Rotation-oriented Collaborative Self-supervised Learning for Retinal Disease Diagnosis

Xiaomeng Li, Xiaowei Hu, Xiaojuan Qi, Lequan Yu, Wei Zhao, Pheng-Ann Heng and Lei Xing

***Abstract*—The automatic diagnosis of various conventional ophthalmic diseases from fundus images is important in clinical practice. However, developing such automatic solutions is challenging due to the requirement of a large amount of training data and the expensive annotations for medical images. This paper presents a novel self-supervised learning framework for retinal disease diagnosis to reduce the annotation efforts by learning the visual features from the unlabeled images. To achieve this, we present a rotation-oriented collaborative method that explores rotation-related and rotation-invariant features, which capture discriminative structures from fundus images and also explore the invariant property used for retinal disease classification. We evaluate the proposed method on two public benchmark datasets for retinal disease classification. The experimental results demonstrate that our method outperforms other self-supervised feature learning methods (around 4.2% area under the curve (AUC)). With a large amount of unlabeled data available, our method can surpass the supervised baseline for pathologic myopia (PM) and is very close to the supervised baseline for age-related macular degeneration (AMD), showing the potential benefit of our method in clinical practice.**

***Index Terms*—Self-supervised learning, retinal disease classification**

I. Introduction

Fundus photography is a valuable clinical tool for evaluating various ophthalmic diseases, *e.g.*, aged-related macular degeneration (AMD) [1, 2], glaucoma (GON) [3–7], pathologic myopia (PM) [8], and diabetic retinopathy (DR) [9]. Recently, computer-aided detection techniques help ophthalmologists to automatically diagnose these retinal diseases by learning the representative features from fundus images though the deep convolutional neural networks (CNNs) [10, 9, 11–13]. These CNN-based methods require annotations of diseases in the fundus images. However, annotating the fundus image is tedious and expensive, where the professional knowledge is also required. Self-supervised learning (SSL), also called unsupervised visual representation learning, can help in this regard by providing a strategy to pre-train a neural network with unlabeled data, followed by fine-tuning for a downstream task with limited annotations. Hence, in this paper, our goal is to present a self-supervised method, which learns the representative features from the data itself without any human annotations. Then, the learned representation is evaluated on retinal disease classification tasks.

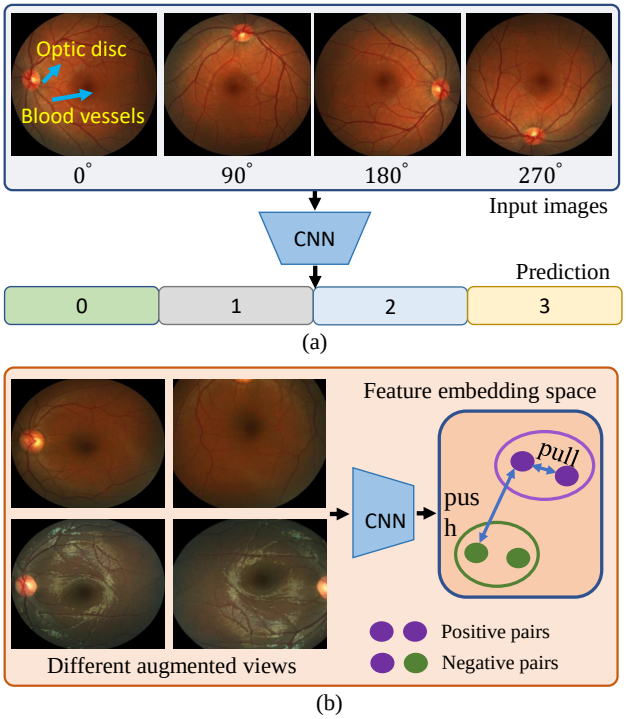

Figure 1: (a) shows the rotation prediction task. Each fundus image contains the obvious structures, *i.e.*, optic disc, and blood vessels. Rotating a fundus image by 90° will change the obvious orientation of these structures; (b) Images generated from one patient image under different augmentations (*positive pairs*) should be similar in the embedding space, while images from different patients (*negative pairs*) should be dissimilar.

Recently, self-supervised learning has attracted increasing attention in the medical imaging domain, due to it is free from human-annotated supervision and its potential of leveraging the massive amount of unlabeled data. Various types of self-supervised methods have been developed for multiple medical applications, such as subject identification from spinal MRI [14], cardiac MR image segmentation [15], lung lobe

This work was supported in part by a HKUST start-up research grant.

X. Li is with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, SAR, China. X. Li, L. Yu, W. Zhao and L. Xing are with the Department of Radiation Oncology, Stanford University, CA, 94305, USA (e-mail: eexmli@ust.hk; lei@stanford.edu). X. Hu and P.-A. Heng are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China. L. Yu is also with the Department of Statistics and Actuarial Science, the University of Hong Kong, Hong Kong SAR, China. X. Qi is with the Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong SAR, China. (Corresponding author: Xiaomeng Li and Lei Xing.)

segmentation and nodule detection [16], brain hemorrhage classification and brain tumor segmentation [17, 18]. The main idea is to pre-define a handcrafted pretext task, which is used to train a deep neural network to learn the visual features. The pretext task usually performs a transformation to the input images and requires the trained model to learn to predict properties of the transformation from the transformed image.

In this work, we formulate a rotation prediction task by adopting the rotation transformation to learn the rotation-related features. This is based on the crucial observation that the growth of abnormal blood vessels behind the macula tend to hemorrhage or leak fluid, which is an important cause of AMD [19]. Hence, exploring the low-level structure (*e.g*., vessel structures) information would be beneficial for the observation of retinal diseases. We observe that the structures of fundus images are sensitive to the orientations, *e.g*., the optic disc, and the blood vessels have specific directions, as shown in Figure 1(a). Hence, learning to predict rotations helps discover the vessel structures of the fundus images, which benefit self-supervised feature learning and then improve the retinal disease diagnosis.

Although the features learned by the rotation prediction task have explored the representative features for fundus images, these features make the diagnosis (classification) results sensitive to the image orientation, *i.e*., the network will produce different results for the input image with different rotations. In our task, the goal is to differentiate abnormal retinal diseases, which is invariant to image rotations. Hence, in addition to learn salient features, we present multi-view instance discrimination task to learn rotation-invariant features. Specifically, as shown in Figure 1(b), the multi-view instance discrimination aims to learn the feature representations that are similar to the representation of transformed versions of the input image and different from other images, where the transformations include rotation, randomly scaling, cropping, and the adjustment of the image brightness, contrast, and saturation. *By formulating the collaborative learning tasks, i.e., rotation prediction and multi-view instance discrimination, we encourage the network to discover the discriminative structures of fundus images and explore the robust representation used for fundus disease diagnosis, and then use the learned features to improve the overall performance of fundus disease diagnosis.* Three public datasets are employed to validate the effectiveness of our self-supervised method for retinal disease diagnosis. We summarize the main contributions of this work as follows:

- We present a novel rotation-oriented collaborative self-supervised learning method for disease classification from fundus images. Our method does not require any human-annotated labels during feature learning. With a large amount of unlabeled data available, our method can surpass the supervised baseline for PM and is very close to the supervised baseline for AMD (see Table II and III).
- We formulate a collaborative learning task that splits features to learn rotation-related and -invariant representations, which not only discover the discriminative structures from fundus images but also explores the invariant property used for retinal disease classification.
- Various experiments on two common eye diseases classification tasks demonstrate the superiority of our method than other state-of-the-art self-supervised methods (4.2% absolute improvement on AUC for AMD). Our code is publicly available at https://github.com/xmengli999/Rotation-oriented-self-supervised

II. Related Works

In this section, we first review related works on automatic ophthalmic disease diagnosis from fundus photography and then discuss some recent literatures on self-supervised feature learning.

