Self-Supervised Feature Learning via Exploiting Multi-Modal Data for Retinal Disease Diagnosis

Xiaomeng Li, Mengyu Jia, Md Tauhidul Islam, *Member, IEEE*, Lequan Yu, *Member, IEEE*, and Lei Xing, *Member, IEEE*

***Abstract*— The automatic diagnosis of various retinal diseases from fundus images is important to support clinical decision-making. However, developing such automatic solutions is challenging due to the requirement of a large amount of human-annotated data. Recently, unsupervised/self-supervised feature learning techniques receive a lot of attention, as they do not need massive annotations. Most of the current self-supervised methods are analyzed with single imaging modality and there is no method currently utilize multi-modal images for better results. Considering that the diagnostics of various vitreoretinal diseases can greatly benefit from another imaging modality, *e.g.*, FFA, this paper presents a novel self-supervised feature learning method by effectively exploiting multi-modal data for retinal disease diagnosis. To achieve this, we first synthesize the corresponding FFA modality and then formulate a patient feature-based softmax embedding objective. Our objective learns both modality-invariant features and patient-similarity features. Through this mechanism, the neural network captures the semantically shared information across different modalities and the apparent visual similarity between patients. We evaluate our method on two public benchmark datasets for retinal disease diagnosis. The experimental results demonstrate that our method clearly outperforms other self-supervised feature learning methods and is comparable to the supervised baseline. Our code is available at GitHub.**

***Index Terms*— Retinal disease diagnosis, self-supervised learning, multi-modal data.**

I. Introduction

COLOR fundus photography has been widely used in clinical practice to evaluate various conventional ophthalmic diseases, *e.g.*, age-related macular degeneration (AMD) [1], pathologic myopia (PM) [2], and diabetic retinopathy [3], [4]. Recently, deep learning has shown very good performance on a variety of automatic ophthalmic disease detection problems from fundus images [5]–[7], and these techniques can help ophthalmologists in decision making. The success is attributed to the learned representative features from fundus images, which requires a large amount of training data with massive human annotations. However, it is tedious and expensive to annotate the fundus images, since experts are needed to provide reliable labels. Hence, in this paper, our goal is to learn the representative features from data itself, without any human annotations. Then, the learned representations are evaluated on the fundus image classification tasks.

The self-supervised feature learning methods have been explored in the medical imaging domain for several tasks, such as subject identification from spinal MRI [10], cardiac MR image segmentation [11], brain hemorrhage classification [12] and lung lobe segmentation and nodule detection task [13]. Most previous works focused on developing novel pretext tasks as the supervisory signals to train the network to learn feature representation. For example, [11] proposed to learn self-supervised features by predicting anatomical positions from MR images. Reference [12] designed a new pretext task, *i.e.*, Rubik's cube recovery, as a supervisory signal to train the network to predict these transformations. The common idea of these works is to exploit internal structures of data and encourages the network to predict such structures. However, most previous works are focused on learning self-supervised features with single modality data, while none of them investigate the role of multi-modal data and how it could be utilized in self-supervised learning.

Fundus fluorescein angiography (FFA) is an imaging modality that can provide useful information regarding the retinal vasculature in the retina [14]. This information can help ophthalmologists better understand the structures of fundus lesions, microangioma, and capillary non-perfusion area, which are crucial for the diagnosis and treatment of some vitreoretinal diseases like AMD and PM [14]–[16]. The retinal vasculature information presented in FFA is complementary to color fundus images since FFA could identify fundus lesions that were not discovered by color fundus images [15]. Moreover, as shown in [14], compared to using color fundus photographs alone, there is a significant improvement in diagnostic sensitivity when using color fundus photographs with the corresponding FFA images. To utilize the mutual information in these two modalities, we propose to learn *the general feature representations for fundus disease classification via both color fundus and the corresponding FFA images.*

However, FFA is an invasive and time-consuming procedure, which is difficult to collect in many clinical sites. To the best of our knowledge, the Fundus-FFA dataset [9] is the only publicly available color fundus images with the corresponding FFA. Hence, we obtain the FFA modality through a generative

Manuscript received June 16, 2020; accepted July 8, 2020. Date of publication July 13, 2020; date of current version November 30, 2020. This work was supported in part by a Faculty Research Award from Google Inc. *(Corresponding author: Lei Xing.)*

The authors are with the Department of Radiation Oncology, Stanford University, Stanford, CA 94305 USA (e-mail: xmengli@stanford.edu; jeremy18@stanford.edu; tauhid@stanford.edu; lequany@stanford.edu; lei@stanford.edu).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2020.3008871

adversarial network on the public fundus-FFA dataset [9], such that our method is still applicable even when only color fundus images are available. Some examples of the color fundus and the corresponding synthesized FFA images are shown in Figure 1. Naively concatenating more datasets is a simple solution to utilize multiple modalities. However, as has been discovered in supervised learning, it is not an efficient way to use multiple modalities, which can even lead to reduced performance on the dataset of interest. We also observed this phenomenon for self-supervised learning, as results in Figure 5(a) in Section IV. By enlarging the dataset with corresponding FFA images, self-supervised learning performance decreases by around 6%. This fact suggests that the presence of domain difference can affect the performance and cross-modality relationship has to be considered.

To address this issue, in this paper, we formulate a novel patient feature-based softmax embedding to learn general feature representation from multi-modal data by learning the *positive concentrated* and *negative separated* properties. *The positive concentrated property* refers to learn **transformation-invariant and modality-invariant features** for individual patients. This is motivated by the fact that our downstream task is disease classification and a patient's disease classification result would not change due to image transformations, thus the expected feature representation should be invariant to transformations. Similarly, a patient's two modalities, *i.e.*, a color fundus image and the corresponding FFA image, should share the same semantic meaning, thus their feature representations should be coherent. Hence, we propose to mine the shared cross-modality information by learning modality-invariant features. *The negative separated property* refers to learn **patient-similarity features** by separating patients from each other. This is based on the observation that class-wise supervised learning can retain apparent similarity among classes in the representation space. For an image from a class leopard, the classes that get the highest responses from a trained neural net classifier are all visually correlated, e.g., jaguar and cheetah. It is not the semantic labeling, but the apparent similarity in the data themselves that brings some classes closer than others. Hence, if we treat each patient as a class and learn to separate him/her from others, we may end up with a representation that captures apparent similarity among patients. With these constraints as our learning objective, the network encodes both *modality-invariant* features and *patient-similarity* features into high-level representations, which can capture the semantically shared information across different modalities and apparent visual similarity between patients; see demonstrations in Figure 5 and Figure 7. A "patient" is a triplet, consisting of color fundus, transformed fundus, and FFA images, where all these images are obtained from the same patient. "patient feature-based" denotes that our loss function is calculated directly based on the patients' features. Hence, we name our method as "patient feature-based softmax embedding."

To demonstrate the effectiveness of our method, we employed two public fundus image datasets for disease classification, *i.e.*, Ichallenge-AMD dataset [17], and Ichallenge-PM dataset [18]. Given that the proposed self-supervised method does not use any label information, a direct comparison between our method and the state-of-the-art retinal disease diagnosis methods might not be fair, but extensive experiments still demonstrate the superiority of our method against the state-of-the-art self-supervised methods [19]–[21] on two retinal disease datasets. Notably, our method also achieves comparable performance with the supervised baseline.

