Few-Shot Transfer Learning for Hereditary Retinal Diseases Recognition

Siwei Mai[1], Qian Li[2], Qi Zhao[2], and Mingchen Gao[1(✉)]

[1] Department of Computer Science and Engineering, University at Buffalo, Buffalo, USA
{siweimai,mgao8}@buffalo.edu

[2] Department of Ophthalmology, Beijing Tongren Eye Center, Beijing Tongren Hospital, Beijing Key Laboratory of Ophthalmology and Visual Sciences, Capital Medical University, Beijing, China

Abstract. This project aims to recognize a group of rare retinal diseases, the hereditary macular dystrophies, based on Optical Coherence Tomography (OCT) images, whose primary manifestation is the interruption, disruption, and loss of the layers of the retina. The challenge of using machine learning models to recognize those diseases arises from the limited number of collected images due to their rareness. We formulate the problems caused by lacking labeled data as a Student-Teacher learning task with a discriminative feature space and knowledge distillation (KD). OCT images have large variations due to different types of macular structural changes, capturing devices, and angles. To alleviate such issues, a pipeline of preprocessing is first utilized for image alignment. Tissue images at different angles can be roughly calibrated to a horizontal state for better feature representation. Extensive experiments on our dataset demonstrate the effectiveness of the proposed approach.

Keywords: Hereditary Retinal Diseases Recognition · Student-Teacher learning · Knowledge distillation · Transfer learning

1 Introduction

Visual impairment and blindness caused by inherited retinal diseases (IRDs) are increasing due to the global prolonged life expectancy. There was no treatment for IRDs until recently, a number of therapeutic approaches such as gene replacement and induced pluripotent stem cell transplantation have been proposed, developed, and shown promising potential in some of the ongoing therapeutic clinical trials. Spectral-domain Optical coherence tomography (SD-OCT) has been playing a crucial role in the evaluation of the retina of IRDs in diagnosis, progression surveillance as well as strategy exploration and response assessment of the treatment. However, the recognition, interpretation, and comparison of the minimal changes on OCT as shown in IRDs sometimes could be difficult and time-consuming for retinal physicians. Recently automated image analysis

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12908, pp. 97–107, 2021.
https://doi.org/10.1007/978-3-030-87237-3_10

has been successfully applied in the detection of changes on fundus and OCT images of multiple retinal diseases such as diabetic retinopathy and age-related macular degeneration, which are of higher prevalence among the population to enable the acquisition of large volumes of training data for traditional machine learning approaches including deep learning.

On the contrary, for rare diseases like IRDs, acquiring a large volume of high-quality data representative of the patient cohorts is challenging. These datasets also require accompanying annotations generated by experts which are time-consuming to produce. This hinders the applying state-of-the-art image classification, which usually requires a relatively large number of images with annotations for training. The goal of this project is to design a computer-aided diagnosis algorithm when only a very limited number of rare disease samples can be collected.

The methods of diagnosing ocular diseases like age-related macular degeneration (AMD), diabetic macular edema (DME), etc., through the spectral domain OCT images can be roughly categorized into the traditional machine learning methods and deep learning-based methods. There are lots of works on OCT image analysis based on the traditional machine learning methods like Principal Components Analysis (PCA) [1,15], Support Vector Machine (SVM) [12,17], or Random Forest [7], segmenting each layer of the OCT images [18] or learns global representation directly [19,21].

Lots of previous work also focus on the deep-learning-based methods including supervised and unsupervised ways. Existing mature and pre-trained frameworks such as Inception-v3 [8,22], VGG16 [16,22], PCANet [4], GoogLeNet [9,10], ResNet [9,13], DenseNet [9] have been deployed to classify OCT images. Others unify multiple networks together to make classification more robust for diagnosing, for example four-parallel-ResNet system [13] and multi-stage network [14]. Supervised learning has the advantage of learning hierarchical features compared to traditional feature engineering. However, for supervised learning of OCT medical images, satisfactory results are still dependent on large amounts of data.

In addition to the supervised learning methods above, we also try to address the few-shot learning problem based on the contrastive learning. We empirically show that the Siamese network architecture represented by Simple Siamese [2] is able to learn meaningful representations and bypass the limitation of sample size. Also, the embedding features in feature space obtained by the contrastive learning [5] leads to more knowledge learned by the student network in the subsequent S-T architecture.

