. Then, it is followed by a dropout layer with a dropout rate is 0.6. Finally, a softmax classifier is employed to classify the feature vector into specific categories, and we denote the output possibility as a positive case for each input multimodal sample (x^f, X^O) as p. We use focal loss [51] as the loss function. The focal loss can be calculated by:

$$\mathcal{L} = -\sum_{i=1}^{N}(1 - p^t_i)^\gamma \log(p^t_i), \tag{10}$$

where

$$p^t_i = \begin{cases} p_i & \text{if } \ y_i = 1, \\ 1 - p_i & \text{otherwise.} \end{cases} \tag{11}$$

and i denotes the i-th sample, γ is the focusing parameter that controls the easily classified category's weight. When $\gamma = 0$, Equation (10) equals the normal cross-entropy loss function.

IV. Experimental Study

A. Dataset

In this study, we employ two multimodal ocular imaging datasets, focused on different ocular diseases. We used a DME dataset derived from the DRCR Retina Network clinical trials [52], as well as the Glaucoma grAding from Multi-Modality imAges (GAMMA) dataset from the GAMMA challenge [53] to evaluate the models' performance. Both datasets contain two imaging modalities, i.e., CFP images and OCT volumes.

The DME dataset is a large multicenter dataset including diabetic patients with and without DME, derived from 8 different clinical trial protocols of the DRCR Retina Network. These subjects had paired CFP and OCT images from the same time points, as well as accompanying clinical data such as demographic data and best-corrected visual acuity (VA) from subjective manifest refraction. CFP images are captured by Canon, Kowa, Optos, Topcon and Zeiss cameras. OCT volumes are captured by Heidelberg and Zeiss machines. After removing incorrect, corrupted, incorrectly formatted, duplicate, and incomplete data from the original dataset, we obtained 1007 samples (eye-level) from 820 patients. Each eye has one CFP image and one B-scan OCT volume containing 49 slices. The size of each CFP image is between 863×1100 and 4000×6000 pixels, and the size of each OCT slice is 496×512 pixels. Each instance from this dataset is annotated for the presence of center-involved-DME (referred to as "DME" in this paper), and impaired visual acuity (VA) based on previously published clinical trial protocol guidelines [52], [54]. The label for DME is a binary label to indicate the presence or absence of center-involved-DME, which is based on established sex- and machine-specific thresholds for retinal thickening, used throughout the DRCR Retina Network clinical trials [54], [55]. The label for impaired VA is based on a VA score that is determined by how many letters a patient is able to correctly read on a standard ETDRS vision chart. This VA score is an integer that ranges from 0 to 100. Most clinical guidelines only recommend treatment for DME, where there is center-involved-DME, together with impaired VA, which corresponds to a VA score of less than or equal to 78 letters (a "positive" case); otherwise, we label it as a negative case [55], [56]. Then, we split the entire dataset into three sub-sets: the training, validation, and testing subsets at the patient level to ensure that one patient's images only exist in one subset. In our experiment, we select the model with the highest validation AUROC score on the validation set for testing and report the results on the testing set. The detailed data statistics of the three subsets are 490, 164 and 166 patients in training, validation and testing sets, respectively. At the eye level, there are 601, 201 and 205 samples in training, validation and testing sets, respectively. The positive and negative sample ratios are around $1:1$ and $1:3$ for DME and impaired VA in the three subsets, respectively.

The GAMMA dataset is the first publicly available multimodal eye image dataset, containing 200 samples, 100 samples in training and 100 samples in testing sets. Each sample has one CFP image and one OCT volume of 256 slices. For the

CFP images, there are sizes of 2000×2992 and 1934×1956 pixels. The size of each OCT slice is 992×512 pixels. This dataset provides the label for grading three categories of glaucoma: no glaucoma, early glaucoma, and moderate or advanced glaucoma [53]. Since we can only access the label of the training set, we split the training data into the training and validation sets for model training with a ratio of $4:1$. Then, we upload prediction results on the test set into the official platform to obtain the final accuracy results.

B. Experimental Settings

Since the raw data of CFP images are saved in very high resolutions, we resize the CFP images into 410×410 pixels and the OCT slices are resized into 224×224 for both datasets. We use all 49 slices of each sample for the DME dataset and 64 slices of each sample with sampling every 4 slices for the GAMMA dataset to extract the feature of OCT volume by considering the GPU memory consumption. We also use several commonly used data augmentation ways to augment the input images as mentioned in reference [11], except for CFP images in which the random rotation angle is in the range of $\phi \in [-15°, 15°]$. We implement our method using the PyTorch framework. All the experiments are conducted on a DGX workstation using one NVIDIA A100 GPU card with 40 GB memory. During the training process, we use the Adam optimizer [58] with the learning rate $lr = 1e-4$, the batch size $B = 5$ and the maximum iteration epoch $E = 400$. We also adopt a multi-step learning rate scheduler to adjust the learning rate with milestone $[60, 160, 260]$, and the multiplicative factor of learning rate decay is 0.4. When training the DME classification model, the setting of γ in the focal loss is 0 since DME is a balance distribution. For VA classification, we set γ as 2. The hyperparameters are chosen through an iterative trial-and-error process, aiming to identify the configuration that yields optimal performance on the validation set. This selected configuration is then employed to assess the methods on the test data. By conducting a backbone selection experiment, we choose DenseNet-121 [10] as the backbones for both modalities. The results are shown in Table I.

