Multi-modal Genotype and Phenotype Mutual Learning to Enhance Single-Modal Input Based Longitudinal Outcome Prediction

Alireza Ganjdanesh[1], Jipeng Zhang[2], Wei Chen[2,3,4](✉), and Heng Huang[1](✉)

[1] Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA
{alireza.ganjdanesh,heng.huang}@pitt.edu

[2] Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA
jiz214@pitt.edu

[3] Department of Pediatrics, UPMC Children's Hospital of Pittsburgh, Pittsburgh, PA, USA
wei.chen@chp.edu

[4] Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA, USA

Abstract. In recent years, due to the advance of modern sensory devices, the collection of multiple biomedical data modalities such as imaging genetics has gotten feasible, and multimodal data analysis has attracted significant attention in bioinformatics. Although existing multimodal learning methods have shown superior ability in combining data from multiple sources, they are not directly applicable for many real-world biological and biomedical studies that suffer from missing data modalities due to the high expenses of collecting all modalities. Thus, in practice, usually, only a *main* modality containing a major 'diagnostic signal' is used for decision making as *auxiliary* modalities are not available. In addition, during the examination of a subject regarding a chronic disease (with longitudinal progression) in a visit, typically, two diagnosis-related questions are of main interest that are what their status currently is (diagnosis) and how it will change before their next visit (longitudinal outcome) if they maintain their disease trajectory and lifestyle. Accurate answers to these questions can distinguish vulnerable subjects and enable clinicians to start early treatments for them. In this paper, we propose a new adversarial mutual learning framework for longitudinal prediction of disease progression such that we properly leverage several modalities of data available in training set to develop a more accurate model using single-modal for prediction. Specifically, in our framework, a single-modal

This work was partially supported by NSF IIS 1845666, 1852606, 1838627, 1837956, 1956002, IIA 2040588.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-04749-7_13.

I. Pe'er (Ed.): RECOMB 2022, LNBI 13278, pp. 209–229, 2022.
https://doi.org/10.1007/978-3-031-04749-7_13

model (that utilizes the *main* modality) learns from a pretrained multimodal model (which takes both *main* and *auxiliary* modalities as input) in a mutual learning manner to 1) infer outcome-related representations of the *auxiliary* modalities based on its own representations for the *main* modality during adversarial training and 2) effectively combine them to predict the longitudinal outcome. We apply our new method to analyze the retinal imaging genetics for the early diagnosis of Age-related Macular Degeneration (AMD) disease in which we formulate prediction of longitudinal AMD progression outcome of subjects as a classification problem of simultaneously grading their current AMD severity as well as predicting their condition in their next visit with a preselected time duration between visits. Our experiments on the Age-Related Eye Disease Study (AREDS) dataset demonstrate the superiority of our model compared to baselines for simultaneously grading and predicting future AMD severity of subjects.

1 Introduction

Recent advances in multimodal biomedical imaging and high throughput genotyping and sequencing techniques allow us to study integrative imaging genetics and provide exciting new opportunities to ultimately improve our understanding of different disease mechanisms. Although many multimodal learning methods have been developed and shown superior ability in integrative analysis of imaging genetics data, the following two challenging problems are still desired to address for practical applications:

Input Data with Missing Modalities: An ideal case is that the researchers or clinicians have access to all of the informative data modalities for decision making, *i.e.*, be able to perform multimodal data based diagnosis. However, due to the high cost of collecting all data modalities, typically, only a single *main* modality that provides the majority of 'signal' about a subject's status is examined in practice. For instance, it has been established that genetic factors play an essential role in the progression of Age-related Macular Degeneration (AMD) pathogenesis [20,21,71,77]. Thanks to advances in sequencing technologies [1,48,49], the determination of whole-genome sequence is feasible nowadays and can provide valuable information for AMD diagnosis, but AMD severity score [19] is usually only determined by exploring characteristics of subjects' Color Fundus Photographs (CFP) - that is the most accessible retinal image modality globally - in practice due to lack of expensive facilities required for sequencing, especially in low-resourced areas.

Diagnosis and Prediction of Longitudinal Outcome: Many diseases have several stages in terms of severity, and a subject may progress to advanced ones through time. Predicting the disease progression can help understand the disease's dynamics and thus, advise physicians on medication intake. Two questions of main interest when studying a subject's condition in clinical practice are that given their examination records, "how is current severity status of them?" (*diagnosis*), and "how will their disease severity change until their next visit?" (*i.e.*, *longitudinal outcome prediction*) Accurate answers to these questions can comprehensively predict a subject's current status as well as their future disease trajectory and enable clinicians to start early treatment for highly vulnerable ones to decelerate their disease progression. However, it is often prohibited to collect the time series biomedical data (from multiple years visits) to predict the disease progression in practical applications, especially for low-resourced areas. The researchers and clinicians often want to make the diagnosis and longitudinal outcome prediction only using the data at the current visit, which makes the disease progression prediction more challenging.

We aim to solve both challenging tasks in the second aspect while considering the constraints mentioned in the first one. To do so, firstly, our intuition is that single-modal input based models that benefit from the *main* and *auxiliary* data modalities collected in multi-modal datasets during training and rely on the *main* modality in their inference phase better mimic clinical practice. Therefore, we train such a model in our framework. Secondly, we can overcome the longitudinal prediction challenge by leveraging records collected at the current visit to make predictions for the current and next visits if the time gap between them is not too large compared to the typical pace of the disease progression.

Multimodal learning (MML) [22,23,44,69,79] and Deep Mutual Learning (DML) [31,81] methods have shown significant results recently. On the one hand, MML methods can effectively utilize the supervision from several modalities to improve the classification performance in tasks such as visual question answering and video categorization. However, they require that all input modalities be available for their inference, which limits their practicality for biomedical applications that usually suffer from missing modalities. On the other hand, DML methods have demonstrated that two models that are trained together and get feedback from their peers have better generalization performance compared to their baseline models that are trained separately. Thus, our intuition is to overcome the missing modality problem of multimodal learning methods for our task by developing a single-modal model while leveraging the benefits of mutual learning by training the model mutually with a multimodal one.