A. Automatic Disease Diagnosis from Fundus Photography

Automated identification of retinal diseases is a big step towards early diagnosis and prevention of exacerbation of the disease. Early works for automatic retinal disease diagnosis from fundus photography are mainly based on the handcrafted features, such as AMD detection through texture analysis [20] or color filter based features [21]. Recently, a large portion of research is dedicated to supervised methods that show remarkable results with convolutional neural networks for automatic retinal disease recognitions [22, 23, 9, 24–27, 12, 28, 29]. For example, Burlina *et al*. [23] proposed a pretrained OverFeat feature for AMD classification from color fundus photos. Grassmann [27] classified AMD diseases into 13 classes through ensembling several convolutional neural networks. Recently, Peng *et al*. [12] developed a DeepSeeNet based on an Inception-v3 architecture [30] to identify patient-level AMD severity. Their method first detects individual risk factors and then the results are obtained by combining values from both eyes. For PM classification, Freire *et al*. [28] employed Xception [31] with ImageNet pretrained weights to classify PM and Non-PM from fundus images. Additional data such as RIGA and REFUGE datasets are also utilized as the training data. Xie *et al*. [32] trained the ImageNet pretrained ResNet50 with the labeled training data to classify PM from fundus images and this method achieved the highest result (99.74%) on a PM classification challenge [33]. Guo *et al*. [34] proposed a lesion-aware segmentation network to simultaneously classify and segment lesions.

However, these works are based on supervised learning, which adopts a massive amount of labeled data for training, and annotating fundus photography requires the substantial effort of human experts. Different from the previous works, in this paper, we focus on developing the self-supervised method for retinal disease diagnosis to reduce the annotation efforts.

B. Self-supervised Learning

Recently, self-supervised/unsupervised visual representation learning has attracted increasing attention due to its enormous potential of being free from human-annotated supervision and its extraordinary capability of leveraging the boundless unlabeled data. Various types of self-supervised methods have shown promising results in multiple application fields. In this section, we discuss some related self-supervised techniques in the domain of medical images and natural images.

Medical images. The key challenge for self-supervised learning is identifying a suitable self-supervision task, *i.e*., pretext task, to train the neural networks. Notable pretext tasks used in medical images include Rubik's cube and Rubik's cube+ recovery [17, 18], anatomical position prediction [15], reconstructing part of the image like image completion [35, 36], 3D distance prediction [37], image-intrinsic spatial offset prediction [38]. The common principle of these works is to construct different pretext tasks by discovering supervisory signals directly from the input data itself and train the deep network to predict this supervisory information, from which the high-level representation of the input is learned. For example, Zhuang *et al*. [17] proposed the Rubiks cube recovery task, *i.e*., cube ordering, and orientation pretext tasks, for the brain hemorrhage classification and tumor segmentation from CT and MR images. Zhu *et al*. [18] further improved this method and proposed Rubik's cube+ recovery task, which contains an additional masking identification pretext task. Bai *et al*. [15] formulated an anatomical position prediction pretext task to learn self-supervised features for cardiac MR image segmentation. Spitzer *et al*. [37] introduced a pretext task, which aims at predicting 3D distance between two patches sampled from the same brain. Recently, Taleb *et al*. [39] developed a series of 3D self-supervised methods for 3D medical images.

Natural images. Most of the above self-supervised methods defined a handcrafted pretext task to learn visual representation. This kind of idea has also been explored in the natural images, such as relative patch prediction [40, 41], image inpainting [42], colorizing gray-scale images [43], image jigsaw puzzle [44], geometric transformations [45, 46]. These methods are shown to be useful in various natural images. Yet, even with suitable architectures, these methods are being outperformed by contrastive methods [47].

Recently, contrastive methods [48–51], which are based on the task of instance discrimination, currently achieve state-of-the-art performance in self-supervised learning. The main idea of contrastive approaches is to bring representations of different views of the same image closer ('positive pairs') and spread representations of views from different images ('negative pairs') apart. For example, Dosovitskiy *et al*. [52] proposed to use the Softmax embedding with classifier weights to calculate the feature similarity, however, it prevents explicitly comparison over features, which results in limited efficiency and discriminability. Wu *et al*. [49] developed a memory bank to memorizes features of each instance. Ye *et al*. [50] calculated the positive concentrated property based on the "real" instance feature, instead of classifier weights [52] or memory bank [49]. He *et al*. [48] used a moving average network (momentum encoder) to maintain consistent representations of negative pairs drawn from a memory bank.

Most of the existing methods focus on designing a single pretext task to learn visual feature representation. In contrast, we present a novel collaborative method to learn the complementary information, *i.e*., rotation-related features and rotation-invariant features, from different pretext tasks, thus discovering the vessel structures in fundus images and discriminative features for retinal disease diagnosis, respectively.

C. Learning Rotation-invariant Features

Some methods learn rotation-invariant features by designing the network architecture to be rotation-invariant [53–55]. For example, Cheng [53] introduced a rotation-invariant layer and a Fisher discriminative layer and embedded them into a neural network. Our method learns rotation-invariant features by learning to predict rotations. Different from these related works, our method is a self-supervised method without modifying network architecture.

III. Methodology

Figure 2 shows the workflow of the overall architecture of our self-supervised method for retinal disease diagnosis. At the beginning, we randomly sample m images from the training dataset $S = \{x_i\}_{i=1}^{N}$. For each image x_i, we apply random data augmentation twice to generate $\widehat{x}_i$ and $\widetilde{x}_i$; see the x_1 and x_2 as examples in the Figure 2. Then, we generate the rotated images by rotating these augmented images by $0°, 90°, 180°, 270°$, and each image is assigned with a rotation label $0, 1, 2, 3$, correspondingly. After that, a feature embedding network $F(\cdot;\theta)$ is utilized to map the input x_i to a high-level feature vector $\mathbf{f}_i$, which is then decoupled into $\mathbf{f}_i^{(d)}$ and $\mathbf{f}_i^{(r)}$. These two decoupled features are collaboratively optimized by a multi-view instance discrimination task and a rotation prediction task. Finally, we employ the features learned from the multi-view instance discrimination task to perform the retinal disease classification. Below, we will elaborate on the rotation prediction, multi-view instance discrimination, and other network details.