For the DME dataset, to evaluate the performance of our model, we employ accuracy (ACC), the area under the ROC (receiver operating characteristic) curve (AUROC), precision, recall, specificity, and F1 score as the evaluation metrics. The larger the model's value for the five metrics indicates higher performance. Among these metrics, the primary metric is the AUROC score since it is not sensitive to data distribution. On the contrary, the scores of other metrics may be affected by the threshold. In this section, all reported results are based on the optimal threshold, which is selected using the precision-recall curve on the validation set. For the GAMMA dataset, we use the official metric, Cohen's Kappa coefficient, to evaluate the model's performance. These metrics are defined as follows:

$$\begin{cases} ACC = \dfrac{TP+TN}{TP+FP+TN+FN}, & Pre. = \dfrac{TP}{TP+FP}, \\ Rec. = \dfrac{TP}{TP+FN}, & Spe. = \dfrac{TN}{TN+FP}, \\ F1 = 2 \times \dfrac{Pre. \times Rec.}{Pre. + Rec.}, & Kappa = \dfrac{p_o - p_e}{1-p_e}, \end{cases} \tag{12}$$

where TP, FP, TN, and FN are the true positive, false positive, true negative, and false negative numbers, respectively. p_o and p_e denote the accuracy and the probability of predicting the correct categories by chance, respectively. Moreover, we utilize GPU memory requirements (Memo) and execution time (Time) to demonstrate the resource and time demands of various methods [59].

C. Comparison with state-of-the-art methods on the DME dataset

To demonstrate the effectiveness of GeCoM-Net, we first compare it with seven peer methods on the DME dataset, including five single-modal-based and four multimodal-based methods. For single modal-based methods, the baselines are 1) three different backbones, i.e., DenseNet-121 [10], ResNet-50 [9] and VGG-16 [57], for classifying CFP images by considering it is widely used for medical image analysis and has achieved promising performance; 2) 3D-CNN for classifying OCT volume images as mentioned in reference [19]; 3) MIMS-CNN [17], which is the current state-of-the-art (SOTA) for DME detection using OCT images. The four multimodal-based methods are 1) high-level feature concatenation (FCon) refers to the process of concatenating the feature vectors from different modalities as the final feature vector of one sample for the classification task. This is the most commonly used multimodal feature fusion strategy [50]; 2) Late fusion of predictions from multiple classifiers (LateFusion), which is also a commonly used strategy to fuse multimodal results at the decision level, using the procedure in reference [20] – averaging the output of their softmax layers; 3) MM-CNN [20], which is proposed to train a two-stream CNN (ResNet-18 as the backbone) that can extract features from both CFP and OCT images at the same time.; and 4) MM-MIL [19], which is a recently published multimodal method for retinal disease recognition based on ResNet-50. We reimplement these baselines according to the original papers but fine-tuned the hyper-parameters on the DME dataset.

The experimental results for all the methods are reported in Table I, from which we have the following observations: 1) The methods using OCT inputs outperform the methods that use CFP images, especially for DME detection (the AUROC score has been improved by more than 13%). The reason for this is that in clinical practice, OCT imaging provides more diagnostically relevant information than CFP for DME diagnosis, and the presence or absence of DME (including the labels in this dataset) is defined as OCT-measured retinal thickness. Therefore, it is unsurprising that single-modality OCT outperforms CFP imaging in this regard. 2) Multimodal learning methods can achieve higher accuracy than single-modal-based methods in most scenarios. It indicates that the two modalities can provide complementary information to each other for the detection of impaired VA and DME. However, FCon and LateFusion obtain lower AUROC scores for impaired VA diagnosis than single-modal-based methods that use OCT images. One potential reason is that these two methods have not fully captured mutually beneficial information from both modalities. This underscores the significance

TABLE I

COMPARISON OF THE RESULTS OBTAINED BY GECOM-NET AND ITS PEER METHODS. MEMO AND TIME STAND FOR GPU MEMORY CONSUMPTION (IN MB) AND INFERENCE TIME (IN MILLISECONDS) OF ONE SAMPLE, RESPECTIVELY. THE BOLD NUMBER AND UNDERLINED NUMBER INDICATE THE HIGHEST SCORE AND THE SECOND-HIGHEST SCORE IN A COLUMN, RESPECTIVELY.

Modality	Methods	VA						DME						Memo	Time
		Pre.	Rec.	Spe.	F1	ACC	AUROC	Pre.	Rec.	Spe.	F1	ACC	AUROC		
Fundus	Densenet-121 [10]	0.562	0.719	0.784	0.631	0.766	0.780	0.706	0.923	0.693	0.800	0.795	0.832	1150	13.9
	ResNet-50 [9]	0.403	**0.877**	0.500	0.552	0.605	0.762	0.695	0.802	0.719	0.745	0.756	0.809	1176	5.6
	VGG-16 [57]	0.387	0.719	0.561	0.503	0.605	0.700	0.533	0.890	0.377	0.667	0.605	0.692	**1030**	**2.5**
OCT	MIMS-CNN [17]	0.709	0.684	0.892	0.696	0.834	0.832	0.989	0.956	0.991	0.972	0.976	0.987	23424	158.3
	3D CNN [19]	0.581	0.754	0.791	0.656	0.780	0.829	0.902	0.912	0.921	0.907	0.917	0.967	2750	12.0
	Ours (MIFS)	0.655	0.667	0.865	0.661	0.810	0.832	0.966	0.923	0.974	0.944	0.951	0.992	6880	25.8
Fundus + OCT	FCon [50]	0.661	0.684	0.865	0.672	0.815	0.816	0.989	0.945	**0.991**	0.966	0.971	0.993	13870	41.0
	LateFusion [20]	0.655	0.632	0.872	0.643	0.805	0.817	0.926	0.824	0.947	0.872	0.893	0.961	7480	36.3
	MM-CNN [20]	**0.829**	0.596	**0.953**	0.694	**0.854**	0.847	0.978	0.956	0.982	0.967	0.971	0.992	2390	9.2
	MM-MIL [19]	0.620	0.772	0.818	0.688	0.805	0.851	0.830	0.857	0.860	0.843	0.859	0.902	8978	20.9
	Ours (GeCoM-Net)	0.732	0.719	0.899	**0.726**	0.849	**0.870**	**0.989**	**0.967**	**0.991**	**0.978**	**0.980**	**0.997**	7174	38.6