In this paper, we introduce a novel framework based on deep mutual learning [31,81] in which a single-modal model – our model only need the *main* diagnostic modality (*e.g.* CFP) of a target disease (*e.g.* AMD) to conduct the predictions – and a pretrained multimodal model that takes the *main* and *auxiliary* (genetics and age) data modalities as input evolve together during training. Both models learn to solve our formulated classification problem to simultaneously 1) grade the current disease status of a subject (Advanced or not) and 2) predict their future condition in their next visit (Advanced or not, with a predefined time-gap between visits, *e.g.* 3 years). Further, we hypothesize that genetics and demographics (age) information can provide 'complementary knowledge' for a model for longitudinal outcome prediction, especially in the subjects with similar fundus images that may have different future trajectories due to their genetic differences. Therefore, we design our framework such that the single-modal model learns to infer outcome-related representations of *auxiliary* modalities using its representations for the *main* modality from its multimodal colleague using a

Riemannian adversarial training scheme. After that, it combines them to make the predictions. In addition, we use entropy regularization during the pretraining stage of the multimodal model to prevent it from neglecting noisy auxiliary modalities and focusing only on the main one. Our contributions can be summarized as follows:

- We introduce a new framework to simultaneously diagnose current status and predict the longitudinal outcome of subjects for disease progression by developing a model that only requires the *main* diagnostic modality – collected at current visit – for its predictions while properly leveraging *auxiliary* modalities available in the training set to enhance final model's performance.
- We propose to model the complex relationship of representations of the main modality and auxiliary ones by Riemannian Generative Adversarial Networks.
- We design a functional entropy regularized pretraining scheme for the multimodal model to prevent it from shortcut learning to discard the auxiliary modality and only use the more informative main modality.

2 Related Work

Multi-Modal Learning (MML): MML combines knowledge from several modalities to enhance predictions for a target task. It has achieved significant results in domains such as video understanding and visual question answering that leverage several types of visual, audial, or textual data [2,17,22–24,28,35,39,44,50,56,67,70,79]. However, these works assume that all modalities are present during training and inference which limits their direct application in medical problems that missing modalities are a common challenge in them. A popular workaround is to reconstruct and impute missing modalities using available ones [14,47,57,61,64,66,76]. However, reconstruction of extremely high-dimensional modalities such as genetics ($\sim$1.6 $\times 10^5$ dimensional in our problem) is not practical in healthcare problems with limited training data. Further, predicting some modalities from others may not always be feasible. For instance, prediction of one of RGB and thermal images [76] from the other is sensible, but reconstruction of whole-genome sequence from fundus images of eyes is not. Another group of methods proposes variational approaches to deal with missing modalities and model the joint posterior of representations of modalities as a product-of-experts [74]. Lee and Van der Schaar [42] use this method to integrate multi-omics data and train modality-specific predictors to ensure representations of individual modalities are learned faithfully. Nevertheless, a modality-specific predictor is not reasonable in the longitudinal prediction of disease outcome for modalities such as genetics that are *static* while the disease status of a subject may change in time. This is the case for the method of Wang et al. [69] as well that trains modality-specific classifiers with incomplete data pairs and train a final multi-modal model using limited complete pairs while distilling [27,34,45] the knowledge of pretrained models in it.

Deep Mutual Learning (DML): In a nutshell, two or several models are trained simultaneously in DML such that each model gets supervision from training labels and predictions/representations of other models. Zhang et al. [81] introduced DML and showed it has better image classification performance compared to knowledge distillation [27,34,45] methods. Since then, different types of DML for various applications such as image classification [31,41,59,73], semi-supervised learning [75], self-supervised learning [8,68], and object detection [54] have been proposed. These models are not suitable for our problem as they train two models with the same input modality. Recently, Zhang et al. [80] proposed a multimodal image segmentation model to train two single-modal models in a DML manner. However, their multimodal DML idea is designed for problems that their modalities are two 'views' of the same phenomenon, not 'complementary' modalities such as CFP and genetics for AMD that CFP contains the majority of the diagnostic signal while noisy genetics input only complements the knowledge from CFP.

Age-Related Macular Degeneration (AMD): In this paper, we analyze the retinal imaging genetics data which were collected to study the AMD disease and are a good testing platform to evaluate our new method. AMD is a chronic disease [46] that causes the progressive decline of vision due to the dysfunction of the central retina in older adults and is the major root of blindness in elder Caucasians [9,16,65]. Based on a scale called AMD severity score, three stages are defined for AMD: early, intermediate, and late (advanced) [19]. The severity score is determined by exploring characteristics of the Color Fundus Photographs (CFP) of subjects. The main symptom of the early and intermediate stages is the presence of yellowish deposits called 'drusen' in the retina, and most patients are asymptomatic in them [5,29]. The irreversible stage that is accompanied by severe vision loss is late AMD that appears in two forms: 'Dry' and 'Wet'. In Dry AMD (Geographic Atrophy), accumulation of drusen in the retina decreases its sensitivity to light stimuli and causes gradual loss of central vision. In Wet AMD (Choroidal Neovascularization), the growth of leaky blood vessels under the retina damages photoreceptor cells and affects visual acuity. GWAS studies have shown that genetic and environmental factors are critical elements associated with AMD [20,21,71] and its progression time [77]. In recent years, multiple deep learning based predictive models are proposed for AMD. They have two categories: 1) diagnostic models that predict AMD severity of a subject based on their CFP taken at their current visit [11–13,29,38,52]. Although these models have shown convincing performance for the *diagnosis* task, they cannot predict subjects' *longitudinal outcome* that is crucial information for clinicians to start preventive treatments for vulnerable subjects. 2) Models predicting whether a subject progresses into late AMD in less than *'n'* years [10,53,78], where *'n'* is a predefined value. Nonetheless, if their answer is yes, they do not provide any information about whether the subject is already in advanced AMD or they will progress to it in the future. Furthermore, the majority of previous works are single-modal based on CFPs that waste genetic modality in training datasets or they are multi-modal [53,78] taking CFPs and 52 AMD-associated variants [77] which limits their practicality because they need genetic modality in their inference phase.

3 Proposed Method

We develop an adversarial mutual framework capable of utilizing *auxiliary* modalities (genetics and age) available in training set to improve the training of a single-modal model (using only *main* modality (CFP)) that simultaneously addresses main queries regarding a subject's status when a chronic disease is concerned that are: 1) the current status of a subject (*e.g.*, current AMD severity) and 2) how their status will change until their next visit (*e.g.*, how their AMD severity score will change in the near future, i.e., longitudinal outcome) if they maintain their current lifestyle and disease progression trajectory. This knowledge empowers practitioners to start early treatment to decelerate the disease progression for susceptible subjects. We explain the intuitions behind our model step by step in the following subsections using AMD terminologies, but as we noted, it is applicable for similar diseases as well. Our procedure can be seen in Fig. 1.