A. Rotation prediction task

To discover the salient structures of fundus images, we perform the rotation prediction task to learn the rotation-related features. The input x_i is fed into a neural network (*e.g*., ResNet18) and we denote the output of the last residual block as feature $\mathbf{f}_i$. Note that each image x_i is rotated to obtain $x_{i,y}$ as the inputs, the actual feature should be $\mathbf{f}_{i,y}, y \in \{0, 1, 2, 3\}$. To simplify the description, here, we use $\mathbf{f}_i$ to represent $\mathbf{f}_{i,y}$. Then, to reduce feature dimension and get a high-level representation, two modules with a fully connected layer, followed by a BN and a ReLU, are applied sequentially after $\mathbf{f}_i$. Then, the feature is equally decoupled to $\mathbf{f}_i^{(r)}$ and $\mathbf{f}_i^{(d)}$ along the channel dimension. Finally, a fully connected layer, denoted as $F_c(\cdot;\theta_c)$, takes the feature $\mathbf{f}_i^{(r)}$ as the input and generates four probabilities, followed by a Softmax operation. As mentioned above, each image is assigned with a rotation label, *i.e*., 0, 1, 2, 3. Then, the rotation prediction loss is denoted as:

$$\mathcal{L}_r = \frac{1}{4N}\sum_{i=1}^{N}\sum_{y=0}^{3} l(F_c(\mathbf{f}_{i,y}{}^{(r)};\theta_c), y), \tag{1}$$

where l is the cross-entropy loss [56] used for the classification task and $y \in \{0, 1, 2, 3\}$ is the rotation label.

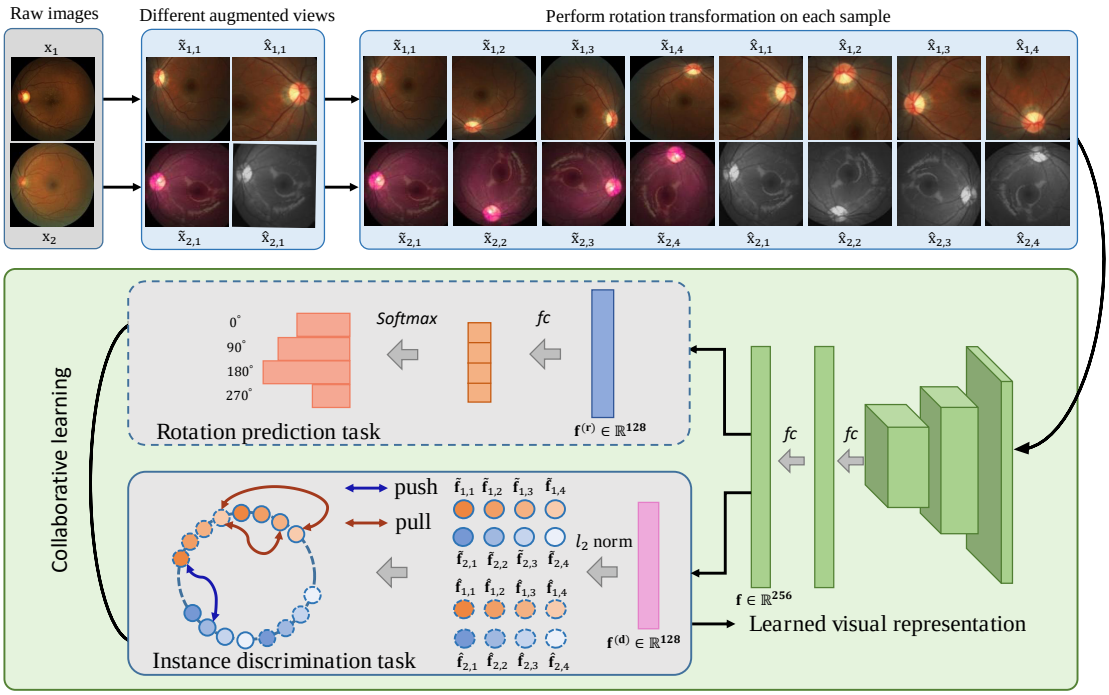

Figure 2: The illustration of the proposed method. We randomly select m raw images in one mini-batch and random data augmentation is applied twice to generate $\widetilde{x}_i$ and $\widehat{x}_i$. We visualize the case that $m = 2$ for visualization. Each image is rotated by $\{0°, 90°, 180°, 270°\}$ to derive the rotation-transformed images, *i.e.*, $\widetilde{x}_{i,1}$, $\widetilde{x}_{i,2}$, $\widetilde{x}_{i,3}$, $\widetilde{x}_{i,4}$,$\widehat{x}_{i,1}$, $\widehat{x}_{i,2}$, $\widehat{x}_{i,3}$, $\widehat{x}_{i,4}$. These images are fed into the neural network to learn the high-level feature $\mathbf{f}$, which is decoupled and then jointly optimized by two pretext tasks, *i.e.*, a multi-view instance discrimination task and a rotation prediction task. Finally, we adopt the features learned by the multi-view instance discrimination task and evaluate the learned features on retinal disease classification, based on a kNN classifier.

B. Multi-view instance discrimination task

To reveal the transformation-invariant representation for retinal disease diagnosis, we present the multi-view instance discrimination. As shown in Figure 2, $\widetilde{x}_{i,y}$ and $\widehat{x}_{i,y}$ denote different data augmented views of image x_i. The key hypothesis of the instance discrimination task is that the good features are shared between *multiple views* of the same fundus image. Hence, the objective is that different data augmented views (positive pairs) of a single image should be *invariant* in the embedding space, while images from different patients (negative pairs) should be *dissimilar*, as illustrations in Figure 1(b).

After obtaining feature $\mathbf{f}_i^{(d)}$ for image x_i, we first use l_2 normalization to normalize $\mathbf{f}_i^{(d)}$, *i.e.*, $\left\|\mathbf{f}_i^{(d)}\right\|_2 = 1$. For simplicity, in this section, we use $\mathbf{f}_i$ to represent $\mathbf{f}_i^{(d)}$. The positive pair is represented as $\left(\widehat{\mathbf{f}}_{i,y}, \widetilde{\mathbf{f}}_{i,k}\right)$, where y and k denote the rotation label, *i.e.*, $y, k \in \{0, 1, 2, 3\}$. The negative pair is denoted as $\left(\widehat{\mathbf{f}}_{i,y}, \widetilde{\mathbf{f}}_{j,k}\right)$, where $i \neq j$; see color illustrations in Figure 2. For each iteration, we randomly sample m images from the dataset. For each image x_i, the augmented samples should be classified into class i and the other images derived from x_j should not be classified to class i. Formally, the probability of $\widehat{x}_{i,y}$ being recognized as class i is defined by

$$P(i|\widehat{x}_{i,y}) = \frac{\exp\left(\sum_{k=0}^{3} \widetilde{\mathbf{f}}_{i,y}^T \widehat{\mathbf{f}}_{i,k}/\tau\right)}{\sum_{j=1}^{m} \exp\left(\sum_{k=0}^{3} \widetilde{\mathbf{f}}_{j,k}^T \widehat{\mathbf{f}}_{i,y}/\tau\right)}, \tag{2}$$

where τ is the temperature parameter that controls the concentration level of the sample distribution and τ is set to 0.1 by default [49]. $\widetilde{\mathbf{f}}_{i,y}^T \widehat{\mathbf{f}}_{i,k}$ denotes the cosine similarity between positive pairs while $\widetilde{\mathbf{f}}_{j,k}^T \widehat{\mathbf{f}}_{i,y}$ denotes the cosine similarity between negative pairs. Through the Softmax embedding function in Eq. (2), the network pushes "negative pairs" away and pulls "positive pairs" together. The final objective is minimizing the sum of the negative log likelihood over all the images within the batch, which is described as:

$$\mathcal{L}_d = -\sum_i \sum_y \log P(i|\widehat{x}_{i,y}) - \sum_i \sum_{j \neq i} \sum_y \log\{(1 - P(i|\widetilde{x}_{j,y}))\}, \tag{3}$$

where $P(i|\widehat{x}_{i,y})$ is the probability of $\widehat{x}_{i,y}$ being recognized as class i, and $1 - P(i|\widetilde{x}_{j,y})$ is the probability of $\widetilde{x}_{j,y}$ not being recognized as class i.