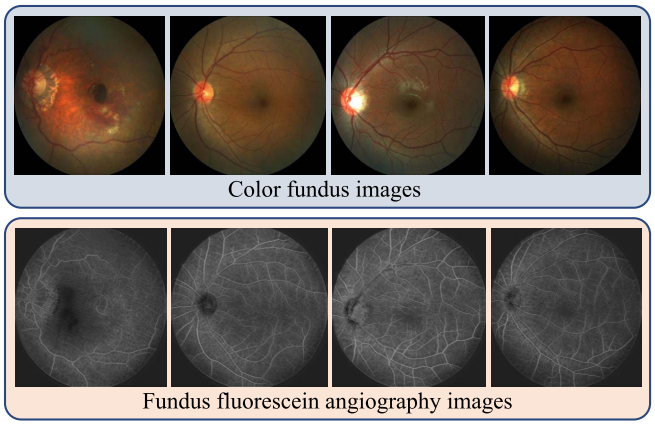


Fig. 1. Examples of color fundus images and the corresponding synthesized FFA images. Fundus images are selected from the Ichallenge-AMD dataset [26], and FFA images are synthesized by training a CycleGAN [8] on a public Fundus-FFA dataset dataset [9], and then tested on the unseen fundus images. Our goal is to perform self-supervised learning using data from these two related modalities.

The main contributions of this paper are:

- We present a novel self-supervised learning method by effectively exploiting multi-modal data for retinal disease diagnosis. Our method contains a network to synthesize another modality, thus it is still applicable even though only color fundus images are available. To the best of our knowledge, this is the first work for self-supervised disease diagnosis from fundus images.
- We formulate the patient feature-based softmax embedding as a self-supervised signal to capture the mutual information across modalities and patient-similarity features from multi-modal data, which learns effective representation for fundus image classification.
- Extensive experiments on two common eye diseases, *i.e.*, AMD and PM, demonstrate the superiority of our method than other state-of-the-art self-supervised methods. Our method also achieves comparable results with the supervised baseline.[1]

II. Related Works

In this section, we mainly review automatic disease diagnosis works from fundus photography and literatures related to self-supervised feature learning.

A. Automatic Disease Diagnosis From Fundus Photography

Many conventional ophthalmic diseases can be examined from fundus photography, such as age-related macular

[1] Code is available at https://github.com/xmengli999/self_supervised

degeneration (AMD), diabetic retinopathy (DR), glaucoma, and pathological myopia (PM). With the advances of deep learning, considerable efforts have been devoted to developing convolutional neural networks for automatic eye disease recognitions [4], [6], [22]–[28]. As for AMD diagnosis, [23] employed a CNN that is pre-trained on OverFeat features to perform AMD classification from fundus images. Reference [27] ensembled several convolutional neural networks to classify AMD diseases into 13 classes. Reference [6] developed a DeepSeeNet based on an Inception-v3 architecture [29] to identify patient-level AMD severity, by first detecting individual risk factors and then combining values from both eyes to assign a severity result. As for PM classification, [28] employed Xception [30] as the baseline architecture with ImageNet pre-train weights to diagnose PM from fundus images. However, most previous works on disease diagnosis from fundus photography are based on supervised learning, which requires a massive amount of labeled data. Different from previous works, in this paper, we focus on developing the self-supervised method for fundus disease diagnosis.

B. Self-Supervised Feature Learning

Self-supervised feature learning is becoming a popular topic and has been studied in several medical image recognition tasks. The common principle of existing works is to construct different pretext tasks by discovering supervisory signals directly from the input data itself, and then training the deep network to predict this supervisory information, such that the high-level representation of the input is learned. Notable pretext tasks include Rubik's cube recovery [12], anatomical position prediction [11], reconstructing part of the image like image completion [31], [32], 3D distance prediction [33], image-intrinsic spatial offset prediction [34]. For example, [12] proposed Rubik's cube recovery task for brain hemorrhage classification and tumor segmentation from CT and MR images, respectively. Reference [11] proposed to learn self-supervised features by predicting anatomical positions for cardiac MR image segmentation. Reference [34] designed image-intrinsic spatial offset relations task to learn self-supervised features. [33] introduced predicting 3D distance between two patches sampled from the same brain as a pretext task.

Recently, instance discrimination [19], [21], [35], an effective pretext task, achieves promising results on unsupervised feature representation learning. For example, [36] proposed to use softmax embedding with classifier weights to calculate the feature similarity, however, it prevents explicitly comparison over features, which results in limited efficiency and discriminability. Reference [19] developed memory bank to memorizes features of each instance. Reference [21] calculated the positive concentrated property based on the "real" instance feature, instead of classifier weights [36] or memory bank [19]. However, this method treated the optimization as the binary classification problem via maximum likelihood estimation, which is infeasible to learn the feature embedding from multi-modal data.

Unlike the previous works that explored self-supervised learning from the single modality data, we present to effectively exploit multi-modal data to improve self-supervised feature learning. Multi-modality data has been widely utilized for many medical image recognition tasks and there are several related works [37]–[42]. For example, [38] proposed to identify Alzheimer's disease (AD) - relevant biomarkers by learning from two modalities, *i.e.*, genetic information and brain scans. Reference [39] developed a hybrid fusion network for multi-modal MR image synthesis. Reference [40] proposed a latent representation learning method for multi-modality (*i.e.*, PET and MRI) based AD diagnosis. However, all these works are supervised learning or image synthesis methods, while we are investigating an unsupervised learning strategy for disease classification.

From this perspective, in this work, we employ the instance discrimination [19], [21] as the pretext task, and propose to learn features by learning both modality-invariant and patient-similarity features from multi-modal data.

III. Method

A. Overview

Figure 2 depicts the workflow of our self-supervised method for retinal disease diagnosis. We first train a GAN model on the Fundus-FFA dataset [9] to learn the mapping function between the color fundus and FFA, and then obtain the synthesized FFA modality on the Ichallenge-AMD and Ichallenge-PM dataset. Secondly, to learn the self-supervised features, we randomly sample n triplets, and each triplet is derived from each patient, consisting of color fundus image, the transformed image, and the corresponding FFA. The triplets are fed into the neural network to learn the high-level feature representations, which are optimized by the proposed patient feature-based softmax embedding objective. Our learning objective encourages the network to learn transformation- and modality-invariant features, while also capture the patient-similarity features. Finally, we evaluate the network on unseen fundus images, following the standard evaluation protocol in most self-supervised works [19], [35]. The final classification result is obtained by applying a K-Nearest Neighbor (KNN) classifier. Below, we will elaborate on the FFA image synthesization, patient feature-based softmax embedding, and technical details.

B. FFA Image Synthesization

FFA is invasive and it is difficult to collect in many clinical sites [43]. Hence, we propose to synthesize FFA images, such that our method can still be utilized to perform self-supervised learning even though only color fundus are available. Specifically, we train a generative model on the Fundus-FFA dataset [9] to learn the mapping function between the color fundus and FFA images, and then synthesize the corresponding FFA modality in the target fundus datasets, *i.e.*, Ichallenge-AMD and Ichallenge-PM dataset, to perform self-supervised feature learning.