of proficient multimodal learning; otherwise, incorporating additional input modalities might not enhance the accuracy of final predictions. 3) Our proposed MIFS outperforms other single-modality methods, with the exception of MIMS-CNN. Comparing MIFS and MIMS-CNN, they yield comparable results, yet MIMS-CNN demands six times the computational time of MIFS. This demonstrates that our proposed feature selection module MIFS can extract discriminative features from input OCT volume efficiently. 4) GeCoM-Net achieves the highest or second-highest scores in most metrics for both impaired VA and DME diagnoses, especially the AUROC scores on two tasks. Specifically, the probability of predicting the correct categories by GeCoM-Net improves the AUROC score of the SOTA for impaired VA and DME diagnoses from 85.1% to 87.0% and 99.3% to 99.7%, respectively. It verifies the effectiveness of our proposed geometric correspondence-based multimodal learning strategy.

During inference, the majority of methods can process a single sample within 50 milliseconds. Our MIFS and GeCoM-Net both cost less than 40 milliseconds. In contrast, the MIMS-CNN necessitates over 158 milliseconds due to its requirement to compute features at three different scales. Comparatively, CFP-based techniques exhibit significantly lower inference times when pitted against OCT-based or multimodal-based methods. The latter two necessitate simultaneous computation of multiple OCT slices, contributing to increased processing time. Additionally, FCon, LateFusion, and GeCoM-Net are based on our MIFS for feature extraction, which results in them spending a similar amount of time for inference of one sample. The latency exhibited by our proposed method proves suitable across numerous real-world scenarios. Considering GPU memory consumption, as depicted in Table I, our MIFS and GeCoM-Net demonstrate comparable performance to OCT-based and multimodal-based methods. Notably, MIMS-CNN commands the highest GPU memory usage per sample due to its three-scale feature extraction. In contrast, MM-CNN exhibits modest GPU memory requirements by leveraging ResNet-18 as its backbone architecture. Furthermore, the FCon method consumes more GPU memory than other multimodal approaches. This discrepancy can be attributed to an additional fully connected layer with 128 neurons introduced before the classifier during the feature concatenation stage. While this layer streamlines training convergence, it concurrently escalates the parameter count of the fully connected layer.

TABLE II

COMPARISON OF THE RESULTS OBTAINED BY GECOM-NET AND THE PEER METHODS ON THE GAMMA DATASET. "ADD. INFO." AND "ENSEMBLE" IN THE THIRD AND FOURTH COLUMNS STAND FOR WHETHER USING ADDITIONAL DISC REGION ANNOTATIONS AND ENSEMBLE MULTIPLE MODELS OR NOT. ✗AND ✓DENOTE USE AND NOT USE THE OPERATE, RESPECTIVELY.

Modality	Method	Add. info.	Ensemble	Kappa
Fundus	Single-modality [53]	✗	✗	0.673
		✓	✗	0.677
OCT	Single-modality [53]	✗	✗	0.575
		✓	✗	0.732
Fundus + OCT	Multi-modality [53]	✗	✗	0.702
		✓	✗	0.770
	SmartDSP*	✗	✓	0.855
	VoxelCloud*	✗	✓	0.850
	EyeStar*	✗	✓	0.848
	HZL*	✗	✓	0.840
	IBME*	✗	–	0.826
	MedIPBIT*	✗	✓	0.805
	WZMedTech*	✗	✓	0.795
	DIAGNOS-ETS*	✗	✓	0.754
	MedICAL*	✗	✓	0.729
	FATRI-AI*	✗	✓	0.696
	Ours (GeCoM-Net)	✗	✗	0.860
	Ours (GeCoM-Net)	✗	✓	**0.884**

* The results are from the official report of the GAMMA challenge [53].

D. Comparison with the SOTA methods on the GAMMA dataset

Then, to further demonstrate the effectiveness of our proposed method, we conduct our method on the publicly available GAMMA dataset. There are several multimodal models [53], [60], [61] and fundus-based models [62], [63] have been developed based on the GAMMA dataset. We compare the results of our method with the results of six official models and the top-10 models on the final stage of the GAMMA challenge [53]. Among six baselines, there are two models for a single CFP modality, two models for a single OCT modality and two models for both modalities. The difference between the two models for different modalities is that one model uses additional Disc region annotations to help the model learn the Glaucoma features. For the ten methods, the results are the ensemble results, except for IBME, which does not provide the ensemble information. The detailed ensemble strategies of existing methods can be found in [53]. For a fair comparison, we also provide the ensemble result of our method. We choose three models that attained the highest AUROC scores during separate training processes, each trained with a distinct random seed. Subsequently, we create an ensemble by averaging the predictions generated by these three models. Besides, the best single model result of our method is also provided. The detailed results are reported in Table II. From the results in Table II, one can see that our ensemble result is the highest score which improves the kappa score of the SOTA for grading glaucoma from 85.5% to 88.4%. Besides, it is worth noting that our model, without using additional Disc annotations and ensemble strategy, can also achieve a higher kappa score (86.0%) than the previous SOTA method. It further verifies the effectiveness of our proposed geometric correspondence-based multimodal learning strategy.