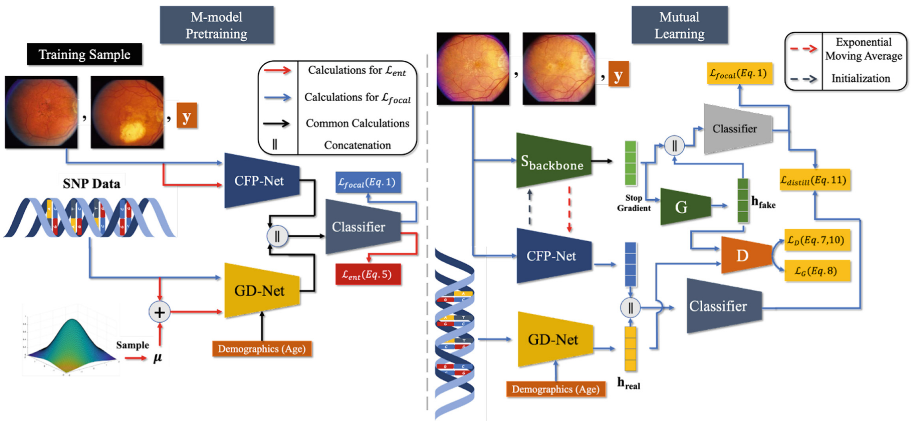

Fig. 1. Overview of our framework. **Left:** pretraining of our multimodal M-model. Color Fundus Photographs (CFP) and genetics information of subjects and are used to train the model. CFP contains the majority of the 'diagnostic signal' related to AMD. Thus, to prevent the model to get biased toward CFP and discard the genetic modality, we impose entropy regularization using a Gaussian measure on the model during training. (Sect. 3.3) **Right:** mutual learning of our single-modal S-model (top) with the pretrained M-model (bottom). S-model learns from the M-model to infer joint AMD-related representations of the genetics and demographics modalities - using its representations for an input CFP - using a Riemannian GAN model. The backbone of the S-model gets initialized by the weights of the CFP-Net of the M-model, and the M-model evolves during training by updating its CFP-Net using the exponential moving average of the weights of the S-model's backbone. (Sect. 3.3)

3.1 Problem Formulation

We formulate our prediction task as a classification problem. Considering AMD severity condition of a subject in their current and next visits (with a pre-defined time gap T_{gap} between them e.g., $T_{gap} = 3$ years), we define three classes: 1) $y = 0$ if a subject is not in the advanced AMD condition and will not progress to it until their next visit. 2) $y = 1$ if they are not currently in the advanced stage but will progress to advanced AMD until their next visit. 3) $y = 2$ if they have already progressed to the advanced phase. As there is no treatment for late AMD yet, [65] the fourth case for (current, next) $\sim$ (advanced, not advanced) is not possible. Our goal is to develop a model that accurately classifies subjects into one of the mentioned classes based on their current visit's CFP images. This formulation enables us to overcome the challenge of heterogeneity of time gaps between consecutive visits for subjects in longitudinal datasets. For instance, we can use records of a subject at visit numbers $\{1, 3, 7, 9\}$ to train a model with $T_{gap} = 2$ with pairs $\{(1, 3), (7, 9)\}$, but a sequence model should handle uneven time gaps $(2, 4, 2)$ between successive visits.

3.2 Notation

Let us assume that we have a longitudinal dataset such as AREDS [60] in which each subject has a random number of records corresponding to the visit time points that their data is collected during the study. We denote the training dataset as $D = \{(x_{i_g}, \{(x_{i_f,t_j}, y_{i,t_j})\})|i \in [N], t_j \in T_i, T_i \subseteq T\}$ where N is the number of subjects, T is the set of all possible visit indices during the study, T_i is the set of available visit indices for the *i-th* subject, x_{i_f,t_j} is the fundus image of the subject taken during the visit with index t_j, and x_{i_g} is the genetic modality of the subject, which is static. For example, in the AREDS dataset [60], examinations are performed every six months, and the maximum follow-up study length for a subject is 13 years (26 visits). Thus, $T = \{1, 2, \cdots, 26\}$ is the set of all possible visit numbers. In addition, we denote our single-modal model as S-model and multimodal on as M-model in the rest of the paper.

3.3 Longitudinal Predictive Model

We introduce an adversarial mutual learning framework in which the single-modal S-model learns from a pretrained multi-modal M-model model to 1) infer outcome-related joint representation of genetics and demographics (age) from its representations for input CFPs using a Riemannian GAN model - inspired by studies [20,21,71,77] that have established high association between these modalities and AMD severity outcome that make it reasonable to incorporate such prior in our model - and 2) combining the predicted representation and the one for the visual modality to solve longitudinal outcome classification task in the course of a mutual training scheme [31,81] that benefits both models. In summary, our algorithm consists of pretraining the multimodal M-model and Mutual training the S-model along with the M-model. We describe details of each one in the following.

M-model Pretraining: We use a multimodal M-model to guide the training process of the S-model in a mutual learning fashion. The architecture of the M-model is shown in Fig.1. It consists of two sub-networks: 1) **CFP-net:** ResNet [33] backbone for CFP modality and 2) **GD-net:** a feed-forward model that combines genetics as well as demographics (age) modalities to obtain a joint outcome-related representation for them. Finally, obtained representations are combined in an early fusion [7] scheme and passed to a classifier to perform prediction.

As the number of samples in the case group (advanced AMD condition) is far less than the control group in our problem, our classification problem is imbalanced. We use Focal loss [43] to train the M-model because it down-weights the contribution of 'simple' examples from majority classes (e.g., control cases without any symptoms that the model can easily classify) in the loss function that the model is already confident about them. Formally, given y_i is the correct class corresponding to a sample x and $p_i = \mathcal{P}_{model}(\mathbf{y} = y_i|\mathbf{x} = x)$ be the predicted conditional probability of our teacher model for class y_i given x, Focal loss for the training sample (x, y) is calculated as

$$L_{focal}(x, y) = -(1 - p_i)^{\gamma} \log(p_i) \tag{1}$$

where γ is a hyperparameter controlling the down-weighting factor. As can be seen, Focal loss is a scaled version of Cross-Entropy loss that has a lower value for confident predictions of the model.

As we mentioned, the CFP of subjects contains the majority of the 'diagnostic singal' regarding their AMD status, and the genetics modality provides complementary knowledge with a much lower signal-to-noise ratio compared to the CFP modality. Therefore, directly training the model with Focal loss and standard regularization schemes for deep learning training such as ℓ_2-norm of weights that prefers networks with simpler structures may bias the model to discard the genetic modality and only focus on the CFP one. This phenomenon has been observed in the literature for domains such as visual question answering [2,17,28]. To overcome this problem, we use functional entropy regularization that balances the contribution of modalities. The intuition is that if our model's predictions show high entropy when we perturb a modality, then it is not bypassing the modality. Formally, given a probability measure μ over the space of input x of a non-negative function $g(x)$, functional entropy of g is defined as [6]:

$$Ent(g) = \int g(x) \log(g(x)) d\mu(x) - \int g(x) d\mu(x) \log(\int g(x) d\mu(x)) \tag{2}$$

However, the calculation of the RHS of this equation is intractable. As a workaround, Logarithmic Sobolev Inequality [6,24] is calculated as an upper bound of the functional entropy for Gaussian measures μ:

$$Ent(g) \leq \frac{1}{2} \int \frac{||\nabla g(x)||^2}{g(x)} d\mu(x) \tag{3}$$

In our problem, we define g as a measure of a discrepancy between the softmax output distribution of the M-model when the original genetics modality and its