The goal of this research is to classify a group of macular-involved IRDs from different stages by a limited number of OCT images. Given the limited training data, we plan to assist the classification with an auxiliary dataset in a related task where labeled data are abundant. We propose a Student-Teacher Learning framework to leverage the knowledge from the auxiliary dataset. In the teacher part, the teacher model is firstly trained on a large-scale labeled auxiliary OCT dataset [11] which contains 3 common retinal degenerative diseases with 84484 images. Soft Nearest Neighbor Loss (SNNL) [5] is utilized to maximize the representation entanglement of different classes to help generalization. Transfer

Learning methods is then applied to adapt the teacher model to the target label space. While for the student model, the collected OCT samples are used to serve as the hard labels. The student model can learn from both the teacher model and the hard label information by Knowledge Distillation [6]. We have collected 1128 diseased OCT images (185 of them are normal) from 60 patients (15 of them are normal) with IRDs. The experiments on the collected dataset demonstrated that, even under the circumstance of limited training samples, the student model can catch a better performance than the teacher model and some common few-shot learning methods [24].

2 Methods

The overview of the proposed method is shown in Fig. 1. The proposed pipeline consists of three parts: image preprocessing, training for the teacher model and then the student model. The OCT images are first normalized to reduce the effect of noise on the model during training. The teacher model is designed to adapt to the target OCT dataset based on a projector trained with the auxiliary OCT dataset by Soft Nearest Neighbor Loss (SNNL). The student model learns the knowledge from "soft" labels from the teacher model and the "hard" labels from target dataset. In general, the structure of the student model is smaller than that of the teacher model to prevent overfitting and to increase the training speed. The source code is available in https://github.com/hatute/FSTL4HRDR.

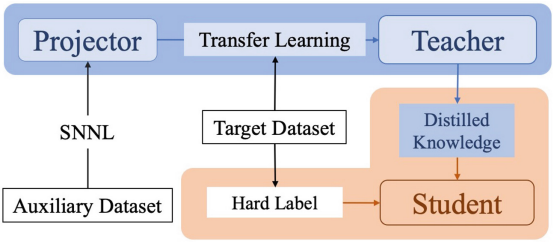


Fig. 1. The overview of the proposed method

The problem of identifying congenital diseases is modeled as a problem of few-shot learning. Unlike the unilateral optimization of model training methods, our approach combines contrastive learning, transfer learning and knowledge distillation to enhance the fast learning ability of the model from various aspects. Second, the teacher-student learning model allows the student to learn more "knowledge" and achieve better performance than the teacher, compared to fine-tuning the original model directly.

2.1 Image Preprocessing

As shown in Fig. 3, the original OCT images show different angles, noise distribution, and size diversity due to the acquisition machine and the patient. This

will distract the neural network from the focal area and increase the training time due to the useless data input during training. We adjust all images to the horizontal position without destroying the original pathological information following a adapted OCT image preprocessing strategy from [19]. The main idea of process is to generate a mask to attain retina layered structure.

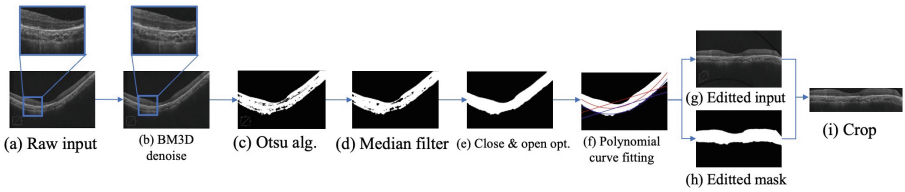


Fig. 2. Pipeline for the image alignment. The process begins with noise reduction as shown in (b) to reduce the irregularly distributed noise with Block-matching and 3D filtering (BM3D) [3] for better capturing the retina structure. (c) The Otsu algorithm allocates the location and morphology of the black background. (d) The median filter further reduces the noise area within the tissue. (e) The morphological operations opening and closing clean noises inside and outside the tissue area. (f) After the contours are obtained, we use a polynomial curve fitting to represent the curvature of the tissue area for adjusting and cropping both the mask and the original image as shown in (i).

2.2 Feature Space Learning

Teacher model is the backbone structure for absorbing and learning the information from an auxiliary dataset, more specifically, the textures, patterns, and pixel distributions in the end-level convolutional layers. Soft Nearest Neighbor Loss (SNNL) [5] is applied for the teacher model training in the feature space before the classifier. SNNL is designed to enhance the separation of class manifolds in representation space. There are bound to be objective differences between the target and auxiliary datasets, and the great separation between the categories in the feature space will facilitate the subsequent transfer learning [5,20] with the target dataset.