The Fundus-FFA dataset [9] contains color fundus images with the corresponding FFA images, which is not pixel-aligned. Based on this consideration, we trained a CycleGAN

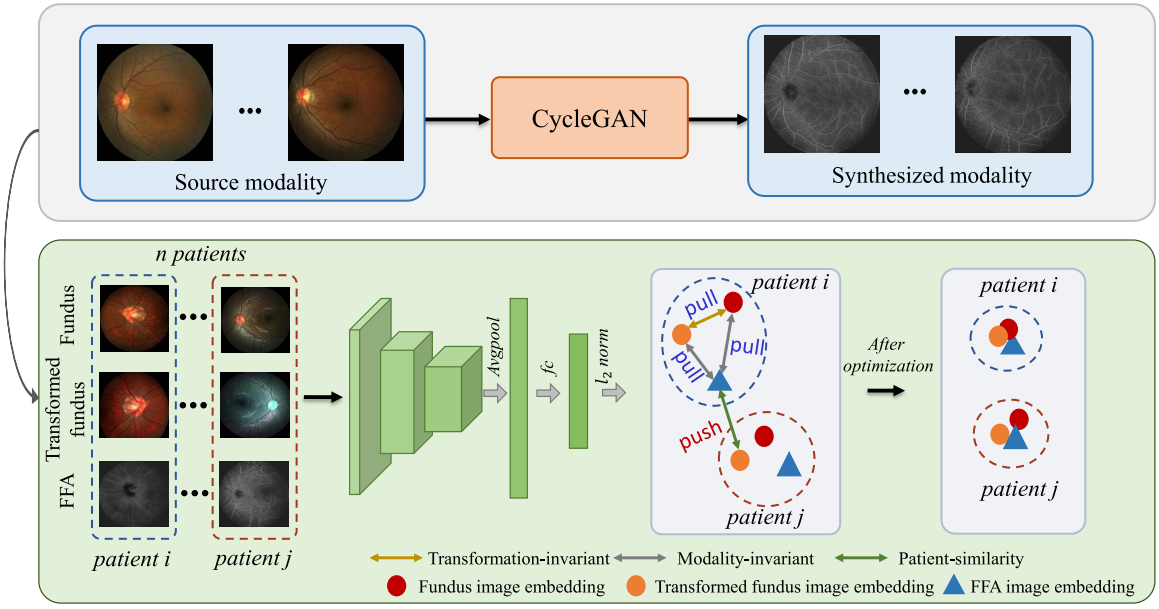


Fig. 2. The illustration of the proposed method. We first train a generative network (CycleGAN) on the fundus-FFA dataset to learn the mapping function between color fundus image and FFA, and then synthesize the corresponding FFA on target color fundus datasets. Secondly, triplets are derived from each patient, consisting of the randomly selected fundus image, transformed image, and the corresponding FFA. These triplets are fed into the neural network to learn the high-level representation with our proposed patient feature-based softmax embedding objective. Finally, the network is evaluated on unseen fundus images and the final classification result is obtained by applying a KNN on the features.

model [8] and we followed the original setting to train the network with both adversarial loss and cycle-consistency loss. To adapt the network to our task, we modified the learning rate to 0.0001 and trained it for 500 epochs. After network optimization, we tested the model on unseen fundus datasets, and the synthesized results can be seen in Figure 1 and Figure 6. Since there is no ground-truth FFA provided in these datasets, the synthesization quality is measured by running supervised learning for fundus image classification; see results in Table V.

C. Patient Feature-Based Softmax Embedding

Let $C = \{c_i\}$ and $S = \{s_i\}$, where c_i and s_i denote the color fundus image and the corresponding FFA image of patient i, respectively. Our goal is to learn a feature embedding network $f_\theta(\cdot)$ that maps an unlabeled image c_i or s_i to a low-dimensional feature embedding $f_\theta(c_i)$ or $f_\theta(s_i) \in \mathbb{R}^d$, where d is the feature dimension. For simplicity, we use $\mathbf{f}_i = f_\theta(c_i)$, $\mathbf{g}_i = f_\theta(s_i)$ to represent the feature of patient i from fundus and FFA, respectively. We normalize all the features by l_2 normalization, *i.e.*, $\|\mathbf{f}_i\|_2 = 1$, $\|\mathbf{g}_i\|_2 = 1$. Without the ground-truth category labels, we need to form self-supervised learning constraints to facilitate model optimization.

1) Transformation- and Modality-Invariant Features: A patient's disease diagnosis result would not change due to image transformations. Thus, a good feature embedding should satisfy that the representation of a color fundus image c_i of patient i should be invariant under random data augmentations. Intuitively, a color fundus image c_i from a patient i should share the same semantic meaning with the corresponding FFA image s_i, thus the representations of a patient should be coherent. Hence, the network requires to learn *transformation-* and *modality-invariant* features. To achieve this, we randomly sample n patients from the datasets, and each patient consists of both color fundus image and the corresponding synthesized FFA. The selected samples are denoted by $\{c_1, s_1, \cdots, c_n, s_n\}$. To learn the transformation-invariant features, a random data augmentation is applied to slightly modify the original fundus c_i to $\hat{c}_i$, and we can obtain the batch denoted by $\{c_1, \hat{c}_1, s_1, \cdots, c_n, \hat{c}_n, s_n\}$. These images are fed into the network to get high-level representations, *i.e.*, $\{\mathbf{f}_1, \hat{\mathbf{f}}_1, \mathbf{g}_1, \cdots, \mathbf{f}_n, \hat{\mathbf{f}}_n, \mathbf{g}_n\}$. The probability of $\hat{c}_i$ being recognized as patient i is defined as

$$P(i|\hat{c}_i) = \frac{\exp\left(\mathbf{f}_i^T \hat{\mathbf{f}}_i / \tau\right)}{\sum_{k=1}^{n} \exp\left(\mathbf{f}_k^T \hat{\mathbf{f}}_i / \tau\right)}, \tag{1}$$

where τ is the temperature parameter controlling the concentration level of the sample distribution [44]. The probability of s_i being recognized as patient i is defined by

$$P(i|s_i) = \frac{\exp\left(\mathbf{f}_i^T \mathbf{g}_i / \tau\right)}{\sum_{k=1}^{n} \exp\left(\mathbf{f}_k^T \mathbf{g}_i / \tau\right)}, \tag{2}$$

where $\mathbf{f}_i^T \hat{\mathbf{f}}_i$, $\mathbf{f}_i^T \mathbf{g}_i$ denote the cosine similarity between positive pairs, as shown in Figure 3.

2) Patient-Similarity Features: To learn patient-similarity features, we need to treat each patient as a class and learn to separate him/her from other patients. As shown in Figure 3, the distance between negative pairs should be enlarged in the representation space. The probability of c_j being recognized

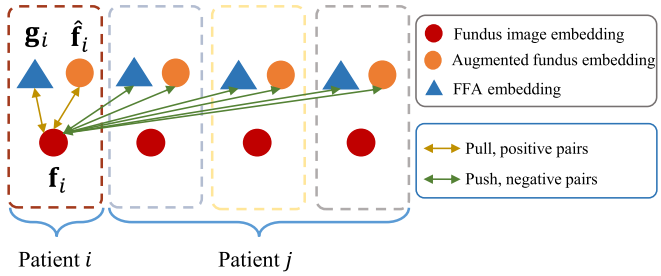


Fig. 3. The illustration of the proposed patient-based softmax feature objective.

as patient i is defined by

$$P(i|c_j) = \frac{\exp\left(\mathbf{f}_i^T \mathbf{f}_j/\tau\right)}{\sum_{k=1}^{n} \exp\left(\mathbf{f}_k^T \mathbf{f}_j/\tau\right)}, j \neq i. \quad (3)$$

This equation also holds for s_j. We assume different image samples being recognized as patient i are independent, the joint probability of $\hat{c}_i, s_i$ being recognized as patient i and c_j, s_j being not classified to patient i is

$$P_i = P(i|\hat{c}_i)P(i|s_i)\prod_{j\neq i}(1 - P(i|c_j))\prod_{j\neq i}(1 - P(i|s_j)), \quad (4)$$

3) Learning Objective: The above probability is optimized by the negative log likelihood, which is defined as

$$\mathcal{L}_i = -\log P(i|\hat{c}_i) - \log P(i|s_i) - \sum_{j\neq i}\log(1-P(i|c_j)) - \sum_{j\neq i}\log(1-P(i|s_j)) \quad (5)$$

The final loss function is defined by minimizing the mean of the negative log likelihood over all patients n within the batch. Our learning objective is defined as

$$\mathcal{L} = \frac{1}{n}\sum_i \mathcal{L}_i. \quad (6)$$

Hence, we learn the *modality-invariant* and *patient-similarity* features by simultaneously doing positive concentration and negative separation. Since the loss function is calculated on the feature of the patient, we name it as *patient feature-based softmax embedding*.