C. Network details

1) Loss function: The total objective is the weighted combination of a rotation prediction task and a multi-view instance discrimination task. The objective is denoted by

$$\mathcal{L} = \mathcal{L}_d + \lambda \mathcal{L}_r, \tag{4}$$

where λ is a weighting factor, indicating the importance of the rotation prediction task. In our experiment, we set $\lambda = 1$. We also analyze the effects of λ in Table V.

2) Network architecture: Our framework is based on the ResNet18 [57], following the same setting as the previous works [49, 50]. We apply a max pooling on the output of the last residual block in ResNet18. Then, the feature is flattened to a vector, and a fully connected layer, batch normalization, and ReLU are sequentially applied to reduce the feature dimension to 256. Then, $\mathbf{f}$ is equally split into $\mathbf{f}^{(r)}$ and $\mathbf{f}^{(d)}$ to learn the rotation prediction task and the multi-view discrimination task. A fully connected layer with the output channel 4 is applied on $\mathbf{f}^{(r)}$ to generate the probabilities for each rotation type, while a l_2 normalization layer is employed on the $\mathbf{f}^{(d)}$ to calculate the cosine similarities among features.

3) Implementation details: At each training iteration, m images are randomly selected, and random data augmentation is applied twice to the selected images, resulting in $2m$ generated images. Then each image is rotated by $\{0°, 90°, 180°, 270°\}$ to derive the rotation-transformed images, hence, the final batch size is $8m$. In our experiments, m is set to 64. This training strategy also takes the full advantage of relationships among all instances sampled in a mini-batch. To evaluate the learned feature, we apply the k-nearest neighbors (kNN) classifier based on the $\mathbf{f}^{(d)}$ and k is empirically set to 100.

The whole framework is built on PyTorch [58] with an NVIDIA Tesla V100 32GB GPU. We resized images to 320×320 resolution. For data augmentation, we randomly scaled and cropped images into the patches of size 224×224, with a random scaling factor chosen from [0.2, 1.0]. Our algorithm performs a 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 is chosen uniformly from [0.6, 1.4], following the setting in [50]. The network is optimized with Adam optimizer [59], the learning rate is set to 0.0001 and the weight decay is 0.0001.

IV. Experimental Results

A. Datasets

To evaluate the effectiveness of our method, we perform normal and abnormal classification to diagnose age-related macular degeneration (AMD) and pathological myopia (PM) on two public ophthalmic disease datasets, *i.e.*, Ichallenge-AMD and Ichallenge-PM, respectively. To the best of our knowledge, these two datasets are the only publicly available datasets for AMD and PM screening, respectively.

Ichallenge-AMD dataset. Ichallenge-AMD dataset [60] contains 1200 annotated retinal fundus images from 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. Labels of AMD/non-AMD, disc boundaries, and fovea locations, as well as boundaries of kinds of lesions are provided in this dataset. More detailed information about the dataset can be seen from Ichallenge-AMD website[1]. During the feature learning stage, we do not use any label information. Only the image-level labels are used in the AMD/non-AMD accuracy evaluation stage. The training, validation, and test dataset has 400 fundus images, respectively. Since only training data is released with annotations, we perform 5-fold cross-validation on the training dataset.

Ichallenge-PM dataset. Ichallenge-PM dataset [33] contains 1200 annotated color fundus images with labels, including both PM and non-PM cases. All the photos were captured with Zeiss Visucam 500. More detailed information can be found in the Ichallenge-PM website[2]. Note that the training stage does not need any annotations and only the image-level annotations are utilized during the evaluation stage. We also perform 5-fold cross-validation on this dataset.

EyePACS dataset. To evaluate the transfer learning capacity of our model among different diseases, we train the self-supervised model on a large diabetic retinopathy (DR) dataset, *i.e.*, Kaggles Diabetic Retinopathy Detection Challenge (EyePACS) dataset[3], and report the classification result on the Ichallenge-AMD and Ichallenge-PM datasets, respectively. EyePACS dataset is sponsored by the California Healthcare Foundation and the images are captured under various conditions and various devices. The left and right fields are provided for every subject, and an ophthalmologist rate the presence of DR in each image on a scale of 0 to 4. We use the training dataset (35,126 images) in this dataset to train our self-supervised model. Note that we do not use any human-annotated labels in this dataset.

B. Evaluation Metrics

We use AUC, Accuracy, Precision, Recall, F1-score to measure the classification performance. AUC stands for area under the receiver operating characteristic (ROC) curve, which measures the entire two-dimensional area underneath the entire ROC curve. ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The definitions of Accuracy, Precision, Recall and F1score are shown as follows.

$$\begin{aligned} \text{Accuracy} &= (TP + TN)/(TP + TN + FP + FN), \\ \text{Precision} &= TP/(FP + TP), \\ \text{Recall} &= TP/(TP + FN), \\ \text{F1} &= 2 * (\text{Recall} * \text{Precision})/(\text{Recall} + \text{Precision}), \end{aligned} \tag{5}$$

where TP, TN, FP, FN refer to true positive, true negative, false positive, false negative, respectively.

C. Comparisons with others on the Ichallenge-AMD dataset

We compare our method with the state-of-the-art unsupervised feature learning methods on the Ichallenge-AMD dataset. The results are shown in Table I.

[1] http://ai.baidu.com/broad/introduction?dataset=amd
[2] http://ai.baidu.com/broad/introduction?dataset=pm
[3] https://www.kaggle.com/c/diabetic-retinopathy-detection/data

Table I: Comparisons with other self-supervised methods on the Ichallenge-AMD dataset (Unit: %). Backbone: ResNet18.

	AUC	Accuracy	Precision	Recall	F1-score
Supervised	77.19	87.60	84.51	77.19	79.71
Rotation [46]	52.26	78.73	59.53	52.26	48.69
MemoryBank [49]	66.49	82.02	74.63	66.49	68.69
Contrastive [61]	68.06	82.45	73.48	68.06	69.84
Decouple [51]	69.19	83.80	79.72	69.19	71.62
Invariant [50]	71.42	84.31	77.99	71.42	73.67
Multi-modal [62] †	74.58	86.58	83.20	74.58	77.33
Ours	**75.64**	**87.09**	**83.96**	**75.64**	**78.51**

† need additional data to synthesize another modality.

Experimental settings. To have a fair comparison, all models were trained on the ResNet18 backbone [57] with 5-fold cross-validation. In the "Supervised" baseline, we modified the output channel of the last fully connected layer in ResNet18 to 2 and the model was trained with cross-entropy loss for binary classification.

We compare with other self-supervised methods, including rotation prediction task [46], instance discrimination methods [49, 61, 50] and collaborative method [51]. We run these methods with their released code on the Ichallenge-AMD dataset. For [46], we modified the output channel of the last fully connected layer of ResNet18 [57] to 4 and trained the network with cross-entropy loss for four rotation type predictions. To compare with instance discrimination based methods [49, 61, 50], we used the same training strategies and the only difference is the optimization method. Wu *et al*. [49] proposed an instance discrimination method to compute the similarity among instances. However, the memory bank saves the memorized feature and is only updated per epoch, which is inefficient and would cumber the training process. Chen *et al*. [61] showed that contrastive learning can be beneficial to unsupervised feature learning and Ye *et al*. [50] proposed a positive concentrated and negative spread out method. To fairly compare with these unsupervised methods, we trained all the models for 2000 epochs on the Ichallenge-AMD dataset. For simplicity, we perform kNN on all the unsupervised feature learning methods to evaluate the final performance for classification.