Gaussian perturbed version of it are inputted to the model while keeping the input CFP fixed. In other words, given an input sample $x = (x_f, x_g)$:

$$\begin{gathered} \mathcal{P}_{model}(\mathbf{y}|\mathbf{x} = (x_f, x_g)) = (p_1, \cdots, p_K) \\ \mathcal{P}_{model}(\mathbf{y}|\mathbf{x} = (x_f, x_g + \epsilon)) = (p'_1, \cdots, p'_K),\ \epsilon \sim \mathcal{N}(0, \Sigma_{x_g}), \\ g(x, \epsilon) \triangleq \tfrac{1}{K} \sum_{j=1}^{K} BCE(p_j, p'_j) \end{gathered} \tag{4}$$

The function g defined in Eq. (4) can represent the sensitivity of the model's predictions to Gaussian perturbations of the genetic modality. Now, we plug g into Eq. (3) and define a loss function $\mathcal{L}_{ent}$ which encourages the model to have high functional entropy *w.r.t* its genetics input:

$$\mathcal{L}_{ent} = -\frac{1}{2} \int \frac{||\nabla g(x, \epsilon)||^2}{g(x, \epsilon)} d\mu(\epsilon)\,. \tag{5}$$

In practice, we estimate the integral using Monte Carlo sampling, i.e., we approximate it with one σ for each sample. In addition, we set Σ_{x_g} as a diagonal covariance matrix with diagonal elements being the empirical variance of samples in the batch in each iteration.

Mutual Learning of S-model and M-model: After pretraining the M-model, we develop a training scheme based on mutual learning to train the S-model. As shown in Fig. 1, S-model has a backbone identical to CFP-net in M-model and a 'predictor' module. We aim to embed two prior medical knowledge into the inductive bias of our model that are: 1) high association between AMD severity and genetic variants [20,21,71,77]. 2) the ability of fundus images to predict the age of subjects [72]. To do so, we use the predictor module inside the S-model to predict representations of GD-net of the M-model. This prediction will be in a much lower dimensional space than reconstructing/imputing the whole genetic and age modalities together [14,47,57,61,64,76], and thus, is more sample efficient. The distribution of joint representation of genetics and age given the representation of CFP images may be multimodal, i.e., the mapping between them not necessarily be bijective. Thus, we train the predictor sub-network of the S-model using Generative Adversarial Networks (GAN) that are capable of modeling complex high dimensional distributions [3,26,30].

Modeling Interactions Between Representation of a CFP and Corresponding Joint Representation of Genetics and Age: We formulate learning such complex interaction with Riemannian GAN [51,58] training. In summary, GAN [3,26,30,51,58] models are trained using a two-player game in which a generator model G aims to learn the underlying distribution of a set of samples in the training set to trick a discriminator model D that distinguishes whether its input is real or a fake one generated by G. As the training process advances, the generator learns the distribution of training samples, and the discriminator will not be able to differentiate between real and fake samples generated by G. Conventional GAN models' discriminators [26] measure the distance between real and fake samples using Euclidean distance between their low dimensional embeddings. However, it is shown that [4,18] such distance may not faithfully reflect distances of data points as it is well-known that high dimensional real-world data is not randomly distributed in the ambient space and are often restricted to a nonlinear low-dimensional manifold [63] with unknown intrinsic dimension. Therefore, Riemannian GAN models' discriminators, project low dimensional representations of samples on a Riemannian manifold such as hypersphere [51,58] and calculate distances between them with the length of geodesics connecting them on the manifold. Distances on hypersphere are limited which makes the training stable, and it is shown that [51] training GAN with geodesic distances on hypersphere is equivalent to minimizing high order Wasserstein distances between real and fake distributions and generalizes methods that minimize the 1-Wasserstein distance [3,30].

Formally, we define a unit hypersphere with a center c and the main axis direction u ($c, u \in \mathbb{R}^d$) that are learnable. Given a joint representation on genetics and age (can be real predicted by GD-net of M-model or fake one by predictor of S-model) input $h \in \mathbb{R}^D$ $(D > d)$ to the discriminator, it projects h into a d-dimensional space using nonlinear layers to obtain an embedding g. Then, it projects g on the unit sphere with center c such that $g_{proj} = \frac{g-c}{||g-c||}$. Now, let's consider circular cross-sections of the hypersphere that the main axis u of the hypersphere is the normal vector of the surface that they lie in. The idea is that if the discriminator gets designed to distinguish between real and fake samples based on the closeness of the cross-section that they lie on to the greatest circle of the hypersphere - i.e., the larger the radius of the cross-section that a sample lies on, more realness score is assigned to it - then the generator will attempt to generate samples that are on the largest circle of the hypersphere. Therefore, it will be able to generate more diverse samples, which prevents mode collapse. Given a batch of samples $H = \{h^i\}_{i=1}^B$, we calculate g^j_{proj} for each sample h^j and decompose it as $g^j_{proj} = g^j_{proj,u} + g^j_{proj,u^\perp}$. The output score of the discriminator for a sample h^j is calculated as:

$$D(h_j) = -\frac{||g^j_{proj,u}||}{\sigma_{proj,u}} + \frac{||g^j_{proj,u^\perp}||}{\sigma_{proj,u^\perp}} \tag{6}$$

where $\sigma_{proj,u}$ and $\sigma_{proj,u^\perp}$ are empirical variances of $||g^j_{proj,u}||$ and $||g^j_{proj,u^\perp}||$ respectively. We use the relativistic objective [37] to train the GAN model. In a nutshell, it is designed such that the generator not only attempts to increase the score of the discriminator for fake samples, but also aims to decrease its score for real samples. If we denote joint representations of GD-net in M-model by $h \sim \mathcal{P}_{GD}$ and the ones predicted by the predictor model of S-model with $h' \sim \mathcal{P}_{pred}$, objectives of G (predictor in S-model) and discriminator D are as follows:

$$\mathcal{L}_D = \max_D \mathbb{E}_{h \sim \mathcal{P}_{GD}}[\log(f(D(h) - \mathbb{E}_{h' \sim \mathcal{P}_{pred}}[D(h')]))] \\ + \mathbb{E}_{h' \sim \mathcal{P}_{pred}}[\log(f(\mathbb{E}_{h \sim \mathcal{P}_{GD}}[D(h)] - D(h')))] \tag{7}$$

$$\mathcal{L}_G = \max_G \mathbb{E}_{h' \sim \mathcal{P}_{pred}}[\log(f(D(h') - \mathbb{E}_{h \sim \mathcal{P}_{GD}}[D(h)]))] \\ + \mathbb{E}_{h \sim \mathcal{P}_{GD}}[\log(f(\mathbb{E}_{h' \sim \mathcal{P}_{pred}}[D(h')] - D(h)))] \tag{8}$$

where $f(z) = sigmoid(\lambda z)$ calculates the discriminator's estimated probability that one/batch of real sample[s] is/are more realistic than a batch/one fake one[s], and λ is a hyperparameter [37]. We train the parameters for the main axis u and center c as follows. In each iteration, given a batch of real and fake samples $H = \{h^i\}_{i=1}^{B}$, at first, we update the center parameter with:

$$\mathcal{L}_c = \frac{1}{|B|} \sum_{j=1}^{|B|} \mathcal{H}(||g^j_{proj} - c||_2) \tag{9}$$