D. Technical Details

1) Network Architecture: Our framework is based on the ResNet18 backbone [45], following the same setting as the previous works [19], [21]. We apply an average pooling on the output of the last residual block in ResNet18. Then, the feature is flattened to a vector and a fully connected layer, a batch normalization layer, and a ReLU are sequentially applied to reduce the feature dimension to 128. Finally, the feature is normalized by l_2 normalization to the embedding space. The proposed patient feature-based softmax embedding loss function is utilized to train the neural network.

2) Implementation Details: The whole framework is built on PyTorch [46] with an NVIDIA Tesla V100 32GB GPU. We resize images to 320×320 resolution. For data augmentation, we randomly scale and crop images into the patches of size 224×224, with a random scaling factor chosen from [0.2, 1.0]. Our algorithm performs randomly horizontal flip and has a probability of 0.2 to randomly grayscale the input. The algorithm also randomly blends the image to some extent with its black version, grayscale version. This operation changes the brightness, contrast, and saturation of the input image with a random factor chosen uniformly from [0.6, 1.4], following the setting in [19], [21]. Note that data augmentation is also applied in each image in the triplet to enrich the training samples. For implementation, each image has two positive samples and $2n - 2$ negative samples to compute Eq. (6), where n is the number of sampled patients and τ is set to 0.1. In each feed forward, we sample 75 patients, *i.e.*, $n = 75$. The network is optimized with Adam optimizer [47]. The initial learning rate is set to 0.0001 and is dropped by a factor of 0.1 every 1000 epochs. All the experiments are equally trained for 2000 epochs and the reported results are conducted on 5-fold cross-validation.

3) Evaluation Protocol: We verify our method by applying a KNN classifier on frozen features, following a common protocol [19], [21]. To investigate the transfer learning capacity, we unfreeze the features and train a supervised linear classifier (a fully-connected layer followed by softmax) on the target datasets.

IV. Experiments

A. Datasets

We employ two public retinal disease datasets, *i.e.*, Ichallenge-AMD [2] (task 1) and Ichallenge-PM [3] (task 1), and evaluate the effectiveness of our method by performing normal and abnormal fundus image classification.

1) Ichallenge-AMD Dataset: Ichallenge-AMD dataset [17] contains 1200 annotated retinal fundus images, including both non-AMD subjects (77%) and AMD patients (23%). Typical signs of AMD that can be found in these photos include drusen, exudation, hemorrhage, etc. Since only training data is released with annotations, we use the training data in the Ichallenge-AMD dataset and perform 5-fold cross-validation.

2) Ichallenge-PM Dataset: Ichallenge-PM dataset [18] contains 1200 annotated color fundus photos with Non-PM (50%) and PM cases (50%). All the photos were captured with Zeiss Visucam 500. We use the training data in the Ichallenge-PM dataset and perform 5-fold cross-validation. In these two datasets, the image-level annotation is provided, where 0 denotes normal and 1 denotes abnormal cases. However, we do not utilize any human-annotated labels information during network training. To evaluate the classification accuracy of our method, we employ AUC, accuracy, precision, recall, and F1-score as the evaluation metrics.

[2] https://ichallenges.grand-challenge.org/iChallenge-AMD/
[3] https://ichallenges.grand-challenge.org/iChallenge-PM/

3) EyePACS Dataset: To evaluate the transfer learning capacity of our model, we train the self-supervised model on the Kaggle's Diabetic Retinopathy Detection Challenge (EyePACS) dataset[4] and report the classification result on the AMD dataset. This dataset is sponsored by the California Healthcare Foundation. It provides a totally 88,702 images, captured under various conditions and various devices. The Left and right fields are provided for every subject, and an ophthalmologist rated the presence of diabetic retinopathy in each image on a scale of 0 to 4. We use all the images in this dataset to train our self-supervised model. Note that we did not utilize any human-annotated labels in this dataset.

4) Fundus-FFA Dataset: To the best of our knowledge, Fundus-FFA dataset [9] is the only publicly available dataset that contains color fundus images and corresponding FFA images. It has 30 healthy persons and 29 patients with diabetic retinopathy. Each patient has a color fundus photo and corresponding FFA. The dataset is very limited and is not suitable to train an unsupervised model. As an alternative, we train a generative model on the Fundus-FFA dataset to learn the mapping function between the color fundus images and the corresponding FFA images [8], [48]–[50].

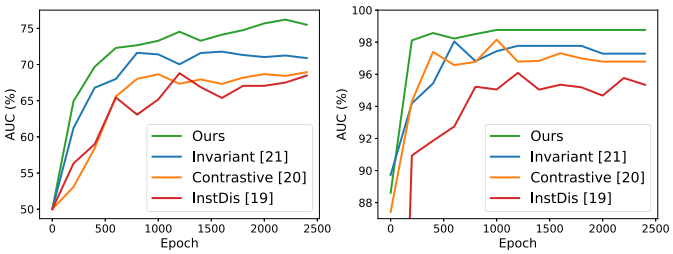
B. Comparison on the Ichallenge-AMD Dataset

To show the effectiveness of our method, we compare it with state-of-the-art self-supervised learning methods on the Ichallenge-AMD dataset.

1) Experimental Settings: To have a fair comparison, all of the models are trained on the ResNet18 backbone [45] with 5-fold cross-validation. In the *Supervised* baseline, we modified the output channel of the original fully connected layer of ResNet18 to 2 for two-class classification. The supervised model is trained by the cross-entropy loss with human-annotated labels. To compare with self-supervised methods, unlike other self-supervised methods that learn the 2D or 3D correspondences by predicting rotation or anatomical positions [11], [12], our fundus image classification task is invariant to the image transformation. To the best of our knowledge, there are no related self-supervised methods that learn self-supervised transformation-invariant features in the medical imaging domain, we compare with several state-of-the-art instance discrimination methods in the computer vision domain [19]–[21]. Finally, we perform KNN on all the unsupervised feature learning methods to evaluate the feature performance for classification and $k = 100$. Note that we run these methods with the same backbone, learning strategies and trained all the models for 2000 epochs on 5-fold cross-validation.

2) Results: The results of different unsupervised methods are shown in Table I. It is observed that *Contrastive* [20] and *Invariant* [21] achieve better results than *InstDisc* [19], showing that contrastive learning can be beneficial to unsupervised feature learning. From the comparison, we can see that *Invariant* [21] performs slightly better than *Contrastive* [20]. This is because the heavy data augmentation proposed in *Contrastive* [20] would hurt the performance in our fundus image classification task. It is also observed in Table I that our method excels all other unsupervised feature learning methods by at least around 3.16% on AUC, which demonstrates the effectiveness of our method in the unsupervised feature learning. Figure 4(a) visualizes the learning curve of the validation results and we can see our method consistently outperforms other methods. Notably, without any annotation during training, our method is approaching to the supervised learning baseline, *e.g.*, 74.58% vs 77.19% on AUC. The results further demonstrate the effectiveness of our self-supervised learned features.

[4]https://www.kaggle.com/c/diabetic-retinopathy-detection/data



(a) Ichallenge-AMD dataset (b) Ichallenge-PM dataset

Fig. 4. Comparison of AUC results on the (a) Ichallenge-AMD dataset and (b) Ichallenge-PM dataset.