Results. From Table I, we can see that our method excels other state-of-the-art unsupervised feature learning methods by around 4.2% on AUC, which demonstrates that our method is very promising in the unsupervised feature learning. Compared to [62] that relies on additional data, our method still have a higher result (around 1.0% improvement). Notably, without any annotation during training, our method achieves comparable results with the supervised learning baseline, *i.e*., 75.64% vs 77.19% on AUC and 87.09% vs 87.60% on Accuracy. The results further demonstrated the effectiveness of our self-supervised learned features.

D. Comparison on the generalization among datasets

To show the generalization capability of our method, we trained the self-supervised model on the EyePACS dataset (*source dataset*) and evaluated on the Ichallenge-AMD and Ichallenge-PM datasets (*target datasets*), respectively.

Table II: Results obtained by first training a self-supervised model on the EyePACS dataset (source dataset) and then evaluating on the Ichallenge-AMD dataset (target dataset) with the kNN classifier. *Random weights* denotes the network weights are randomly initialized (Unit: %).

Method	Ichallenge-AMD				
	AUC	Accuracy	Precision	Recall	F1-score
Supervised	77.19	87.60	84.51	77.19	79.71
Random weights	50.00	78.23	39.11	50.00	43.86
Moco v2 [63]	62.93	83.29	**88.11**	62.93	65.66
Moco v1 [48]	65.39	83.04	77.94	65.39	67.81
Invariant [50]	74.43	86.07	81.71	74.43	76.95
Ours	**78.11**	**87.85**	85.58	**78.11**	**80.78**

Table III: Results obtained by first training a self-supervised model on the EyePACS dataset (source dataset) and then evaluating on the Ichallenge-PM dataset (target dataset) with the kNN classifier. *Random weights* denotes the network weights are randomly initialized (Unit: %).

Method	Ichallenge-PM				
	AUC	Accuracy	Precision	Recall	F1-score
Supervised	98.04	97.66	97.30	98.04	97.53
Random weights	91.83	91.35	90.93	91.83	91.11
Moco v2 [63]	97.97	98.11	98.18	97.97	98.06
Moco v1 [48]	96.60	97.30	97.40	96.60	96.91
Invariant [50]	97.83	98.11	98.35	97.83	98.05
Ours	**99.12**	**99.19**	**99.27**	**99.12**	**99.18**

Experimental settings. To adapt the method to the EyePACS dataset, we first resized the images to 256×256 and trained all the unsupervised methods for 150 epochs for a fair comparison. We then froze model parameters and *only* evaluated the model performance on the target datasets by the kNN classifier (k=100), respectively. The reported results are the 5-fold cross-validation results on the target datasets, *i.e*., Ichallenge-AMD and Ichallenge-PM datasets. The results are shown in Table II and Table III. "Random weights" denotes that the network weights are randomly initialized. To reproduce [50, 48, 63] on this dataset, we run these methods with the same network backbone (ResNet18), the same batch size ($b = 64$), learning strategies (Adam optimizer with learning rate 0.0001) and trained for the same epochs (150 epochs). All the kNN classifiers are evaluated at the features from the last fully connected layer. For the "Supervised" baseline, we trained the model on the target datasets with image-level labels, as procedures described in Section IV. B.

Results. From Table II and Table III, we can see that "Random weights" achieves a random result (50% AUC) on the Ichallenge-AMD dataset and a higher result (91.83%) on the Ichallenge-PM dataset, which indicates that pathological myopia classification is a much easier task. It is observed that Moco v2 [63] achieves limited results on AMD and PM classification tasks and this is due to that this method employs heavy data augmentations that may not be suitable for the fundus imaging. Moco v1 [48] gets a slightly higher result on the Ichallenge-AMD dataset, but a worse result on the Ichallenge-PM dataset. We can see that our method outperforms the other state-of-the-art method (*i.e*., Invariant [50])

Table IV: Results obtained by fine-tuning with different pretrained models on the Ichallenge-AMD dataset (Unit: %).

Pretrain	Ichallenge-AMD				
	AUC	Accuracy	Precision	Recall	F1-score
ImageNet	85.50	91.39	89.16	85.50	86.88
Unsupervised (**ours**)	86.99	90.64	86.34	86.99	86.40

AMD and PM classification tasks, respectively. These results further demonstrated the effectiveness of our method in terms of generalizing among datasets. Notably, we can see that our self-supervised method can achieve higher performance (around 1.1% improvement) than supervised baseline for PM diagnosis. Compared to our results in Table I, we can see that with more unlabeled fundus photos available, the performance of our self-supervised method can be further increased.

E. Comparison with the ImageNet pretrain

Our method provides an alternative approach to the ImageNet pretrained model. To validate this argument, we train the model on a large unlabeled fundus dataset, *i.e.*, EyePACS dataset, and fine-tune the model on the target dataset, *i.e.*, Ichallenge-AMD dataset. We compare this unsupervised model with the ImageNet pretrained model and both models have been fine-tuned on the target dataset in the same way.

Table IV shows the comparison of using the ImageNet pretrained model and the unsupervised pretrained model. We can find that our method achieves a higher AUC (86.99%) than the ImageNet pretrained model (85.5%). Note that in our approach, we trained the model with 35,126 fundus images without any annotations. However, in the ImageNet pretrained model, 1.2 million natural images with labels are employed. Compared to the ImageNet pretrained model, our method does not have annotation cost and achieves a higher AUC, showing the practical value of the proposed method.

F. Ablation Study

1) Importance of the rotation prediction task: Our framework collaboratively trains two pretext tasks and utilizes the output for the multi-view instance discrimination task as the final feature, *i.e.*, $\mathbf{f}^{(d)}$. The rotation prediction task serves as an auxiliary task to provide rich structure features during feature training, and then the multi-view instance discrimination task learns the transformation-invariant features that can be employed in the retinal disease diagnosis. Then, we analyze the importance of the auxiliary rotation prediction task in our framework.

Experimental settings. We trained our framework with different λ, where λ is the weight in Eq. (4) and indicates the importance of the rotation prediction task. $\lambda = 0.0$ denotes that the network is trained with only a multi-view instance discrimination task. As λ increases, the more important of the rotation prediction task in the network training. We trained models with different λ and each model was trained for 150 epochs. We used the same learning rate (0.0001) and batch size ($b = 64$). All the models were trained with the same network backbone (ResNet18).

Table V: The importance of the rotation prediction task. λ indicates the weight of the rotation prediction task, which is defined in Eq. (4). The models are trained on the EyePACS dataset and evaluated on the Ichallenge-AMD dataset. $\lambda = 0.0$ denotes using the vanilla rotation augmentation without the rotation prediction task.