$\mathcal{H}$ is the Huber function [36], and the objective estimates the center of the hypersphere given a batch of samples. Then, we fix the center parameter, and to make the training of the center parameter stable, we encourage the discriminator to map samples to embeddings with similar distances relative to the center, i.e.,

$$\mathcal{L}_{dist} = \frac{1}{|B|} \sum_{j=1}^{|B|} \mathcal{H}(||g^j_{proj} - c||_2 - \sigma_h) \tag{10}$$

where σ_h is the empirical standard deviation of $||g^j_{proj} - c||_2$ distances from projected embeddings to the center. Parameters of the main axis u and discriminator are updated with backpropagated gradients from loss functions in Eqs. (7, 10).

We train the S-model's classifier to combine its representation for CFP and the predicted joint one for genetics and demographics modalities to accurately classify subjects' status. Firstly, we use Focal loss [43] defined in Eq. (1) to leverage training labels. Secondly, we use a distillation loss [34] to guide the S-model using predictions of the M-model:

$$\mathcal{L}_{distill} = KL(\mathcal{P}_S(\mathbf{y}|\mathbf{x}; T), \mathcal{P}_M(\mathbf{y}|\mathbf{x}; T)) \tag{11}$$

where the parameter T is a temperature parameter that controls the sharpness of output softmax distributions of models. In summary, the training objective for S-model's training is:

$$\mathcal{L}_S = \mathcal{L}_{focal} + \lambda_1 \mathcal{L}_{distill} + \lambda_2 \mathcal{L}_G \tag{12}$$

Before starting training the S-model, we initialize its backbone with the weights of the pretrained M-model's CFP-net to make the convergence faster.

As adversarial training may cause instability and degradation of the backbone's representations [15,25,62], we do not backpropagate gradients from adversarial training for the backbone's weights. Instead, we train them using supervision from Focal loss and distillation loss. Finally, as shown that mutual learning benefits from both models getting feedback from their peers, we update M-model's CFP-net's weights with exponential moving average (EMA) of the backbone of the S-model, i.e., after each iteration, we update CFP-net's weights as:

$$\theta_{CFP} \leftarrow \alpha\theta_{CFP} + (1-\alpha)\theta_{Backbone} \tag{13}$$

Doing so prevents corruption of the weights of pre-trained M-model happening when using well-known distillation loss from S-model to M-model [31,81] in the starting phase of training as S-model's predictions are not reliable yet. We summarize our training algorithm in supplementary materials.

4 Experiments

In this section, we evaluate the effectiveness of our proposed adversarial mutual learning method on the task of simultaneously grading the current AMD severity of a subject as well as predicting their longitudinal outcome in their next visit when the predefined time gap between visits are 2, 3, and 4 years respectively. We compare our model with baseline methods, provide its interpretations, and perform an ablation study to analyze the effect of its different components.

4.1 Experimental Setup

Data Description: We use Age-related Eye Disease Study (AREDS) dataset [60] for our experiments, which is the largest longitudinal dataset available for AMD collected and maintained by National Eye Institute (NEI). It is available at the dbGaP[1] AREDS contains longitudinal CFPs of 4628 participants, and a subject may have up to 13-year follow-up visits since the baseline. For preprocessing step, we cropped each CFP to a square that encompasses the Macula [13,52] and resized it to 224×224 pixels resolution. As mentioned in Sect. 1, the yellowish color of drusen in the Macula and the red color of leaky blood vessels are important characteristics of dry and wet AMD respectively. Thus, we use a nonlinear Bézier augmentation [82] - previously proposed for CT scans and X-ray data - followed by random vertical and horizontal flip to augment CFPs. In addition to CFPs, genome sequence of 2780 ($\sim$60%) subjects is available in AREDS. We use all the genetic variants that are in the 34 loci regions [21] associated with advanced AMD with minor allele frequency (MAF) > 0.01 [21], and 156,864 SNPs remain after filtering. We then partition the AREDS dataset on the subject level and take all subjects that their genetics information is available

[1] https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000001.v3.p1.

as our train set. We randomly partition the rest into two halves for our validation and test sets. We refer to supplementary materials for more details about our data preparation.

Baselines: We compare our method against previous mutual learning and knowledge distillation methods in the literature. **DML** [81] trains two models from scratch with different initialization such that each model is trained with a loss function that is the sum of two terms, namely Cross-Entropy loss and KL-divergence between the distributions predicted by the model and its peer. **KDCL** [31] improves DML by using 'ensemble' of models' predictions instead of prediction of the peer model in the KL-divergence term. We use two ensemble schemes for KDCL, namely 'min-logit' and 'mean'. **KD** [34] distills the knowledge in the powerful large pretrained model, called teacher model, into a model, student, by training the student model using KL-divergence loss between its predictions and the ones for the teacher model. In addition, to show the effectiveness of leveraging 'complementary' knowledge in the genetics modality, we compare our model with single-modal baselines such that we train a ResNet architecture with Focal loss and Cross-Entropy loss. We denote these two cases in our experiments as **Base-Focal** and **Base-CE**.

Training and Evaluation: We use multi-class Area Under Curve (AUC) introduced by Hand and Till [32] as our evaluation metric because it is suitable for imbalance classification problems and has been used in AMD literature [13,52,53,78]. We pretrain our M-model for 10 epochs with batch size 128. Then, we train S-model mutually with M-model for 10 epochs with batch size 32. We use the same architectures for two sub-networks of all other mutual learning and knowledge distillation methods, and we use the architecture of our S-model for Base-CE/Focal. By doing so, we reduce the effect of architectural design and can more readily compare the methods. For a fair comparison, we train all baseline models for 20 epochs with batch size 128. We use Adam optimizer [40] with learning rate 0.0003, exponential decay rates $(\beta_1, \beta_2) = (0.9, 0.99)$, and weight decay 0.0001 for all models except for the parameters of the S-model's predictor and discriminator that we set $(\beta_1, \beta_2) = (0.5, 0.999)$, and also, initialize their parameters with normal distribution with zero mean and std of 0.02. We refer to supplementary materials for more details of experiments.