E. Ablation Study

1) Effectiveness of pooling strategy in MIFS: As mentioned in Section III-B, both global average pooling and global max pooling are used to select the critical feature maps from the entire OCT volume in our method, ensuring that the feature selection process captures sufficient information. To demonstrate the effectiveness of the feature selection strategy in our proposed MIFS, we compare MIFS with using only these two basic pooling strategies to select features in the MIFS module. Specifically, we compare the following strategies: 1) using global average pooling (AVG. pooling) to select feature maps and 2) using global max pooling (MAX. pooling) to select feature maps. The experimental results are shown in Table III. We can see that global max pooling outperforms global average pooling on both the VA and DME classification tasks. Our MIFS can further improve the results of the approach that uses global max pooling in terms of F1 score, ACC and AUROC. These results verify the effectiveness of our multi-instance feature selection strategy (MIFS).

2) Effectiveness of GeCoA in GeCoM-Net: To verify the effectiveness of our proposed geometric correspondence-based attention module, we construct another variant of GeCoM-Net (denoted as w/o GeCoA). The method w/o GeCoA ignores the geometric relationships between the two modalities (i.e., removing the operations on C_i^f to update the weight parameters of the CFP and OCT branches), but it adopts our proposed MIFS strategy for learning from OCT scans. The comparison results are reported in Table IV. From the results, we see that GeCoM-Net can outperform its variant w/o GeCoA by a large margin, especially for the VA diagnosis, which improves the AUROC score from 81.6% to 87.0%. It indicates the essential role of the geometric correspondence-based attention module for the feature learning of GeCoM-Net.

TABLE III

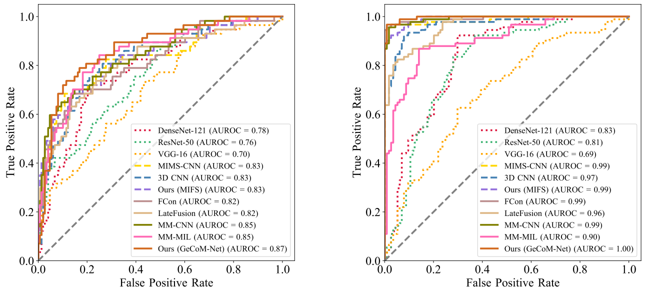
COMPARISON OF THE RESULTS FROM TWO POPULAR POOLING STRATEGIES AND MIFS. AVG. POOLING AND MAX. POOLING DENOTE THE GLOBAL AVERAGE POOLING AND THE GLOBAL MAX POOLING FOR ALL THE FEATURE VECTORS OF OCT SLICES, RESPECTIVELY.

Methods	VA			DME		
	F1	ACC	AUROC	F1	ACC	AUROC
AVG. pooling	0.586	0.766	0.810	0.924	0.937	0.989
MAX. pooling	0.605	0.771	0.828	0.939	0.946	0.990
Ours (MIFS)	**0.661**	**0.810**	**0.832**	**0.944**	**0.951**	**0.992**

TABLE IV

RESULTS OF GECOM-NET AND ITS VARIANT W/O GECOA FOR THE VA AND DME DIAGNOSES.

Method	VA			DME		
	F1	ACC	AUROC	F1	ACC	AUROC
GeCoM-Net w/o GeCoA	0.621	0.810	0.816	0.967	0.971	0.993
Ours (GeCoM-Net)	**0.726**	**0.849**	**0.870**	**0.978**	**0.980**	**0.997**



(a) VA ROC curves

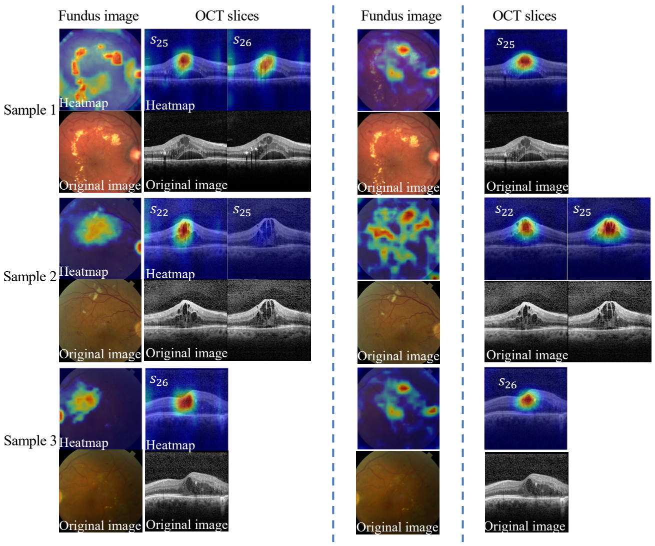
(b) DME ROC curves

Fig. 5. The ROC curves of the different models for VA and DME classifications.

F. Visualization and Discussion

For the classification of VA and DME on the DME dataset, VA classification is an imbalanced classification, and DME is a balanced classification. To demonstrate the performance of different models mentioned in Table I, we visualize the ROC curve, as shown in Fig. 5. The ROC curve is not biased toward the majority or minority class. From Fig. 5, we can see that our GeCoM-Net method consistently attains the highest area under the curve values across all three scenarios - the VA ROC curve and DME ROC curve. Notably, both the VA and DME ROC curves demonstrate the largest areas, underscoring GeCoM-Net's superior diagnostic performance for both VA and DME cases. In conclusion, our method has demonstrated robust performance in both imbalanced and balanced scenarios.