	AUC	Accuracy	Precision	Recall	F1-score
$\lambda = 0.0$	74.43	86.07	81.71	74.43	76.95
$\lambda = 0.5$	75.70	86.84	83.90	75.70	75.41
$\lambda = 1.0$	**78.11**	**87.85**	**85.58**	**78.11**	**80.78**
$\lambda = 2.0$	72.17	85.57	83.63	72.17	75.49
$\lambda = 4.0$	72.08	85.57	81.88	72.08	75.12

Table VI: The importance of the rotation prediction task. λ indicates the weight of the rotation prediction task, which is defined in Eq. (4). The models are trained on the Ichallenge-AMD dataset by 5-fold cross-validation.

	AUC	Accuracy	Precision	Recall	F1-score
$\lambda = 0.5$	69.08	83.29	77.47	69.08	71.39
$\lambda = 1.0$	**75.64**	**87.09**	**83.96**	**75.64**	**78.51**
$\lambda = 2.0$	70.11	83.80	77.69	70.11	72.36

Results. The results are shown in Table V and Table VI. When $\lambda = 0.0$, the network only includes the multi-view instance discrimination task and the result is 74.43% on AUC in Table V. Note that $\lambda = 0.0$ denotes that using vanilla rotation augmentation without the rotation predict task. We found that this setting achieves lower results than our method ($\lambda = 1.0$). The comparison shows the effectiveness of the learning to prediction task, which learns effective features and cannot be replaced by vanilla rotation augmentation.

As λ increases, the classification performance improves and the best performance is reached when $\lambda = 1.0$. When λ continues increasing, the classification performance drops apparently from 78.11% to 72.17%. From Table VI, we can observe that with $\lambda = 1.0$, our method also achieves the best result on the Ichallenge-AMD dataset by 5-fold cross-validation. The results show that both the rotation prediction task and multi-view instance-wise discrimination task are useful in the training.

2) Effects of each individual task: Our method is a collaborative method that decouples features to the rotation-related and rotation-invariant features by formulating two tasks. Here, we analyze the effects of each task in Table VII. The experiments are conducted with 5-fold cross-validation on the Ichallenge-AMD dataset. The experimental setting keeps consistent with those in Table I.

As shown in Table VII, we can see that training the rotation prediction task alone can achieve very limited performance (52.26% AUC), while training with the instance discrimination task can reach a higher result, *i.e.*, 71.42% on AUC. This phenomenon can also be found in [47], which indicated that instance discrimination task, *i.e.*, contrastive method, outperforms other the handcrafted pretext tasks. It is also observed that through collaborative training of these two tasks, our method can achieve a higher result for disease classification.

3) Analysis on the data augmentation: In this section, we

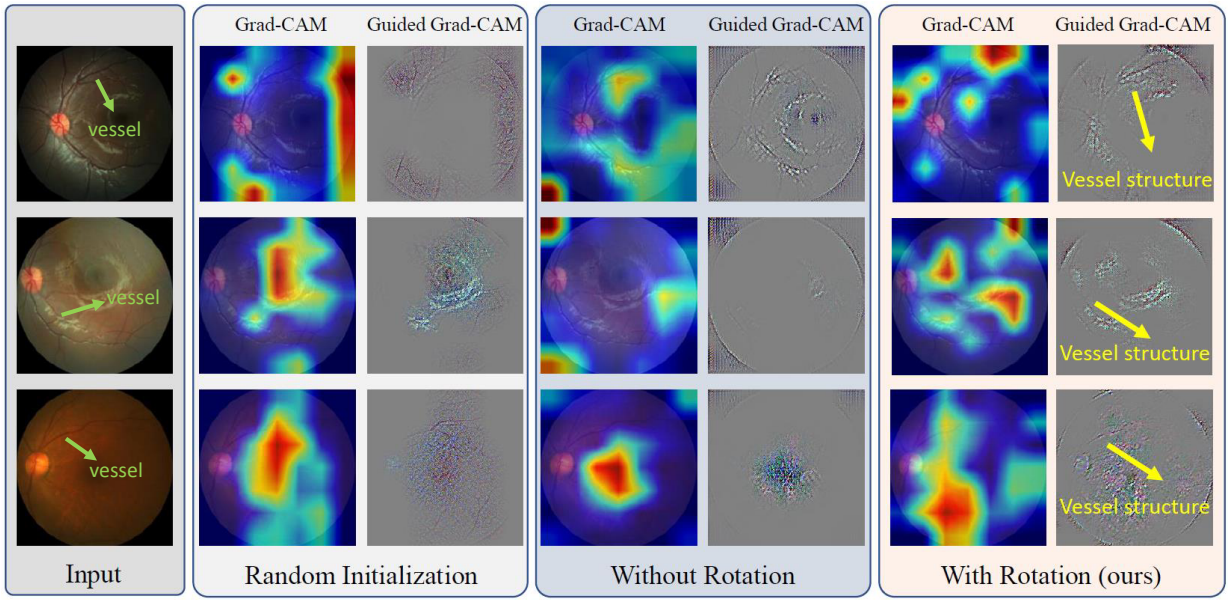

Figure 3: The visualization of the features of three examples from the self-supervised models. "Random weights" denotes that the model weights are randomly initialized, while "Without" and "With" Rotation refer to train the network without/with the rotation task. Grad-CAM [64] (blue heatmap) highlights the discriminative regions (red regions), which correspond to high scores. We can see that our method captures more discriminative regions; see the 6th column. Guided Grad-CAM [65] gives high-resolution discriminative visualizations and we can also find salient structures such as vessels; see the last column. Best viewed in color.

Table VII: Effects of each individual task on the Ichallenge-AMD dataset (Unit: %).

	AUC	Accuracy	Precision	Recall	F1-score
Rotation	52.26	78.73	59.53	52.26	48.69
InstDist	71.42	84.31	77.99	71.42	73.67
Ours	**75.64**	**87.09**	**83.96**	**75.64**	**78.51**

Table VIII: Results by different augmentation strategies on the Ichallenge-AMD dataset (Unit:%).

	AUC	Accuracy	Precision	Recall	F1-score
4n+Rot	61.45	80.00	69.76	61.45	62.98
Ours	**75.64**	**87.09**	**83.96**	**75.64**	**78.51**

analyze the different strategies of data augmentation. Our method first augments the input images with random scaling, random left-right flip, random intensity modifications, *etc*. On top of that, we further augment 4 rotated versions for each image.

4) Feature visualization: In this section, we visualize features to verify whether the rotation prediction task can successfully learn the salient/structure features on fundus images. We have shown the results without and with the rotation prediction task in Table V (see the 1st and 3rd rows). We visualize the features from these two models in Figure 3, and the showed feature is obtained from the 1st layer in the last residual block. The "input" column is randomly selected from the target dataset (Ichallenge-AMD dataset). "Random weights" denotes that the model weights are randomly initialized. We visualize the features through Grad-CAM [64] and Guided Grad-CAM. Grad-CAM [64] (blue heatmap) localizes discriminative regions and represents where the model has to look to make the particular decision. Guided Grad-CAM gives high-resolution discriminative visualizations, which is obtained by pointwise multiplying the heatmap with guided backpropagation [65]. The red region corresponds to a high score (discriminative regions). From observations on Guided Grad-CAM in Figure 3, we can see that compared with the other two alternatives, "With Rotation" can present more clear retinal structures, such as vessel structures. The visualization indicates that the rotation prediction task can help the network in learning salient features or obvious structures.