Equation 1 shows the total loss function, which consists of the cross-entropy loss on logits and the soft nearest neighbor loss for the representation learning controlled by the hyper-parameter α. i for selected samples in the batch. j for another sample in the same category as i. k for another sample in the same batch as i. In Eq. 2, b is the batch size, T is the temperature. When the temperature is large, the distances between widely separated points can influence the soft nearest neighbor loss more. Moreover, the numerator of the log function implies the distance between the target, and similar samples in each category, while the denominator is the distances between the target and other samples in the batch.

Usually, the use of cosine distances in training results in a smoother training process.

$$\mathcal{L} = \sum_{j} y_j \log f^k(x_j) + \alpha \cdot \sum_{i \in k-1} l'_{sn}(f^i(x), y) \tag{1}$$

$$l'_{sn} = \arg\min_{T \in \mathbb{R}} -\frac{1}{b} \sum_{i \in 1 \ldots b} \log \left(\frac{\sum\limits_{\substack{j \in 1 \ldots b \\ j \neq i \\ y_i = y_j}} e^{-\frac{\|x_i - x_j\|^2}{T}}}{\sum\limits_{\substack{k \in 1 \ldots b \\ k \neq i}} e^{-\frac{\|x_i - x_k\|^2}{T}}} \right) \tag{2}$$

2.3 Knowledge Distillation and Student-Teacher Learning

In order to overcome the obstacles caused by the lack of training data, we use the combination of Knowledge Distillation and Student-Teacher Learning for knowledge transfer [6,23]. The S-T architecture is designed to give the small-scale (student) model the ability to adapt to small samples and to absorb knowledge from large-scale auxiliary dataset learned by the teacher model. This is mainly designed to eliminate the overfitting problem, when large models cannot learn parameters effectively with a small amount of data.

$$\mathcal{L}(x; W) = \alpha \cdot H(y, \sigma(z_s; T = 1)) + \beta \cdot H(\sigma(z_t; T = \tau), \sigma(z_s; T = \tau)), \tag{3}$$

where α and β control the balance of information coming from the two sources, which generally add up to 1. H is the loss function, σ is the softmax function parameterized by the temperature T, z_s is the logits from student network and z_t is the logits from teacher network. τ denotes the temperature of adapted softmax function, each probability p_i of class i in the batch is calculated from the logits z_i as:

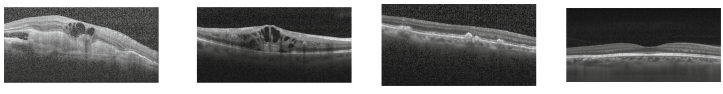
$$p_i = \frac{\exp(\frac{z_i}{T})}{\sum_j \exp(\frac{z_j}{T})}, \tag{4}$$

when T increases, the probability distribution of the output becomes "softer", which means the differences among the probability of each class decreased and more information will provide. By the S-T architecture, the smaller size student model with the blank background is able to accept the knowledge from the fine-tuned teacher as well as information from labels.

3 Experiments

3.1 Datasets

Auxiliary Dataset. We use a publicly available dataset of OCT images as shown in Fig. 3 from Cell dataset [11] and BOE dataset [17] for training. The Cell



(a) CNV (b) DME (c) Drusen (d) Normal

Fig. 3. Four types of samples in the auxiliary dataset.

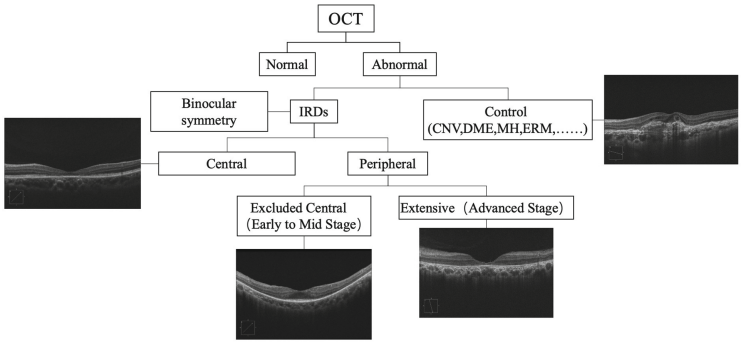
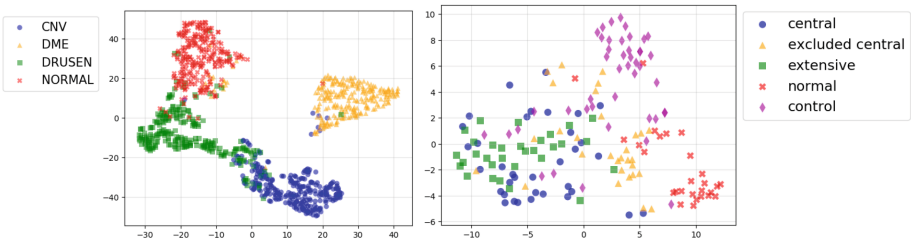