Furthermore, we utilize gradient-weighted class activation maps [64] (Grad-CAM) to show the focus areas of the model for the provided prediction in the input images. Grad-CAM is a technique for producing visual explanations for decisions from a large class of CNN-based models, thereby making them more transparent. We choose several positive samples from the test set of the DME classification task since positive samples have lesion areas that we can see whether the model focuses on the correct regions or not. We visualize the samples from our multimodal model, single CFP image-based model (i.e., DenseNet-121) and our OCT-based model (i.e., MIFS). For the multimodal model, We visualize both CFP and OCT images by computing the heat map separately. The OCT volume is normalized over all the slices to see which slices the model focuses on. For the OCT-based model, we use a similar way to visualize the OCT part of the multimodal model. The visualization results (ignoring the OCT slides whose heatmap pixel values are all less than 0.0002) are shown in Fig. 6. The two columns of Fig. 6 (a) are the CFP images and highlighted OCT slices. From the results, one can see that GeCoM-Net can



(a) Multimodal (b) CFP modal (c) OCT modal

Fig. 6. Demonstration of model's focused areas on three positive samples using Grad-CAM. We normalize the largest heatmap values across 49 slices into 0 to 1 for each sample and only show the slices whose largest heatmap values are larger than 0.0002. In the figure, the bright color denotes the region on which the model makes decisions.

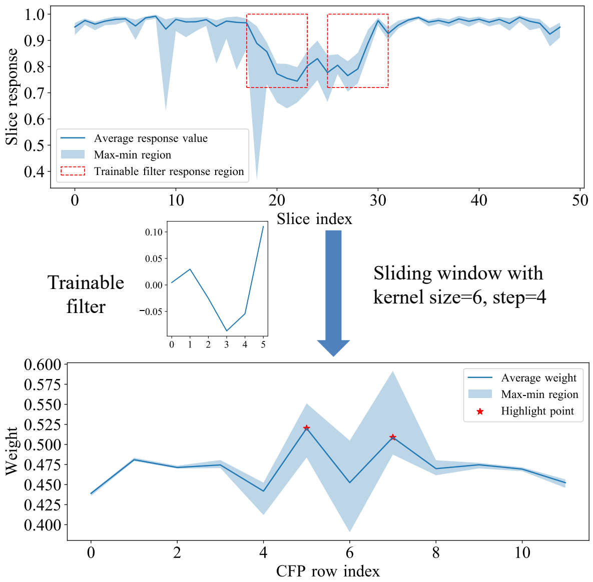


Fig. 7. The illustration of the OCT slice responses mapping to the CFP row weights for a DME positive example. The upper part is the slice response of OCT volume, the middle part is the well-trained filter and the bottom part is the CFP row weights.

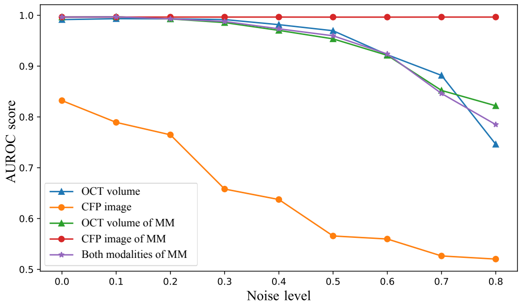


Fig. 8. The performance of different models under different levels of Gaussian noise. "OCT volume" stands for adding noise to the OCT volume of the single OCT-based model, "CFP image" stands for adding noise to the CFP image of the single CFP-based model, "OCT volume of MM" denotes only adding noise to the OCT volume of our multimodal model, "CFP image of MM" denotes only adding noise to the CFP image of our multimodal model, and "Both modalities of MM" denotes adding noise to OCT volume and CFP image of our multimodal model at the same time.

focus on the correct lesion areas to make the right prediction for the correctly classified positive samples. For the second sample, in the results of the multimodal model, the values of highlighted pixels in the heatmaps of s_{22} and s_{25} are larger than 0.0002 and correctly highlight the lesions. This means that the model makes decisions based on both slices but pays more attention to s_{22}. Comparing the results of the multimodal model to the visualization of single modality-based models, we can see that the results of the OCT-based model are very similar to the OCT part of the multimodal model. They focus on the same OCT slices or the adjacent slices of the same sample. The main difference between the multimodal model and the single-modality model is shown in the CFP images. We find that the heatmaps of CFP images of the multimodal model are more concentrated (as shown in Fig. 6 (a)) than those of the CFP modality (as shown in Fig. 6 (b)) and the highlight region are related to the location of highlighted OCT slices. The potential reason for this phenomenon is that the feature learning of CFP images is successfully enhanced by the multimodal model's geometric correspondence information. It has been revealed that the performance improvement is mainly derived from the enhanced complementary information in CFP images.