TABLE I
RESULTS ON THE ICHALLENGE-AMD DATASET (UNIT: %)

	AUC	Accuracy	Precision	Recall	F1-score
Supervised	77.19	87.09	82.98	77.82	79.27
InstDisc [19]	66.49	82.02	74.63	66.49	68.69
Contrastive [20]	68.06	82.45	73.48	68.06	69.84
Invariant [21]	71.42	84.31	77.99	71.42	73.67
Ours	**74.58**	**86.58**	**83.20**	**74.58**	**77.33**

TABLE II
RESULTS ON THE ICHALLENGE-PM DATASET (UNIT: %)

	AUC	Accuracy	Precision	Recall	F1-score
Supervised	98.04	97.66	97.30	98.04	97.53
InstDis [19]	95.49	95.32	95.04	95.49	95.18
Contrastive [20]	96.98	96.94	96.67	96.98	96.68
Invariant [21]	97.26	97.30	97.11	97.26	97.16
Ours	**98.55**	**98.65**	**98.60**	**98.55**	**98.57**

C. Comparison on the Ichallenge-PM Dataset

We also compare our method with the other unsupervised feature learning methods on the Ichallenge-PM dataset. In this dataset, we use the same experimental settings as those in the Ichallenge-AMD dataset. Table II summarizes the results of different methods on the Ichallenge-PM. From Table II we can see that our method excels other methods on all five metrics. In particular, our method outperforms the state-of-the-art method *Invariant* [21] by 1.29% on AUC. It is observed that the results keep consistent with those on the Ichallenge-AMD dataset, showing the effectiveness and generalization of our method. The validation results during learning are visualized in Figure 4(b). We can see that our method consistently surpasses

TABLE III
RESULTS OBTAINED BY FIRST TRAINING A SELF-SUPERVISED MODEL ON THE EYEPACS DATASET AND THEN FINE-TUNING ON THE FOLLOWING TWO DATASETS. *Random init* DENOTES THE NETWORK IS TRAINED WITH RANDOMLY WEIGHT INITIALIZATION (UNIT: %)

Method \ Dataset	Ichallenge-AMD					Ichallenge-PM				
	AUC	Accuracy	Precision	Recall	F1-score	AUC	Accuracy	Precision	Recall	F1-score
Random init	77.19	87.09	82.98	77.82	79.27	98.04	97.66	97.30	98.04	97.53
Invariant [21]	81.62	87.51	81.92	81.62	81.35	98.02	97.84	97.56	98.02	97.75
Ours	**83.17**	**89.37**	**85.71**	**83.17**	**83.67**	**98.41**	**98.38**	**98.31**	**98.41**	**98.33**

other methods. It is worth mentioning that our method achieves higher results than the supervised upper bound in this dataset, which further demonstrates the effectiveness of our method.

D. Comparison on Generalization Capacity

To demonstrate the generalizable features, we show the transfer learning results of our method. We pre-train the self-supervised model on the EyePACS dataset and fine-tune the model on the Ichallenge-AMD and Ichallenge-PM dataset, respectively. In the pre-training stage, we do not utilize any labels while the labels are required in the fine-tuning stage. Our goal is to investigate whether our self-supervised method can learn more generalizable or transferable features that can be easily transferred to other tasks. We compared with the state-of-the-art self-supervised method [21]. During the pre-training stage, we trained all the self-supervised methods until convergences (around 200 epochs) on the EyePACS dataset. Then, the learned network weight is employed as the network initialization and is fine-tuned on the Ichallenge-AMD and Ichallenge-PM dataset, respectively. During the fine-tuning stage, all the models are trained with the same learning strategy and data augmentation, and the only difference is the network initialization.

Table III lists the transfer learning results on the Ichallenge-AMD and Ichallenge-PM dataset. *Random init* denotes the network is trained with randomly weight initialization. From Table III we can see that our method consistently outperforms the state-of-the-art self-supervised method [21] on two benchmark datasets. As for the Ichallenge-AMD dataset, it is observed that our method can achieve around 6.0% and 1.5% improvement on AUC over *Random init* and *Invariant* [21], respectively. Similarly, our result on the Ichallenge-PM dataset also excels *Random init* and *Invariant*. These consistent results demonstrate the excellent transfer learning capacity of our method.

E. Analytical Studies

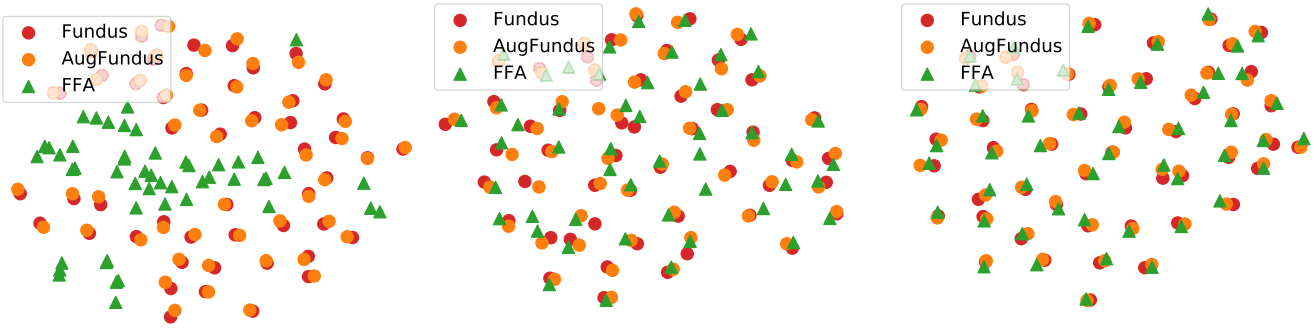
1) Comparison to Other Alternatives: To show the effectiveness of our method, we compare the following variants: **Enlarged-Data**: Train a self-supervised model with the method [21], where the multi-modal data is used by simply enlarging the dataset. For example, the enlarged dataset has $2n$ samples, where n is original color fundus and n is corresponding FFA. All $2n$ samples are used as the training images. **As-Augmentation**: Train a self-supervised model on the instance discrimination task [21] by adding the synthesized modality data as an augmentation. **Ours**: Train a self-supervised model on multi-modal data with constraint in Eq. (6).

TABLE IV
ABLATION STUDY ON THE ICHALLENGE-AMD DATASET. (A) ENLARGED-DATA: TRAIN A SELF-SUPERVISED MODEL WITH MULTI-MODAL DATA BY SIMPLY ENLARGING THE DATASET. (B) AS-AUGMENTATION: TRAIN A SELF-SUPERVISED MODEL ON THE INSTANCE DISCRIMINATION TASK BY ADDING THE SYNTHESIZED MODALITY DATA AS AN AUGMENTATION. (C) OURS: TRAIN A SELF-SUPERVISED MODEL ON MULTI-MODAL DATA WITH CONSTRAINT IN EQ. (6). (UNIT: %)

	AUC	Accuracy	Precision	Recall	F1-score
(a) Enlarged-Data	65.72	80.93	70.97	65.72	67.35
(b) As-Augmentation	70.93	83.21	77.78	70.93	72.65
(c) **Ours**	**74.58**	**86.58**	**83.20**	**74.58**	**77.33**

Figure 5 visualizes the learned feature embedding of three variants. We randomly sample 50 color fundus images from the Ichallenge-AMD dataset with corresponding synthesized FFA. These images along with the randomly augmented samples are fed into the network to get feature representations, followed by reducing the feature dimension to 2 by t-SNE [51]. The closer the fundus and the augmented fundus image embedding, the better transformation-invariant feature is learned. Similarly, the closer the fundus and FFA image embedding, the better modality-invariant feature is learned. We can see from Figure 5(a) that *Enlarged-Data* achieves inferior performance and there is no apparent relationship between color fundus and FFA images. This is because both the fundus image and FFA learn the transformation-invariant features on its own, and the cross-modality information is neglected by the network. *As-Augmentation* can close the distance between the color fundus and the corresponding FFA images (the red circle and the green rectangle in Figure 5(b)), but the performance is still inferior. It is observed from Figure 5(c) that our method can minimize the distance among fundus, transformed image, and FFA image, and at the same time enlarge the distance among different patients. Results for each variant on the Ichallenge-AMD dataset are summarized in Table IV. We can see our method can outperform *Enlarged-Data* and *As-Augmentation* on all five metrics. In particular, our result surpasses *Enlarged-Data* and *As-Augmentation* by around 8.86% and 3.6% on AUC, respectively. These comparisons show that the modality-invariant constraint on multi-modal data is very useful and can contribute to better feature representation.