Fig. 4. Relationship of categories in the target dataset

dataset contains four categories including Normal, Choroidal Neo Vascularisation (CNV), Diabetic Macular (DME) and Drusen. They have a total of 109,309 samples, of which 1,000 are used for testing and the rest are used for training. There are two kinds of sizes in those images, $1536 \times 496 \times 1$ and $1024 \times 496 \times 1$. For the experiments, they are all preprocessed and resized to $224 \times 224 \times 1$. The BOE dataset have smaller size than the Cell one. It works for target testing, which acquired from 45 patients. 15 normal patients, 15 patients with dry AMD, and 15 patients with DME.

Target Dataset and Data Acquisition. In the target dataset as shown in Fig. 4, we have 1128 samples from 60 patients' 94 eyes, of which 236 are central IRDs, 204 are excluded central IRDs, 209 are extensive IRDs, 185 are normal and 294 are control samples (CNV, DME, MH, ERM...). The size of the images is $1180 \times 786 \times 1$. Extracted macular OCTs containing at least one OCT scan providing a cross section of the fovea were included in this study. The B scan OCT images with evidence of retinal disease as determined by two retinal specialists were defined as controls. For the experiments, they are all preprocessed and resized to $224 \times 224 \times 1$ because of the limitation of hardware. The ratio of training, testing and validation is 0.70/0.15/0.15. The data were collected from Beijing Tongren Eye Center with a clinical diagnosis of IRDs involving the macular area were included in the current study. SD-OCT data were acquired using a Cirrus HD-OCT 5000 system (Carl Zeiss Meditec Inc., Dublin, CA, USA). This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Beijing Tongren Eye Center. (No.TRECKY2017-10, Mar.3,2017).



(a) Auxiliary Dataset (b) Target Dataset

Fig. 5. Feature space representation of teacher (processed by T-SNE). After transfer learning, the teacher model still has excellent clustering and fitting ability in the high-dimensional feature space in the target feature space. The red crosses (normal) appearing in both figures and the purple diamonds (containing CNV, DME, etc.) in the right figure still have good clustering performance. (Color figure online)

3.2 Experimental Settings

Baseline and Data Applicability. It has been shown that the core problems of Few-Shot Learning (FSL) in supervised machine learning are Empirical Risk Minimization and Unreliable Empirical Risk Minimizer [24]. To alleviate these, we usually start with three aspects: data, models, and algorithms. For data, we purposefully design the preprocessing pipeline. We conducted baseline experiments on the state-of-the-art Simple Siamese (SimSiam) network [2]. As shown in Table 1: our preprocessing pipeline mitigates the impact on the model itself due to noise diversity. Also, normalizing this allows the model to exclude redundant concerns.

Feature Space Representation. The SNNL loss function enables the model to get a better projection of the input image during training in the designed feature space, which means that inter-class samples can be clustered while intra-class samples can be separated by the distance function. From Fig. 5, we can see that when the Teacher model (ResNet-50) is trained by the auxiliary dataset, it has the ability to project the test samples from the auxiliary dataset. Meanwhile, to the target dataset, the Teacher also can cluster the normal class and control class which becomes the control class in the target dataset before fine-tuning and transfer learning.

Table 1. Baseline accuracy(%)

Methods	Target dataset	
	Raw	Preprocessed
SimSiam	58.26 ± 1.59	60.5 ± 1.40

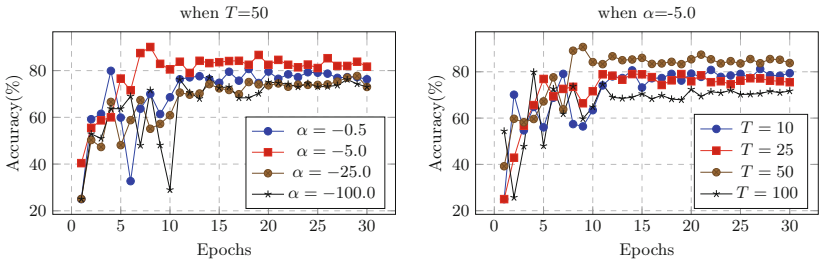


Fig. 6. Teacher model (ResNet-50) Test Experiments

3.3 Student-Teacher Learning

Teacher Model. We choose the ResNet-50 as Teacher Model to handle the auxiliary dataset. The performance is mainly controlled by the hyper-parameter α and Temperature T in Eqs. 1 and 2. In our experiment, we fix the dimension of Feature Space as 128 and pretrain the model with 30 epochs. We decrease the learning rate at epoch 10 and 25 with a factor 0.1. There are two sets of control trials in Fig. 6 optimizing two hyper-parameters α and T. We fix one of them respectively at a time, and optimize the other to get the best accuracy. The best performance is achieved with $T = 50$, and $\alpha = -5.0$.