To demonstrate the operation of responses and weights for both OCT slices and CFP rows, we visualize the average values of OCT responses and CFP row weights for the correctly classified positive DME samples. We present the responses and weights of different indices in Fig. 7. The top subfigure displays the OCT slice responses, the middle subfigure illustrates the trained filter, and the bottom subfigure shows the CFP row weights. The results indicate that the patterns in OCT slices 17-23 and 25-31 are similar to those of the filter, suggesting that slices within these indices are likely to be positive and influential in decision-making. In Fig. 6, the model focuses on slices within this range, which are highlighted to aid in the decision-making process. This highlights the effectiveness of the trainable filter in conjunction with the OCT slice responses in identifying significant slices. By applying the well-trained filter, we map the OCT slice responses to the row weights of CFP feature maps. This mapping results in larger weights being concentrated around the 5th and 7th rows—around the center part of the input CFP image. This distribution indicates a potential registration within the feature space: the important slices of OCT volume are mainly found in slices 17-23 and 25-31, while the CFP feature maps' rows with high weights are rows 5 and 7. These slice indexes and row numbers correspond to the central parts of the image and volume, respectively. It is also consistent with the visualization of the heatmap of the samples as shown in Fig. 6.

To evaluate the robustness of our model, we add different levels of Gaussian noise to the input CFP and OCT images of three models and compare the AUROC scores for the DME diagnosis on the DME dataset. The three models consist of a single CFP-based model, a single OCT-based model, and our multimodal model. For the single modal-based models, we add Gaussian noise to the input image directly. In the case of our multimodal model, we add noise under three different scenarios: 1) only on the CFP image; 2) only on the OCT volume; and 3) on both the CFP image and the OCT volume. The Gaussian noise level ranges from 0 to 0.8 and is sampled at intervals of 0.1. The figure displaying the AUROC score changes of all models on the test set under different levels of Gaussian noise is shown in Fig. 8. We have the following findings: 1) The multimodal model is more robust than the single CFP-based model and similar to the OCT-based model. When the noise level is larger than 0.7, the multimodal method is more robust than the single OCT-based model. 2) The OCT input is more sensitive to noise than the CFP input when the noise level is larger than 0.5, both under the multimodal and single-modal settings. 3) For the multimodal model, adding noise only to the CFP image has no obvious impact on the multimodal decision, as the OCT input can provide complementary information to make the right decision. However, adding noise to the OCT input significantly impacts the decision as the noise level increases.

To assess the impact of location shifts between CFP and OCT, we experiment with varying shift magnitudes on CFP feature maps to simulate the location on two modalities in both the VA and Glaucoma classification tasks. In these experiments, we use the well-trained weights and focused solely on testing different shift sizes. Since it is no longer possible to submit test results to the challenge platform for Glaucoma classification[1], we report the Glaucoma Kappa values based on the validation set. Our dataset primarily includes macula center images, which suggests that shifts between different CFP images and OCT volumes are likely minimal. We implement the shifts of 0, 1, 2, 3, and 4 rows on the CFP feature maps, where a ’0’ shift indicates no movement, and other values correspond to shifts by the respective number of rows on the fundus feature maps. Given that our input images are 410 by 410 pixels, the size of the feature maps is 11×11 after feature extraction. A one-row shift in the feature maps corresponds to approximately a 9% shift (38 pixels) in the CFP image and around 15% in the corresponding OCT volume region of the DME dataset when zoomed to match the input CFP image size. This shift ratio should be considered a rough estimate. This is because a single feature in the feature map of the last layer may correspond to a significantly large perception region in the input. This is due to the fact that the input undergoes multiple pooling and downsampling layers before the final feature maps are obtained. The AUROC scores under these various shift sizes, as shown in Fig. 9, reveal that while smaller shifts do not significantly impact the model’s performance, larger shifts lead to a minor decrease: specifically, a 0.2% decrease in AUROC after a 4-row shift in VA classification and a 3% decrease in Kappa after a 3-row shift in Glaucoma classification. This result is likely due to our method’s use of OCT location information to enhance feature learning in CFP images. After extensive training, the CFP backbone is able to extract superior features compared to when it is trained solely with CFP images. The features in the last feature map layer are derived from a large conception region. Therefore, during testing, the impact of fundus row weights on the CFP features is relatively limited. Besides, our method also utilizes the complementary information of OCT to make decisions together, the information from OCT can maintain a high performance.

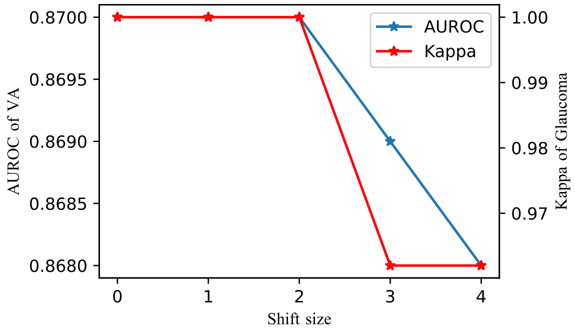


Fig. 9. The AUROC and Kappa scores vary under different shift sizes for VA and glaucoma classifications, respectively. the ‘shift size’ refers to the number of rows shifted on the CFP feature maps.

V. Conclusion

In this work, we proposed a geometric correspondence-based multimodal learning network (GeCoM-Net) to diagnose eye diseases and conditions using CFP and OCT image modalities. It leverages the geometric relationships among the CFP image and its corresponding OCT slices to learn correlated and complimentary features from the two modalities to improve diagnosing accuracy. Moreover, we designed a new feature selection module (MIFS) to select the activated feature maps from OCT slices, which is essential to reduce redundant information and simultaneously preserve discriminative information for the feature extraction of the OCT modality. We also designed a new geometric correspondence-based attention (GeCoA) module to learn the geometry relationship between CFP image and OCT slices, which is essential to learn the correlated features between CFP image and OCT volume. Experiments on a large dataset (DME dataset) and a popular public dataset (GAMMA) demonstrate that GeCoM-Net outperforms the current SOTA methods for DME, impaired VA and glaucoma diagnoses, which verifies the effectiveness of our proposed strategy.