2) Visualization of Patient-Similarity and Modality-Invariant Features: Figure 5 shows the learned feature embedding



(a) Enlarged-Data: AUC 65.72% (b) As-Augmentation: AUC 70.93% (c) Ours: AUC 74.58%

Fig. 5. Visualization of the learned feature embedding of three variants (see definitions in Table IV). We randomly sample 50 color fundus images from the Ichallenge-AMD dataset with corresponding synthesized FFA. These paired two modalities along with the randomly augmented fundus samples are fed into the network to get feature representations, followed by reducing the feature dimension to 2 by t-SNE [51]. The closer the fundus (red circle) and the augmented image (orange circle) embedding, the better transformation-invariant features are learned. The closer the fundus (red circle) and the corresponding FFA (green rectangle) embedding, the better modality-invariant features are learned. The detailed results for each variant are listed in Table IV. Best viewed in color.

TABLE V
ANALYSIS OF SUPERVISED LEARNING ON THE ICHALLENGE-AMD DATASET. "FUNDUS" DENOTES THE COLOR FUNDUS AND "SYN FFA" DENOTES THE SYNTHESIZED FFA (UNIT: %)

Modality	Backbone	AUC	Accuracy	Precision	Recall	F1-score
Fundus	ResNet18	77.19	87.09	82.98	77.82	79.27
Syn FFA		77.05	83.97	76.44	77.05	76.28
Fundus	ResNet34	77.79	86.83	82.78	77.79	79.06
Syn FFA		77.35	84.30	77.48	77.35	77.06
Fundus	ResNet50	75.21	84.89	78.80	75.21	76.00
Syn FFA		75.69	84.98	78.94	75.69	76.33

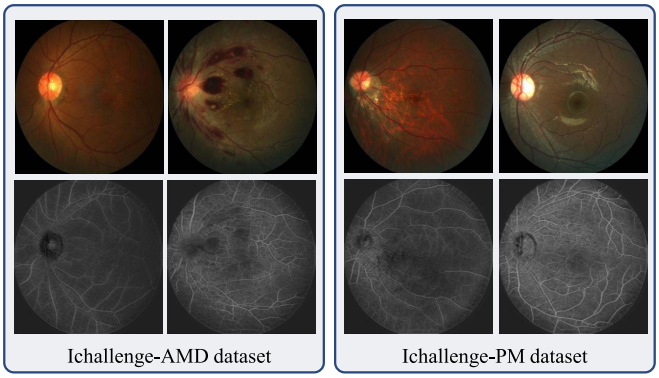


Fig. 6. Visualization of synthesized FFA on the Ichallenge-AMD and Ichallenge-PM dataset. The first row denotes the color fundus images and the second row is the synthesized FFA images through our trained CycleGAN model.

and we found that *color fundus* is very close to *FFA image* in the embedding space, demonstrating the learned modality-invariant features. We also show the similarity score in Figure 7(a), where the maximum similarity score is 1.0. We can observe that FFA achieves a high similarity score with the test image in the embedding space, which further demonstrated the modality-invariant features.

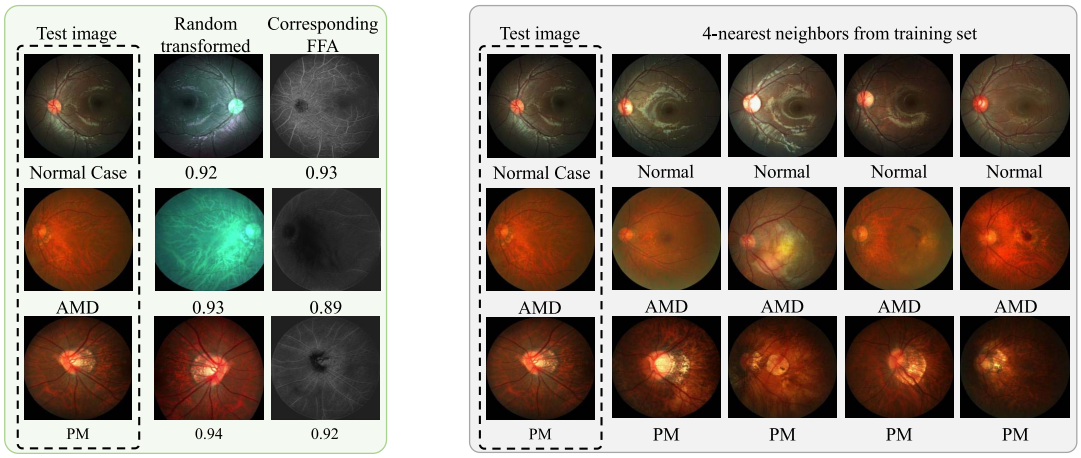
In Section I, we argue that learning to separate patients can learn apparent visual similarity among patients, *i.e.*, patient-similarity features. To demonstrate the patient-similarity features, we visualize the K-Nearest Neighbors. In Figure 7(b), for one test image, we visualize 4-nearest neighbors with a ground-truth label shown below each figure. We found that these neighboring images are very similar to the test image, demonstrating that our method can capture the similarity among patients, *i.e.*, patient-similarity features.

3) Evaluation on the Synthesis Quality: Since there are no ground-truth FFA images presented in both Ichallenge-AMD and Ichallenge-PM datasets, we evaluated the quality of FFA images by running conventional supervised learning experiments with synthesized FFA images. Specifically, we train the supervised baseline with the synthesized FFA images and all the training strategies are the same with those trained for color fundus images. We showed the highest supervised learning results in Table V. It can be observed that our synthesized FFA images can achieve similar classification results, compared to the color fundus images under three different backbones, including ResNet18, ResNet34, and ResNet50. The results demonstrate that the synthesized FFA images can obtain satisfactory quality to perform disease classification tasks. We also visualize the synthesized FFA images in Figure 6. We can see that the retinal vasculature can be observed in the FFA images.

TABLE VI
THE EFFECTS OF THE SYNTHESIZED IMAGES ON OUR METHOD. "SYN1" AND "SYN2" DENOTE THAT OUR METHOD IS TRAINED WITH SYNTHESIZED FFA GENERATED AT EPOCH 450 AND 500, RESPECTIVELY (UNIT: %)

	AUC	Accuracy	Precision	Recall	F1-score
Syn1	74.45	85.06	79.59	74.45	76.31
Syn2	74.58	86.58	83.20	74.58	77.33

To analyze the effects of the synthesized FFA images on our method, we run our method with different synthesized images generated at different models (saved at 450 and 500 epochs



(a) Positive concentration within the patient

(b) Negative separation between patients

Fig. 7. (a) Given the test images with the normal case, AMD, and PM, we first perform the random transformation to obtain the transferred samples, and use the GAN to generate the corresponding FFA images for each test image. Note that both the random transformed images and the corresponding FFA images have high similarity scores with the test sample (numbers below each figure), which indicates the positive concentration in Eq. (1) and (2) can learn the *transformation-* and *modality-invariant* features. (b) We retrieve 4-nearest neighbors from training set for each test images based on the similarity scores through KNN algorithm. The retrieved images have high visual similarity with the test sample, identifying the *negative separation* in Eq. (3) can capture *visual similarity among patients*.