4.2 Experimental Results

Comparison with Baselines Models. Table 1 summarizes the performance of baseline methods and our adversarial mutual learning scheme for simultaneously grading and longitudinal prediction of AMD status of subjects. We explore baseline methods in two settings: 1) genetics modality is incorporated in their training where a multimodal network is trained along with a single-modal one, and we denote them with (M ↔ S). 2) only CFP is used in their training, and two single-modal models are trained together that are shown by (S ↔ S). It can be seen that mutual learning models consistently outperform knowledge distillation and standard single-network training baselines Base-CE/Focal, which is consistent with observations for natural image classification tasks. [31,81] Interestingly, Base-Focal has a competitive or even better performance compared to KD (S ↔ S) and shows better results compared to Base-CE, which shows the superior ability of the Focal loss [43] to handle long-tailed distributions compared to Cross-Entropy loss. In all cases except KDCL-MinLogit with 2 years gap, incorporating the genetics modality in the training procedure of the methods enhances the performance of the final single-modal model in inference, which supports our hypothesis that the genetics modality can provide supervision that is beneficial to the model's training. Furthermore, our model outperforms mutual learning models in all three cases of 2, 3, and 4 years gap between visits that demonstrates our model can more effectively 'denoise' the highly noisy genetics modality during training compared to other baselines and properly learn to predict AMD related joint representation of genetics and demographics modalities from its own one for an input CFP and combine them to perform longitudinal prediction.

Table 1. Comparison of our proposed method with baseline methods. Mean and standard deviation of 5 runs with different initialization are reported.

Time gap		2 years	3 years	4 years
Method	Using auxiliary modality	AUC		
KDCL - MinLogit (M ↔ S) [31]	✓	0.882 ± 0.003	0.881 ± 0.004	0.889 ± 0.003
KDCL - MinLogit (S ↔ S) [31]	×	0.883 ± 0.004	0.880 ± 0.003	0.886 ± 0.004
KDCL - Mean (M ↔ S) [31]	✓	0.876 ± 0.005	0.881 ± 0.003	0.889 ± 0.002
KDCL - Mean (S ↔ S) [31]	×	0.869 ± 0.004	0.874 ± 0.003	0.886 ± 0.005
DML (M ↔ S) [81]	✓	0.879 ± 0.002	0.877 ± 0.004	0.898 ± 0.003
DML (S ↔ S) [81]	×	0.872 ± 0.004	0.874 ± 0.004	0.896 ± 0.004
KD (M ↔ S) [34]	✓	0.872 ± 0.002	0.877 ± 0.003	0.888 ± 0.003
KD (S ↔ S) [34]	×	0.867 ± 0.003	0.873 ± 0.001	0.884 ± 0.001
Base-CE	×	0.862 ± 0.005	0.867 ± 0.005	0.877 ± 0.005
Base-focal	×	0.866 ± 0.003	0.877 ± 0.005	0.881 ± 0.008
AdvML (ours)	✓	**0.896 ± 0.001**	**0.899 ± 0.001**	**0.914 ± 0.001**

Interpretation of S-model's Predictions Figure 2 demonstrates Grad-CAM [55] saliency maps of our S-model. As mentioned in Sect. 1, the main characteristics of AMD in CFPs are the accumulation of yellow deposits called drusen in the Macula of an eye as well as the growth of leaky blood vessels under the retina that cause leakage of blood on photoreceptor cells. Saliency maps in Fig. 2 indicate that our S-model looks for these characteristics in the Macula for decision making, which is aligned with the clinical practice.

Ablation Study: In this section, we perform an ablation study to explore the effect of each component of our model. We remove entropy regularization in M-model's pretraining and the GAN training component in the mutual learning both separately and simultaneously. Table 2 summarizes the results. We can observe that removing entropy regularization for the genetics modality causes more severe performance degradation for our model, which highlights its importance to properly 'debias' the multimodal model to not neglect the genetics modality and only rely on the CFPs and effectively denoise it to extract its discriminative features for classification.

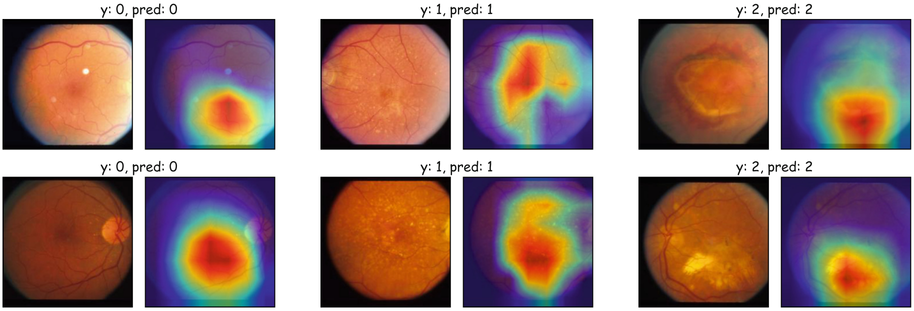

Fig. 2. Grad-CAM [55] saliency maps of our S-model's decisions. It focuses on the Macula region of the eyes and AMD symptoms, namely leaky blood vessels in the retina and yellow deposits in the Macula called drusen, which is aligned with clinical practice. **Left:** neither drusen nor leaky vessels are present in the Macula. **Middle:** Small areas of accumulation of drusen are observable. **Right:** leaked blood in the retina (top) and large areas of drusen (bottom) in the Macula exist.

Table 2. Ablation experiments' results for different components of our method.

Time gap	2 years	3 years	4 years
Ablation experiment	AUC		
W/O Ent Reg	0.880 ± 0.000	0.885 ± 0.001	0.887 ± 0.002
W/O GAN	0.881 ± 0.001	0.889 ± 0.002	0.903 ± 0.002
W/O Ent Reg & GAN	0.871 ± 0.002	0.879 ± 0.003	0.882 ± 0.001

5 Conclusion

In this paper, we introduced a new adversarial mutual learning framework that is capable of leveraging several *auxiliary* diagnostic modalities (containing complementary diagnostic signals that are collected in the training set and missing in inference) to train a more accurate single-modal model which uses the *main* modality (that provides the majority of diagnostic signal and is available in both training and inference) for inference. To do so, the single-modal model is trained with a pretrained multimodal model in a mutual learning manner. We imposed entropy regularization on the multimodal model during its pretraining to encourage it not to neglect the auxiliary modality in its decisions and learn to 'denoise' them to keep their discriminative information. Our single-modal model learns from the multimodal one to infer joint representation of the auxiliary modalities from its representation for the main modality and effectively combine them for its predictions. We modeled the complex interaction between modalities using a Riemannian GAN model and defined our classification task as simultaneously diagnosis of the current status of a subject as well as predicting their longitudinal outcome. We applied our method to the problem of early detection of AMD in which our experiments on the AREDS dataset and our ablation study demonstrated the superiority of our model compared to baselines and the importance of each component for our model.