References

[1] S. R. Flaxman *et al.*, “Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis,” *The Lancet Global Health*, vol. 5, no. 12, pp. e1221–e1234, 2017.

[2] U. Farooq and N. Y. Sattar, “Improved automatic localization of optic disc in retinal fundus using image enhancement techniques and SVM,” in *Proceedings of the IEEE International Conference on Control System, Computing and Engineering*. IEEE, 2015, pp. 532–537.

[1] https://aistudio.baidu.com/competition/detail/119/0/submit-result

[3] T. Li *et al.*, "Applications of Deep Learning in Fundus Images: A Review," *Medical Image Analysis*, vol. 69, p. 101971, apr 2021.

[4] J. G. Fujimoto, C. Pitris, S. A. Boppart, and M. E. Brezinski, "Optical coherence tomography: an emerging technology for biomedical imaging and optical biopsy," *Neoplasia*, vol. 2, no. 1-2, pp. 9–25, 2000.

[5] I. Qureshi, J. Ma, and Q. Abbas, "Diabetic retinopathy detection and stage classification in eye fundus images using active deep learning," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11 691–11 721, 2021.

[6] A. Sharafeldeen *et al.*, "Precise higher-order reflectivity and morphology models for early diagnosis of diabetic retinopathy using oct images," *Scientific Reports*, vol. 11, no. 1, pp. 1–16, 2021.

[7] H. Fu *et al.*, "Disc-aware ensemble network for glaucoma screening from fundus image," *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2493–2501, 2018.

[8] O. Perdomo, S. Otálora, F. A. González, F. Meriaudeau, and H. Müller, "Oct-net: A convolutional network for automatic classification of normal and diabetic macular edema using sd-oct volumes," in *Proceedings of the IEEE International Symposium on Biomedical Imaging*. IEEE, 2018, pp. 1423–1426.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[11] Y. Wang, Z. Wang, Y. Feng, and L. Zhang, "WDCCNet: Weighted double-classifier constraint neural network for mammographic image classification," *IEEE Transactions on Medical Imaging*, vol. 41, no. 3, pp. 559–570, 2021.

[12] Y. Feng *et al.*, "Deep supervised domain adaptation for pneumonia diagnosis from chest x-ray images," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1080–1090, 2021.

[13] X. Xu *et al.*, "MSCS-DeepLN: Evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks," *Medical Image Analysis*, vol. 65, p. 101772, 2020.

[14] Á. S. Hervella, J. Rouco, J. Novo, and M. Ortega, "End-to-end multi-task learning for simultaneous optic disc and cup segmentation and glaucoma classification in eye fundus images," *Applied Soft Computing*, vol. 116, p. 108347, 2022.

[15] Z. Shen, H. Fu, J. Shen, and L. Shao, "Modeling and enhancing low-quality retinal fundus images," *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 996–1006, 2020.

[16] Y. Chen, J. Zhong, and Z. Yi, *Intelligent Analysis of Fundus Images: Methods and Applications*. World Scientific, 2023.

[17] S. Li *et al.*, "Multi-instance multi-scale CNN for medical image classification," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 531–539.

[18] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[19] X. Li *et al.*, "Multi-modal multi-instance learning for retinal disease recognition," in *Proceedings of the ACM International Conference on Multimedia*, 2021, p. 2474–2482.

[20] W. Wang *et al.*, "Two-stream CNN with loose pair training for multi-modal AMD categorization," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 156–164.

[21] L. Zhen, P. Hu, X. Peng, R. S. M. Goh, and J. T. Zhou, "Deep multimodal transfer learning for cross-modal retrieval," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 798–810, 2022.

[22] Y. Wang *et al.*, "Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images," *Medical Image Analysis*, vol. 81, p. 102535, 2022.

[23] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: a review," *Briefings in Bioinformatics*, vol. 23, no. 2, p. bbab569, 2022.

[24] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[25] T. K. Yoo, J. Y. Choi, J. G. Seo, B. Ramasubramanian, S. Selvaperumal, and D. W. Kim, "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.

[26] D. J. Gaddipati and J. Sivaswamy, "Glaucoma assessment from fundus images with fundus to OCT feature space mapping," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–15, 2021.

[27] G. An *et al.*, "Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images," *Journal of Healthcare Engineering*, vol. 2019, 2019.

[28] T. Shehryar *et al.*, "Improved automated detection of glaucoma by correlating fundus and SD-OCT image analysis," *International Journal of Imaging Systems and Technology*, vol. 30, no. 4, pp. 1046–1065, 2020.

[29] S. Aumann, S. Donner, J. Fischer, and F. Müller, "Optical coherence tomography (oct): principle and technical realization," *High resolution imaging in microscopy and ophthalmology: new frontiers in biomedical optics*, pp. 59–85, 2019.

[30] N. Tsiknakis *et al.*, "Deep learning for diabetic retinopathy detection and classification based on fundus images: A review," *Computers in Biology and Medicine*, vol. 135, p. 104599, 2021.

[31] J. J. Gómez-Valverde *et al.*, "Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning," *Biomedical Optics Express*, vol. 10, no. 2, pp. 892–913, 2019.

[32] J. Hu, Y. Chen, J. Zhong, R. Ju, and Z. Yi, "Automated analysis for retinopathy of prematurity by deep neural networks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 1, pp. 269–279, 2018.

[33] K. Shankar, A. R. W. Sait, D. Gupta, S. Lakshmanaprabu, A. Khanna, and H. M. Pandey, "Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model," *Pattern Recognition Letters*, vol. 133, pp. 210–216, 2020.