5) Visualizations of kNN results: The final result is obtained by running kNN on the features from the multi-view instance discrimination task. Hence, we retrieve 5-nearest neighbors from the training set for each test image based on the similarity scores through the kNN algorithm. The label and the similarity score are listed below each image. The higher score is, the more similar to the test image. We can see from Figure 4 that compared with 299th and 300th neighbors, the retrieved images have high visual similarity with the test image, which can help in assigning a correct class to the test image. It is also observed that the majority vote of the 5-nearest neighbors keeps the same with the label of the test sample.

V. Discussion

Fundus photography is an important tool in assessing retinal diseases, such as AMD classification [23, 66, 27, 12], DR grading [9, 11, 10, 67] and PM classification [33], *etc*. With the advances of deep learning techniques, automatic retinal

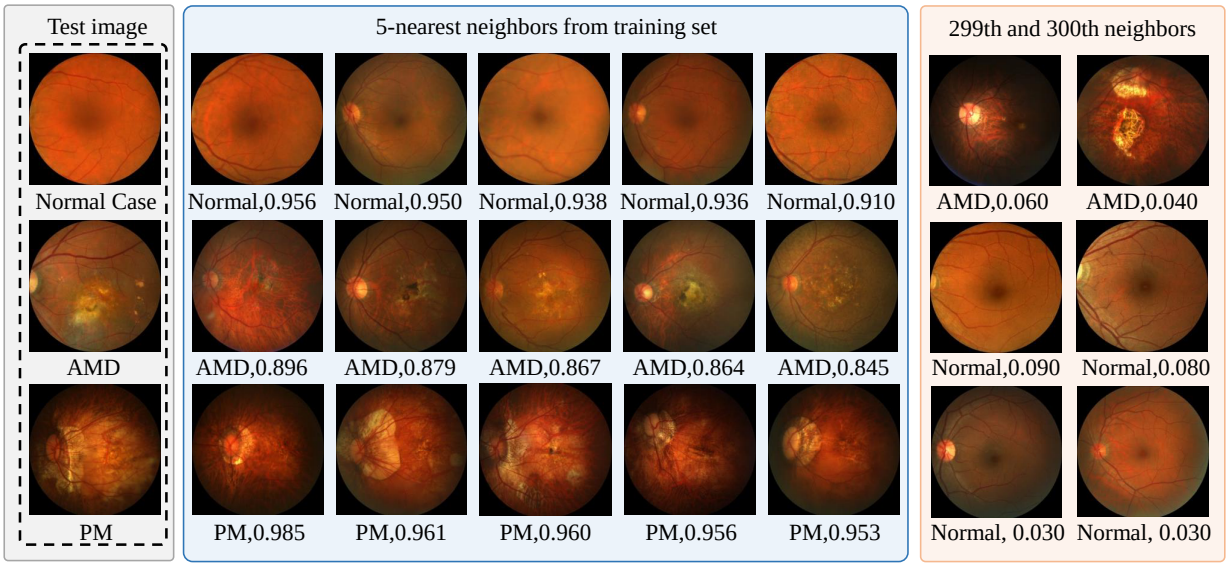

Figure 4: The final result is obtained by a kNN, *i.e.*, majority voting the k-nearest neighbors. Hence, we retrieve 5-nearest neighbors from the training set for each test image based on the cosine similarity scores in the kNN algorithm. We also show dissimilar neighbors for reference. The label and the similarity score are listed below each image. The higher score is, the more similar to the test image. We can see that the retrieved 5-nearest neighbors have high visual similarity with the test sample, which can contribute to assign a correct class to the test image.

disease diagnosis, such as AMD and PM, has been well studied in the research community. Although promising results were obtained on these diagnosis tasks, these methods require human annotations during the model development, which is costly and expensive to obtain. Self-supervised/unsupervised techniques that learn representation from data itself without annotations provide solutions for this issue. In this work, we present a rotation-oriented collaborative self-supervised model for retinal disease diagnosis. Different from previous self-supervised works [48, 63, 50], we formulate collaborative learning pretext tasks, i.e., rotation prediction and multi-view instance discrimination, to decouple features to both *rotation-related* and *rotation-invariant* features, which help discover the discriminative structures of fundus images and reveal the transformation-invariant representation for retinal disease diagnosis, respectively. Our method is validated on two public retinal disease datasets, *i.e.*, Ichallenge-AMD and Ichallenge-PM datasets, in which our method consistently outperforms other self-supervised methods. With a large amount of unlabeled data available, our method can surpass the supervised baseline for PM and very close to the supervised baseline for AMD.

Our method decouples features to rotation-related and rotation-invariant features by collaboratively training two pretext tasks based on the core observations from the color fundus images. The rotation prediction task learns the salient structures and the instance discrimination task is a contrastive learning method that learns transformation-invariant features. Although our method achieves excellent performance, it comes with limitations. There are some other pretexts, such as image painting [42], relative position prediction [44], which may also help decouple features to several types, but are less studied in this work. In the future, we will investigate the advantages of different pretext tasks, and study how to design a better pretext that contributes to the representation feature learning. Another potential research direction is to extend our method to other medical image applications that have obvious orientation characteristics, such as liver, kidney CT, and MR.

Another limitation is that our method only tackle the 2-class classification problem, i.e., the retinal disease cases with obvious lesion patterns, e.g., AMD and PM. The developed rotation prediction task can bring obvious changes when the image rotates. However, DR grading is a challenging task and the grading task requires to know the location and size of different lesions, such as microaneurysms, haemorrhages, microvascular anomalies. In this paper, the method is developed for 2-class normal and abnormal classification, where abnormal case contains obvious lesions. The exploration of the method on more retinal disease diagnosis tasks, including segmentation, grading, and detection would be our future work.

VI. Conclusion

This paper presents a novel self-supervised learning method for retinal disease diagnosis. Our key idea is to learn the visual features from the unlabeled images by developing the rotation-oriented collaborative pretext tasks, *i.e.*, a rotation prediction task, and a multi-view instance discrimination task. The rotation prediction helps to discover the discriminative structures of fundus images by learning the rotation-related features, while the multi-view instance discrimination helps to explore the rotation-invariant features for retinal disease classification. These two features, *i.e.*, rotation-related and rotation-invariant features, are obtained by decoupling features

through collaboratively training two pretexts. Experimental results on two benchmark datasets demonstrate that our method outperforms state-of-the-art self-supervised learning methods. With a large amount of unlabeled data available, our method can surpass the supervised baseline for PM and is very close to the supervised baseline for AMD, showing the potential benefit of our method in clinical practice.

REFERENCES

[1] A.-R. E. D. S. R. Group *et al.*, "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] E. K. de Jong, M. J. Geerlings, and A. I. den Hollander, "Age-related macular degeneration," in *Genetics and Genomics of Eye Disease*. Elsevier, 2020, pp. 155–180.

[3] B. E. Klein, R. Klein, W. E. Sponsel, T. Franke, L. B. Cantor, J. Martone, and M. J. Menage, "Prevalence of glaucoma: the beaver dam eye study," *Ophthalmology*, vol. 99, no. 10, pp. 1499–1504, 1992.

[4] J. Cheng, J. Liu, Y. Xu, F. Yin, D. W. K. Wong, N.-M. Tan *et al.*, "Superpixel classification based optic disc and optic cup segmentation for glaucoma screening," *IEEE transactions on medical imaging*, vol. 32, no. 6, pp. 1019–1032, 2013.

[5] A. Almazroa, S. Alodhayb, E. Osman, E. Ramadan, M. Hummadi, M. Dlaim *et al.*, "Retinal fundus images for glaucoma analysis: the riga dataset," in *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, vol. 10579. International Society for Optics and Photonics, 2018, p. 105790B.