TABLE VII

ABLATION STUDY ON OUR METHOD. THE FIRST ROW DENOTES THAT THE INPUT IS A DOUBLET, CONSISTING OF "COLOR FUNDUS AND TRANSFORMED FUNDUS." THE SECOND ROW DENOTES THAT THE INPUT IS A DOUBLET, CONSISTING OF "COLOR FUNDUS AND CORRESPONDING FFA." THE SECOND ROW DENOTES THAT THE INPUT IS A TRIPLET, CONSISTING OF "COLOR FUNDUS, TRANSFORMED FUNDUS AND CORRESPONDING FFA" (UNIT: %)

Positive		Negative	AUC	Accuracy	Precision	Recall	F1-score
Transformation-invariant	Modality-invariant	Patient-similarity					
✓		✓	71.42	84.31	77.99	71.42	73.67
	✓	✓	70.48	83.80	76.90	70.48	72.68
✓	✓	✓	**74.58**	**86.58**	**83.20**	**74.58**	**77.33**

respectively). The results are shown in Table VI. It is observed that the disease classification results are very similar, which indicates that our method is robust to the synthesized images generated by the trained CycleGAN network.

4) Ablation Study on Our Method: Our method is a contrastive loss function performed on the triplets to optimize "positive pairs" and "negative pairs." This function optimizes the joint probability of positive pairs and negative pairs; see Eq.(4). "positive pairs" are the same as the "correct prediction" that should be minimized while "negative pairs" are similar to the "wrong prediction" that should be maximized. Hence, the optimization must be done with at least one positive and one negative pair.

Since our method has two positive pairs and one negative pair, we conduct ablation study on different combinations of positive and negative pairs, and results are shown in Table VII. The positive pairs denote that the features should be pulled together, where transformation-invariant and modality-invariant features can be learned. The negative pairs denote that the features should be separated away, where patient-similarity features can be learned. It is observed that our method achieves the best performance when learning both transformation-invariant and modality-invariant features.

5) Analysis on Modality-Specific Features: In Eq (2), we encourage the network to learn modality-invariant features.

TABLE VIII

ABLATION STUDY ON MODALITY-SPECIFIC FEATURES (UNIT: %)

	AUC	Accuracy	Precision	Recall	F1-score
multi-task [52]	67.20	79.66	70.08	67.20	68.06
margin0.2	73.54	83.88	78.14	73.54	74.71
margin0.1	74.82	84.81	78.65	74.82	75.97
Ours	**74.58**	**86.58**	**83.20**	**74.58**	**77.33**

However, it is widely known that different modalities present modality-specific information. In this section, we investigate the effectiveness of preserving modality-specific information for unsupervised representation learning. We implemented two variants to persevere the modality-specific information. In the first variant, we followed multi-domain self-supervised learning work [52] and implemented an auxiliary classification branch to differentiate the modalities. We namely this experiment as *multi-task* and the result is shown in Table VIII. However, we found that this multi-task approach would hurt the performance. In the second variant, we added a margin in the numerator in Eq. (2) to control the concentration of modality-invariant features. A large margin denotes less concentration on positive pairs. However, as results in Table VIII, we found there is no apparent performance improvement by learning modality-specific features.

6) Statistic Analysis: To provide the statistic analysis on our method, we perform the independent t-tests on Invariant [21] and our method on the Ichallenge-AMD dataset. We run each experiment three times with randomly initialized seed. Through the t-test, p-value is 0.00064, which is significantly smaller than 0.05. The result indicates strong evidence that there is more than a 95% probability that our method is statistically better than Invariant [21].

V. Discussion

Recently, with the advances of deep learning techniques, automatic retinal disease diagnosis have been well studied in the research community, such as AMD classification [6], [23], [27], [53], DR grading [4], [5], [54], [55] and PM classification [18], etc. Although satisfactory results were achieved on these tasks, these methods require a large amount of labeled data which are difficult and expensive to obtain. In this work, we propose a self-supervised method for retinal disease diagnosis via effectively exploiting multi-modal data. We formulate a patient feature-based softmax embedding learning objective, where modality-invariant features and patient-similarity features are learned. Our method is validated on two public retinal disease datasets, *i.e.*, Ichallenge-AMD and Ichallenge-PM challenge, in which our method consistently outperforms other self-supervised methods and is comparable with the supervised baseline. Our method also surpasses other methods in terms of transfer learning, showing the effectiveness of our method in learning generalizable and transferable features.

Although our method achieves excellent performance, it comes with limitations. Since the number of fundus-FFA images is limited, we compromise to develop a multi-modal self-supervised model by synthesizing the FFA images. In reality, FFA images would provide more information about microaneurysms and hemorrhage, which would be beneficial for the disease diagnosis, such as AMD and PM [14]–[16]. One solution is to collect the datasets with color fundus and corresponding FFA images. Another limitation of our method is that in this paper we focus on unsupervised feature learning. We follow the standard evaluation protocol in most self-supervised and unsupervised learning works [35], [56], [57], where the feature learning stage is unsupervised and the label information is required in the final classifiers, such as KNN or fully connected layer. To make the whole diagnosis process unsupervised, one solution is to investigate joint learning of feature embedding and estimation of cluster assignments (or labels). In particular, we will consider to connect the feature learning with the soft and regularized deep K-means algorithm [58].

The future direction we would like to work on is to better model the mutual information between multi-modal data. Another potential research direction is to extend our method to more multi-modal medical imaging applications, such as multi-modal MRI, CT-MRI recognition tasks, etc. Even though only one modality is available in some cases, we can synthesize another modality through adversarial learning. Through this, we hope to leverage the general feature representation to improve a lot of downstream tasks, such as segmentation, classification, and detection [31], [59]. Also, it might bring some new insights to computer-aided diagnosis in an unsupervised way.

VI. Conclusion

This paper presents a novel self-supervised learning method by effectively exploiting multi-modal data for disease diagnosis from fundus images. Our key idea is to jointly utilize two modalities, *i.e.*, color fundus, and FFA, to learn better feature representation. Our proposed patient feature-based softmax embedding can achieve this goal by learning modality-invariant features and patient-similarity features, which show effective for fundus disease classification. Experimental results on two public datasets demonstrate that our method outperforms other self-supervised methods and achieves comparable performance to the supervised baseline. We also show the excellent performance of our method in learning generalizable features.

References

[1] Age-Related Eye Disease Study Research Group, "Risk factors associated with age-related macular degeneration. A Case-control study in the age-related eye disease study: Age-related eye disease study report number 3," *Ophthalmology*, vol. 107, no. 12, pp. 2224–2232, 2000.

[2] I. G. Morgan, K. Ohno-Matsui, and S.-M. Saw, "Myopia," *Lancet*, vol. 379, no. 9827, pp. 1739–1748, 2012.

[3] K. Zhou *et al.*, "Multi-cell multi-task convolutional neural networks for diabetic retinopathy grading," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 2724–2727.

[4] X. Li, X. Hu, L. Yu, L. Zhu, C.-W. Fu, and P.-A. Heng, "CANet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1483–1493, 2019.

[5] A. Sakaguchi, R. Wu, and S.-I. Kamata, "Fundus image classification for diabetic retinopathy using disease severity grading," in *Proc. Int. Conf. Biomed. Eng. Technol.*, 2019, pp. 190–196.

[6] Y. Peng *et al.*, "DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs," *Ophthalmology*, vol. 126, no. 4, pp. 565–575, Apr. 2019.

[7] J. Virmani *et al.*, "PNN-based classification of retinal diseases using fundus images," in *Sensors for Health Monitoring*. Amsterdam, The Netherlands: Elsevier, 2019, pp. 215–242.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017, pp. 2223–2232.

[9] S. H. M. Alipour, H. Rabbani, and M. R. Akhlaghi, "Diabetic retinopathy grading by digital curvelet transform," *Comput. Math. Methods Med.*, vol. 2012, Sep. 2012, Art. no. 761901.