[34] R. Tennakoon *et al.*, "Classification of volumetric images using multi-instance learning and extreme value theorem," *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 854–865, 2019.

[35] X. Wang *et al.*, "UD-MIL: Uncertainty-driven deep multiple instance learning for OCT image classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3431–3442, 2020.

[36] K. Thakoor, D. Bordbar, J. Yao, O. Moussa, R. Chen, and P. Sajda, "Hybrid 3d-2d deep learning for detection of neovascularage-related macular degeneration using optical coherence tomography b-scans and angiography volumes," in *Proceedings of the IEEE International Symposium on Biomedical Imaging*. IEEE, 2021, pp. 1600–1604.

[37] Y. George, B. J. Antony, H. Ishikawa, G. Wollstein, J. S. Schuman, and R. Garnavi, "Attention-guided 3D-CNN framework for glaucoma detection and structural-functional association using volumetric images," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3421–3430, 2020.

[38] Y. Zhang, X. Ma, M. Li, Z. Ji, S. Yuan, and Q. Chen, "LamNet: A lesion attention maps-guided network for the prediction of choroidal neovascularization volume in SD-OCT images," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1660–1671, 2021.

[39] W. Chen, X. Gong, and Z. Wang, "Neural architecture search on imagenet in four gpu hours: A theoretically inspired perspective," in *Proceedings of the International Conference on Learning Representations*, 2021.

[40] J. Zhao *et al.*, "The fusion strategy of 2d and 3d information based on deep learning: A review," *Remote Sensing*, vol. 13, no. 20, p. 4029, 2021.

[41] J.-S. Lee and T.-H. Park, "Fast road detection by cnn-based camera–lidar fusion and spherical coordinate transformation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 9, pp. 5802–5810, 2020.

[42] X. Lv, Z. Liu, J. Xin, and N. Zheng, "A novel approach for detecting road based on two-stream fusion fully convolutional network," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1464–1469.

[43] C. R. Qi, X. Chen, O. Litany, and L. J. Guibas, "Imvotenet: Boosting 3d object detection in point clouds with image votes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4404–4413.

[44] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 918–927.

[45] M. S. Miri, M. D. Abràmoff, Y. H. Kwon, and M. K. Garvin, "Multimodal registration of sd-oct volumes and fundus photographs using histograms of oriented gradients," *Biomedical Optics Express*, vol. 7, no. 12, pp. 5252–5267, 2016.

[46] M. Mokhtari *et al.*, "Local comparison of cup to disc ratio in right and left eyes based on fusion of color fundus images and oct b-scans," *Information Fusion*, vol. 51, pp. 30–41, 2019.

[47] K. Jin *et al.*, "Multimodal deep learning with feature level fusion for identification of choroidal neovascularization activity in age-related

macular degeneration," *Acta Ophthalmologica*, vol. 100, no. 2, pp. e512–e520, 2022.

[48] W. Wang *et al.*, "Learning two-stream CNN for multi-modal age-related macular degeneration categorization," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[49] X. He, Y. Deng, L. Fang, and Q. Peng, "Multi-modal retinal image classification with modality-specific attention network," *IEEE Transactions on Medical Imaging*, vol. 40, no. 6, pp. 1591–1602, 2021.

[50] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, 2020.

[51] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[52] J. K. Sun and L. M. Jampol, "The diabetic retinopathy clinical research network (DRCR.net) and its contributions to the treatment of diabetic retinopathy," *Ophthalmic Research*, vol. 62, no. 4, pp. 225–230, 2019.

[53] J. Wu *et al.*, "Gamma challenge: glaucoma grading from multi-modality images," *arXiv preprint arXiv:2202.06511*, 2022. [Online]. Available: https://arxiv.org/abs/2202.06511

[54] J. A. Wells *et al.*, "Aflibercept, bevacizumab, or ranibizumab for diabetic macular edema: two-year results from a comparative effectiveness randomized clinical trial," *Ophthalmology*, vol. 123, no. 6, pp. 1351–1359, 2016.

[55] C. W. Baker *et al.*, "Effect of initial management with aflibercept vs laser photocoagulation vs observation on vision loss among patients with diabetic macular edema involving the center of the macula and good visual acuity: a randomized clinical trial," *Jama*, vol. 321, no. 19, pp. 1880–1894, 2019.

[56] T. Y. Wong *et al.*, "Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings," *Ophthalmology*, vol. 125, no. 10, pp. 1608–1622, 2018.

[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[59] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," *CoRR*, vol. abs/1605.07678, 2016. [Online]. Available: http://arxiv.org/abs/1605.07678

[60] K. Zou *et al.*, "Reliable multimodality eye disease screening via mixture of student's t distributions," *CoRR*, vol. abs/2303.09790, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.09790

[61] Y. Li *et al.*, "Multimodal information fusion for glaucoma and diabetic retinopathy classification," in *Ophthalmic Medical Image Analysis*, B. Antony, H. Fu, C. S. Lee, T. MacGillivray, Y. Xu, and Y. Zheng, Eds. Cham: Springer International Publishing, 2022, pp. 53–62.

[62] F. Li *et al.*, "A multicenter clinical study of the automated fundus screening algorithm," *Translational Vision Science & Technology*, vol. 11, no. 7, pp. 22–22, 2022.

[63] R. Hemelings *et al.*, "A generalizable deep learning regression model for automated glaucoma screening from fundus images," *NPJ Digital Medicine*, vol. 6, no. 1, p. 112, 2023.

[64] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.