[6] X. Chen, Y. Xu, D. W. K. Wong, T. Y. Wong, and J. Liu, "Glaucoma detection based on deep convolutional neural network," in *2015 37th annual international conference of the IEEE engineering in medicine and biology society*. IEEE, 2015, pp. 715–718.

[7] T. Saba, S. T. F. Bokhari, M. Sharif, M. Yasmin, and M. Raza, "Fundus image classification methods for the detection of glaucoma: A review," *Microscopy research and technique*, vol. 81, no. 10, pp. 1105–1121, 2018.

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

[9] 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 transactions on medical imaging*, vol. 39, no. 5, pp. 1483–1493, 2019.

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

[11] A. Sakaguchi, R. Wu, and S.-i. Kamata, "Fundus image classification for diabetic retinopathy using disease severity grading," in *International Conference on Biomedical Engineering and Technology*, 2019, pp. 190–196.

[12] Y. Peng, S. Dharssi, Q. Chen, T. D. Keenan, E. Agrón, W. T. Wong *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, 2019.

[13] J. Virmani, G. P. Singh, Y. Singh *et al.*, "Pnn-based classification of retinal diseases using fundus images," in *Sensors for Health Monitoring*. Elsevier, 2019, pp. 215–242.

[14] 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.

[15] W. Bai, C. Chen, G. Tarroni, J. Duan, F. Guitton, S. E. Petersen *et al.*, "Self-supervised learning for cardiac mr image segmentation by anatomical position prediction," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2019, pp. 541–549.

[16] N. Tajbakhsh, Y. Hu, J. Cao, X. Yan, Y. Xiao, Y. Lu *et al.*, "Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data," in *International Symposium on Biomedical Imaging*. IEEE, 2019, pp. 1251–1255.

[17] X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng, "Self-supervised feature learning for 3d medical images by playing a rubiks cube," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2019, pp. 420–428.

[18] J. Zhu, Y. Li, Y. Hu, K. Ma, S. K. Zhou, and Y. Zheng, "Rubiks cube+: A self-supervised feature learning framework for 3d medical image analysis," *Medical Image Analysis*, p. 101746, 2020.

[19] N. Ferrara, "Vascular endothelial growth factor and age-related macular degeneration: from basic science to therapy," *Nature medicine*, vol. 16, no. 10, pp. 1107–1111, 2010.

[20] M. Garnier, T. Hurtut, H. B. Tahar, and F. Cheriet, "Automatic multiresolution age-related macular degeneration detection from fundus images," in *Medical Imaging 2014: Computer-Aided Diagnosis*, vol. 9035. International Society for Optics and Photonics, 2014, p. 903532.

[21] S. Waseem, M. U. Akram, and B. A. Ahmed, "Drusen detection from colored fundus images for diagnosis of age related macular degeneration," in *7th International Conference on Information and Automation for Sustainability*, 2014, pp. 1–5.

[22] H. Fu, J. Cheng, Y. Xu, C. Zhang, D. W. K. Wong, J. Liu, and X. Cao, "Disc-aware ensemble network for glaucoma screening from fundus image," *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2493–2501, 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 *International Symposium on Biomedical Imaging*. IEEE, 2016, pp. 184–188.

[24] P. Yin, Q. Wu, Y. Xu, H. Min, M. Yang, Y. Zhang *et al.*, "Pm-net: Pyramid multi-label network for joint optic disc and cup segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. 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, R. P. Najjar, J. Zhubo, D. Ting, C. Vasseneix, X. Xu *et al.*, "Artificial intelligence to detect papilledema from ocular fundus photographs," *New England Journal of Medicine*, 2020.

[27] F. Grassmann, J. Mengelkamp, C. Brandl, S. Harsch, M. E. Zimmermann, B. Linkohr *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, 2018.

[28] C. R. Freire, J. C. d. C. Moura, D. M. d. S. Barros, and R. A. d. M. Valentim, "Automatic lesion segmentation and pathological myopia classification in fundus images," *arXiv preprint arXiv:2002.06382*, 2020.

[29] M. Badar, M. Haris, and A. Fatima, "Application of deep learning for retinal image analysis: A review," *Computer Science Review*, vol. 35, p. 100203, 2020.

[30] X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," in *International Conference on Image, Vision and Computing*. IEEE, 2017, pp. 783–787.

[31] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[32] R. Xie, L. Liu, J. Liu, and C. S. Qiu, "Pathological myopic image analysis with transfer learning," *arXiv preprint*

arXiv:1908.00410, 2019.

[33] H. Fu, F. Li, J. I. Orlando, H. Bogunovi, X. Sun, J. Liao *et al.*, "Palm: Pathologic myopia challenge," *IEEE Dataport*, 2019.

[34] Y. Guo, R. Wang, X. Zhou, Y. Liu, L. Wang, C. Lv, B. Lv, and G. Xie, "Lesion-aware segmentation network for atrophy and detachment of pathological myopia on fundus images," in *International Symposium on Biomedical Imaging*. IEEE, 2020, pp. 1242–1245.

[35] Z. Zhou, V. Sodha, M. M. R. Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway *et al.*, "Models genesis: Generic autodidactic models for 3d medical image analysis," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2019, pp. 384–393.

[36] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical image analysis*, vol. 58, p. 101539, 2019.

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

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

[39] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert, "3d self-supervised methods for medical imaging," *arXiv preprint arXiv:2006.03829*, 2020.

[40] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *IEEE International Conference on Computer Vision*, 2015, pp. 1422–1430.

[41] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *IEEE International Conference on Computer Vision*, 2017, pp. 2051–2060.

[42] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[43] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European Conference on Computer Vision*. Springer, 2016, pp. 649–666.

[44] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*. Springer, 2016, pp. 69–84.

[45] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with convolutional neural networks," in *Neural Information Processing Systems*, 2014, pp. 766–774.

[46] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *The International Conference on Learning Representations*, 2018.

[47] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *arXiv preprint arXiv:2006.07733*, 2020.

[48] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[49] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.

[50] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6210–6219.

[51] Z. Feng, C. Xu, and D. Tao, "Self-supervised representation learning by rotation feature decoupling," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 364–10 374.

[52] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1734–1747, 2015.

[53] G. Cheng, P. Zhou, and J. Han, "Rifd-cnn: Rotation-invariant and fisher discriminative convolutional neural networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2884–2893.

[54] ——, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016.

[55] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 265–278, 2018.

[56] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, 1958.

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

[58] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito *et al.*, "Automatic differentiation in pytorch," 2017.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *The International Conference on Learning Representations*, 2015.

[60] H. Fu, F. Li, J. I. Orlando, H. Bogunovi, X. Sun, J. Liao *et al.*, "Adam: Automatic detection challenge on age-related macular degeneration," *IEEE Dataport*, 2020.

[61] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[62] X. Li, M. Jia, M. T. Islam, L. Yu, and L. Xing, "Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4023–4033, 2020.

[63] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

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

[65] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[66] 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 ophthalmology*, vol. 135, no. 11, pp. 1170–1176, 2017.

[67] Y. Zhou, X. He, L. Huang, L. Liu, F. Zhu, S. Cui *et al.*, "Collaborative learning of semi-supervised segmentation and classification for medical images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2079–2088.