[10] A. Jamaludin, T. Kadir, and A. Zisserman, "Self-supervised learning for spinal MRIs," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 294–302.

[11] W. Bai *et al.*, "Self-supervised learning for cardiac MR image segmentation by anatomical position prediction," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Springer, 2019, pp. 541–549.

[12] X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng, "Self-supervised feature learning for 3D medical images by playing a Rubik's cube," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Springer, 2019, pp. 420–428.

[13] N. Tajbakhsh *et al.*, "Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data," in *Proc. ISBI*, Apr. 2019, pp. 1251–1255.

[14] S. N. Patel *et al.*, "Color fundus photography versus fluorescein angiography in identification of the macular center and zone in retinopathy of prematurity," *Amer. J. Ophthalmol.*, vol. 159, no. 5, pp. 950–957, 2015.

[15] S. Wang, Y. Zuo, N. Wang, and B. Tong, "Fundus fluorescence angiography in diagnosing diabetic retinopathy," *Pakistan J. Med. Sci.*, vol. 33, no. 6, p. 1328, 2017.

[16] Q. V. Hoang, J. Chua, M. Ang, and L. Schmetterer, "Imaging in myopia," in *Updates Myopia*. Springer, 2020, pp. 219–239.

[17] H. Fu *et al.*, "ADAM: Automatic detection challenge on age-related macular degeneration [data set]," in *Proc. IEEE DataPort*, 2020, doi: 10.21227/DT4F-RT59.

[18] H. Fu *et al.*, "PALM: Pathologic myopia challenge," in *Proc. IEEE Dataport*, 2019.

[19] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. CVPR*, 2018, pp. 3733–3742.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*. [Online]. Available: http://arxiv.org/abs/2002.05709

[21] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proc. CVPR*, 2019, pp. 6210–6219.

[22] H. Fu *et al.*, "Disc-aware ensemble network for glaucoma screening from fundus image," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2493–2501, Nov. 2018.

[23] P. Burlina, D. E. Freund, N. Joshi, Y. Wolfson, and N. M. Bressler, "Detection of age-related macular degeneration via deep learning," in *Proc. ISBI*, 2016, pp. 184–188.

[24] P. Yin *et al.*, "PM-Net: Pyramid multi-label network for joint optic disc and cup segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Springer, 2019, pp. 129–137.

[25] J. Cheng, Z. Li, Z. Gu, H. Fu, D. W. K. Wong, and J. Liu, "Structure-preserving guided retinal image filtering for optic disc analysis," in *Computational Retinal Image Analysis*. Elsevier, 2019, pp. 199–221.

[26] D. Milea *et al.*, "Artificial intelligence to detect papilledema from ocular fundus photographs," *New England J. Med.*, vol. 382, no. 18, pp. 1687–1695, 2020.

[27] F. Grassmann *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*, vol. 125, no. 9, pp. 1410–1420, Sep. 2018.

[28] C. R. Freire, J. C. da Costa Moura, D. M. da Silva Barros, and R. A. de Medeiros Valentim, "Automatic lesion segmentation and pathological myopia classification in fundus images," 2020, *arXiv:2002.06382*. [Online]. Available: http://arxiv.org/abs/2002.06382

[29] X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," in *Proc. Int. Conf. Image, Vis. Comput.*, Jun. 2017, pp. 783–787.

[30] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. CVPR*, 2017, pp. 1251–1258.

[31] Z. Zhou *et al.*, "Models genesis: Generic autodidactic models for 3D medical image analysis," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Springer, 2019, pp. 384–393.

[32] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101539.

[33] H. Spitzer, K. Kiwitz, K. Amunts, S. Harmeling, and T. Dickscheid, "Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Springer, 2018, pp. 663–671.

[34] M. Blendowski, H. Nickisch, and M. P. Heinrich, "How to learn from unlabeled volume data: Self-supervised 3D context feature learning," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Springer, 2019, pp. 649–657.

[35] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," 2019, *arXiv:1911.05722*. [Online]. Available: http://arxiv.org/abs/1911.05722

[36] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1734–1747, Sep. 2016.

[37] X. Li, Q. Dou, H. Chen, C.-W. Fu, and P.-A. Heng, "Multi-scale and modality dropout learning for intervertebral disc localization and segmentation," in *Proc. Int. Workshop Comput. Methods Clin. Appl. Spine Imag.* Springer, 2016, pp. 85–91.

[38] L. Brand, K. Nichols, H. Wang, L. Shen, and H. Huang, "Joint multi-modal longitudinal regression and classification for Alzheimer's disease prediction," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1845–1855, Jun. 2020.

[39] T. Zhou, H. Fu, G. Chen, J. Shen, J. Shen, and L. Shao, "Hi-net: Hybrid-fusion network for multi-modal mr image synthesis," *IEEE Trans. Med. Imag.*, early access, Feb. 20, 2020, doi: 10.1109/TMI.2020.2975344.

[40] T. Zhou, M. Liu, K.-H. Thung, and D. Shen, "Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2411–2422, Oct. 2019.

[41] X. Li *et al.*, "3D multi-scale FCN with random modality voxel dropout learning for intervertebral disc localization and segmentation from multi-modality MR images," *Med. Image Anal.*, vol. 45, pp. 41–54, Apr. 2018.

[42] L. Xing, M. Giger, and J. Min, *Artificial Intelligence in Medicine: Technical Basis and Clinical Applications*. Amsterdam, The Netherlands: Elsevier, 2020. [Online]. Available: https://books.google.com/books?id=i9FKzQEACAAJ

[43] C. Jayadev, N. Jain, S. Sachdev, A. Mohan, and N. Yadav, "Utility of noninvasive imaging modalities in a retina practice," *Indian J. Ophthalmol.*, vol. 64, no. 12, p. 940, 2016.

[44] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: http://arxiv.org/abs/1503.02531

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[46] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran, 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–41.

[48] P. Costa *et al.*, "End-to-end adversarial retinal image synthesis," *IEEE Trans. Med. Imag.*, vol. 37, no. 3, pp. 781–791, Mar. 2018.

[49] Á. Hervella, J. Rouco, J. Novo, and M. Ortega, "Paired and unpaired deep generative models on multimodal retinal image reconstruction," *Proceedings*, vol. 21, no. 1, p. 45, Aug. 2019.

[50] K. Li, L. Yu, S. Wang, and P.-A. Heng, "Unsupervised retina image synthesis via disentangled representation learning," in *Proc. Int. Workshop Simulation Synth. Med. Imag.* Springer, 2019, pp. 32–41.

[51] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[52] Z. Feng, C. Xu, and D. Tao, "Self-supervised representation learning from multi-domain data," in *Proc. CVPR*, 2019, pp. 3245–3255.

[53] P. M. Burlina, N. Joshi, M. Pekala, K. D. Pacheco, D. E. Freund, and N. M. Bressler, "Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks," *JAMA Ophthalmol.*, vol. 135, no. 11, pp. 1170–1176, 2017.

[54] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, "Zoom-in-net: Deep mining lesions for diabetic retinopathy detection," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Springer, 2017, pp. 267–275.

[55] Y. Zhou *et al.*, "Collaborative learning of semi-supervised segmentation and classification for medical images," in *Proc. CVPR*, 2019, pp. 2079–2088.

[56] O. J. Hénaff *et al.*, "Data-efficient image recognition with contrastive predictive coding," 2019, *arXiv:1905.09272*. [Online]. Available: http://arxiv.org/abs/1905.09272

[57] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," 2019, *arXiv:1906.05849*. [Online]. Available: http://arxiv.org/abs/1906.05849

[58] M. Jabi, M. Pedersoli, A. Mitiche, and I. B. Ayed, "Deep clustering: On the link between discriminative models and K-means," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 27, 2020, doi: 10.1109/TPAMI.2019.2962683.

[59] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.