References

1. Aakur, S.N., Narayanan, S., Indla, V., Bagavathi, A., Laguduva Ramnath, V., Ramachandran, A.: MG-NET: leveraging pseudo-imaging for multi-modal metagenome analysis. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 592–602. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_57
2. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4971–4980 (2018)
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning, pp. 214–223. PMLR (2017)
4. Arvanitidis, G., Hauberg, S., Schölkopf, B.: Geometrically enriched latent spaces. In: Banerjee, A., Fukumizu, K. (eds.) The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, 13–15 April 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 130, pp. 631–639. PMLR (2021). http://proceedings.mlr.press/v130/arvanitidis21a.html
5. Ayoub, T., Patel, N.: Age-related macular degeneration. J. R. Soc. Med. **102**(2), 56–61 (2009)
6. Bakry, D., Gentil, I., Ledoux, M., et al.: Analysis and Geometry of Markov Diffusion Operators, vol. 103. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-00227-9
7. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 423–443 (2018)
8. Bhat, P., Arani, E., Zonooz, B.: Distill on the go: online knowledge distillation in self-supervised learning. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19–25, 2021. pp. 2678–2687. Computer Vision Foundation/IEEE (2021). https://doi.org/10.1109/CVPRW53098.2021.00301
9. Bird, A.C., et al.: An international classification and grading system for age-related maculopathy and age-related macular degeneration. Surv. Ophthalmol. **39**(5), 367–374 (1995)
10. Bridge, J., Harding, S., Zheng, Y.: Development and validation of a novel prognostic model for predicting AMD progression using longitudinal fundus images. BMJ Open Ophthal. **5**(1), e000569 (2020)

11. Burlina, P., Freund, D.E., Joshi, N., Wolfson, Y., Bressler, N.M.: Detection of age-related macular degeneration via deep learning. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pp. 184–188. IEEE (2016)
12. Burlina, P.M., Joshi, N., Pacheco, K.D., Freund, D.E., Kong, J., Bressler, N.M.: Use of deep learning for detailed severity characterization and estimation of 5-year risk among patients with age-related macular degeneration. JAMA Ophthalmol. **136**(12), 1359–1366 (2018)
13. Burlina, P.M., Joshi, N., Pekala, M., Pacheco, K.D., Freund, D.E., Bressler, N.M.: Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. JAMA Ophthalmol. **135**(11), 1170–1176 (2017)
14. Cai, L., Wang, Z., Gao, H., Shen, D., Ji, S.: Deep adversarial learning for multi-modality missing data completion. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1158–1166 (2018)
15. Chavdarova, T., Fleuret, F.: SGAN: an alternative training of generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9407–9415 (2018)
16. Congdon, N., et al.: Causes and prevalence of visual impairment among adults in the united states. Arch. Ophthalmol. (Chicago, Ill.: 1960) **122**(4), 477–485 (2004)
17. Dancette, C., Cadene, R., Teney, D., Cord, M.: Beyond question-based biases: assessing multimodal shortcut learning in visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1574–1583, October 2021
18. Edraki, M., Qi, G.J.: Generalized loss-sensitive adversarial learning with manifold margins. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 87–102 (2018)
19. Ferris III, F.L., et al.: Clinical classification of age-related macular degeneration. Ophthalmology **120**(4), 844–851 (2013)
20. Fritsche, L.G., et al.: Seven new loci associated with age-related macular degeneration. Nat. Geneti. **45**(4), 433–439 (2013)
21. Fritsche, L.G., et al.: A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat. Genet. **48**(2), 134–143 (2016)
22. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: action recognition by previewing audio. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10457–10467 (2020)
23. Garcia, N., Nakashima, Y.: Knowledge-based video question answering with unsupervised scene descriptions. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12363, pp. 581–598. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58523-5_34
24. Gat, I., Schwartz, I., Schwing, A.G., Hazan, T.: Removing bias in multi-modal classifiers: regularization by maximizing functional entropies. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, 6–12 December 2020, virtual (2020). https://proceedings.neurips.cc/paper/2020/hash/20d749bc05f47d2bd3026ce457dcfd8e-Abstract.html
25. Goodfellow, I.: NIPS 2016 tutorial: generative adversarial networks. arXiv preprint arXiv:1701.00160 (2016)

26. Goodfellow, I., et al.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. **27** (2014)
27. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: a survey. Int. J. Comput. Vis. **129**(6), 1789–1819 (2021)
28. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6904–6913 (2017)
29. Grassmann, F., et al.: A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. Ophthalmology **125**(9), 1410–1420 (2018)
30. bibitemch13DBLP:confspsnipsspsGulrajaniAADC17 Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: Guyon, I., et al. (eds.) Annual Conference on Neural Information Processing Systems 2017, vol. 30, 4–9 December 2017, Long Beach, CA, USA. pp. 5767–5777 (2017). https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html
31. Guo, Q., et al.: Online knowledge distillation via collaborative learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11020–11029 (2020)
32. Hand, D.J., Till, R.J.: A simple generalisation of the area under the roc curve for multiple class classification problems. Mach. Learn. **45**(2), 171–186 (2001)
33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
34. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
35. Hou, J.C., Wang, S.S., Lai, Y.H., Tsao, Y., Chang, H.W., Wang, H.M.: Audio-visual speech enhancement using multimodal deep convolutional neural networks. IEEE Trans. Emerg. Topics Comput. Intell. **2**(2), 117–128 (2018)
36. Huber, P.J.: Robust estimation of a location parameter. In: Kotz, S., Johnson, N.L. (eds.) Breakthroughs in Statistics. Springer Series in Statistics, pp. 492–518. Springer, New York (1992). https://doi.org/10.1007/978-1-4612-4380-9_35
37. Jolicoeur-Martineau, A.: The relativistic discriminator: a key element missing from standard GAN. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019. OpenReview.net (2019). https://openreview.net/forum?id=S1erHoR5t7
38. Keenan, T.D., et al.: A deep learning approach for automated detection of geographic atrophy from color fundus photographs. Ophthalmology **126**(11), 1533–1540 (2019)
39. Kim, J., Jun, J., Zhang, B.: Bilinear attention networks. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, vol. 31, 3–8 December 2018, Montréal, Canada, pp. 1571–1581 (2018), https://proceedings.neurips.cc/paper/2018/hash/96ea64f3a1aa2fd00c72faacf0cb8ac9-Abstract.html
40. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015 (2015). http://arxiv.org/abs/1412.6980

41. Lan, X., Zhu, X., Gong, S.: Knowledge distillation by on-the-fly native ensemble. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Annual Conference on Neural Information Processing Systems, 2018, vol. 31 NeurIPS 2018, 3–8 December 2018, Montréal, Canada. pp. 7528–7538 (2018). https://proceedings.neurips.cc/paper/2018/hash/94ef7214c4a90790186e255304f8fd1f-Abstract.html
42. Lee, C., Schaar, M.: A variational information bottleneck approach to multi-omics data integration. In: International Conference on Artificial Intelligence and Statistics, pp. 1513–1521. PMLR (2021)
43. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
44. Lin, X., Bertasius, G., Wang, J., Chang, S.F., Parikh, D., Torresani, L.: Vx2text: end-to-end learning of video-based text generation from multimodal inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7005–7015, June 2021
45. Liu, Y., et al.: Unbiased teacher for semi-supervised object detection. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, 3–7 May 2021. OpenReview.net (2021). https://openreview.net/forum?id=MJIve1zgR_
46. Luu, J., Palczewski, K.: Human aging and disease: lessons from age-related macular degeneration. Proc. Natil. Acad. Sci. **115**(12), 2866–2872 (2018)
47. Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X.: SMIL: multimodal learning with severely missing modality. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2302–2310 (2021)
48. Metzker, M.L.: Sequencing technologies-the next generation. Nat. Rev. Genet. **11**(1), 31–46 (2010)
49. Mikheyev, A.S., Tin, M.M.: A first look at the oxford nanopore minion sequencer. Mol. Ecol. Resour. **14**(6), 1097–1102 (2014)
50. Panda, R., et al.: AdaMML adaptive multi-modal learning for efficient video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7576–7585, October 2021
51. Park, S.W., Kwon, J.: Sphere generative adversarial network based on geometric moment matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4292–4301 (2019)
52. Peng, Y., et al.: DeepSeeNet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. Ophthalmology **126**(4), 565–575 (2019)
53. Peng, Y., et al.: Predicting risk of late age-related macular degeneration using deep learning. NPJ Digit. Med. **3**(1), 1–10 (2020)
54. Qi, L., et al,: Multi-scale aligned distillation for low-resolution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14443–14453 (2021)
55. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)

56. Seo, A., Kang, G., Park, J., Zhang, B.: Attend what you need: motion-appearance synergistic networks for video question answering. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, 1–6 August 2021. pp. 6167–6177. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.481
57. Shi, Y., Narayanaswamy, S., Paige, B., Torr, P.H.S.: Variational mixture-of-experts autoencoders for multi-modal deep generative models. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, vol. 32, 8–14 December 2019, Vancouver, BC, Canada. pp. 15692–15703 (2019). https://proceedings.neurips.cc/paper/2019/hash/0ae775a8cb3b499ad1fca944e6f5c836-Abstract.html
58. Shim, W., Cho, M.: CircleGAN: generative adversarial learning across spherical circles. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 21081–21091. Curran Associates, Inc. (2020). https://proceedings.neurips.cc/paper/2020/file/f14bc21be7eaeed046fed206a492e652-Paper.pdf
59. Son, W., Na, J., Choi, J., Hwang, W.: Densely guided knowledge distillation using multiple teacher assistants. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9395–9404 (2021)
60. Study, T.A.R.E.D., et al.: The age-related eye disease study (AREDS): design implications AREDS report no. 1. Control. Clin. Trials **20**(6), 573–600 (1999)
61. Suo, Q., Zhong, W., Ma, F., Yuan, Y., Gao, J., Zhang, A.: Metric learning on healthcare data with incomplete modalities. In: IJCAI, pp. 3534–3540 (2019)
62. Tao, S., Wang, J.: Alleviation of gradient exploding in GANs: fake can be real. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1191–1200 (2020)
63. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)
64. Tran, L., Liu, X., Zhou, J., Jin, R.: Missing modalities imputation via cascaded residual autoencoder. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1405–1414 (2017)
65. Trucco, E., MacGillivray, T., Xu, Y.: Computational retinal image analysis: tools. In: Trucco, E., MacGillivray, T., Xu, Y. (eds.) Applications and Perspectives, Academic Press, New York (2019)
66. Tsai, Y.H., Liang, P.P., Zadeh, A., Morency, L., Salakhutdinov, R.: Learning factorized multimodal representations. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019. OpenReview.net (2019). https://openreview.net/forum?id=rygqqsA9KX
67. Uppal, S., Bhagat, S., Hazarika, D., Majumder, N., Poria, S., Zimmermann, R., Zadeh, A.: Multimodal research in vision and language: a review of current and emerging trends. Inf. Fusion **77**, 149–171 (2021)
68. Wang, J., Li, Y., Hu, J., Yang, X., Ding, Y.: Self-supervised mutual learning for video representation learning. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2021)

69. Wang, Q., Zhan, L., Thompson, P., Zhou, J.: Multimodal learning with incomplete modalities by knowledge distillation. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1828–1838 (2020)
70. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12695–12705 (2020)
71. Wei, Y., Liu, Y., Sun, T., Chen, W., Ding, Y.: Gene-based association analysis for bivariate time-to-event data through functional regression with copula models. Biometrics **76**(2), 619–629 (2020)
72. Wen, Y., Chen, L., Qiao, L., Deng, Y., Zhou, C.: On the deep learning-based age prediction of color fundus images and correlation with ophthalmic diseases. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1171–1175. IEEE (2020)
73. Wu, G., Gong, S.: Peer collaborative learning for online knowledge distillation. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, 2–9 February 2021, pp. 10302–10310. AAAI Press (2021). https://ojs.aaai.org/index.php/AAAI/article/view/17234
74. Wu, M., Goodman, N.D.: Multimodal generative models for scalable weakly-supervised learning. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3–8 December 2018, Montréal, Canada. pp. 5580–5590 (2018). https://proceedings.neurips.cc/paper/2018/hash/1102a326d5f7c9e04fc3c89d0ede88c9-Abstract.html
75. Wu, S., Li, J., Liu, C., Yu, Z., Wong, H.S.: Mutual learning of complementary networks via residual correction for improving semi-supervised classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6500–6509 (2019)
76. Xu, D., Ouyang, W., Ricci, E., Wang, X., Sebe, N.: Learning cross-modal deep representations for robust pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5363–5371 (2017)
77. Yan, Q., et al.: Genome-wide analysis of disease progression in age-related macular degeneration. Hum. Mol. Genet. **27**(5), 929–940 (2018)
78. Yan, Q., et al.: Deep-learning-based prediction of late age-related macular degeneration progression. Nat. Mach. Intell. **2**(2), 141–150 (2020)
79. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6720–6731 (2019)
80. Zhang, Y., et al.: Modality-aware mutual learning for multi-modal medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 589–599. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_56
81. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4320–4328 (2018)
82. Zhou, Z., et al.: Models genesis: generic autodidactic models for 3D medical image analysis. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 384–393. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_42