After training the projector network with the auxiliary dataset by SNNL loss, we perform three different forms of fine-tuning to the target dataset. In the "Features Extraction" way, we freeze the parameters before the last fully-connected layer and replace with a new classifier. In the "High-level" way, we freeze the parameters before the 5th group of convolution layers (the 129_{th} layer in the ResNet-50), left the last group of convolution to learn the high-level features from the new target dataset. In the way of "All Parameters", the model can adjust all the parameters included in ResNet-50. From the data in Table 2, we pick the one with best performance to play the teacher role in the S-T architecture.

Student Model. After accomplished the transfer learning and fine-tuning of Teacher model, We use the ResNet-18 as the Student Model to adapt the smaller size of target data. The ResNet-18 is totally untrained by any data before the S-T learning. From Table 2, we can see that the student model in our trained S-T architecture gets better results than the teacher. This is attributed to the student incorporates knowledge from the teacher's pre-training and information from the hard-label classification. By adjusting the valves of knowledge from both sides, we are able to determine the proportion of prior knowledge that the student model receives from the teacher model versus the feedback from the labels, by adapting this to different data sets. Too much prior knowledge may be counterproductive if the difference between the datasets is greater.

Table 2. Test accuracy (%) comparison of common FSL methods. All the methods based on the pre-trained network with Cell dataset. *Features Extraction* means freezing the pre-trained ResNet-50 and replacing the last layer with a 3-layer classifier. *All Parameters* means the whole ResNet-50 involved in the target-oriented training. *High-level* means only the tail fully connected layers in ResNet-50 involved in the training with the target dataset.

Dataset	Methods				
	Features Extraction	All Parameters	High-level	SimSiam [2]	S-T Learning (ours)
Target(5 Classes)	53.91 ± 1.79	57.17 ± 2.11	59.68 ± 2.59	61.42 ± 2.18	$\mathbf{74.45 \pm 1.59}$
			Teacher[a]		
BOE(3 Classes)[b]	72.11 ± 0.62	97.83 ± 1.55	92.87 ± 1.96	76.82 ± 1.93	$\mathbf{99.69 \pm 0.10}$
		Teacher[a]			

[a] chosen as the teacher model in the S-T learning architecture.
[b] BOE dataset has more duplicate labels with the Cell dataset compared to the target dataset. Therefore, it outperforms the target dataset under the same training and network conditions.

4 Conclusion

In this study, we demonstrate a Student-Teacher Learning based classification model on a small dataset to distinguish several retinal diseases. This framework learns the knowledge from both ground truth labels and pretrained Teacher model to make it possible to handle limited data. Data preprocessing also plays a critical role that cannot be ignored before training.

Acknowledgments. This research is partially funded by NSF-IIS-1910492.

References

1. Anantrasirichai, N., Achim, A., Morgan, J.E., Erchova, I., Nicholson, L.: SVM-based texture classification in Optical Coherence Tomography. In: 2013 IEEE 10th International Symposium on Biomedical Imaging, San Francisco, CA, USA, pp. 1332–1335. IEEE (2013)
2. Chen, X., He, K.: Exploring simple Siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15750–15758 (2021)
3. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Trans. Image Proces. **16**(8), 2080–2095 (2007)
4. Fang, L., Wang, C., Li, S., Yan, J., Chen, X., Rabbani, H.: Automatic classification of retinal three-dimensional optical coherence tomography images using principal component analysis network with composite kernels. J. Biomed. Opt. **22**(11), 116011 (2017)

5. Frosst, N., Papernot, N., Hinton, G.: Analyzing and improving representations with the soft nearest neighbor loss. In: International Conference on Machine Learning, pp. 2012–2020. PMLR (2019)
6. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. stat 1050, 9 (2015)
7. Hussain, M.A., et al.: Classification of healthy and diseased retina using SD-OCT imaging and random forest algorithm. PloS One **13**(6), e0198281 (2018)
8. Ji, Q., He, W., Huang, J., Sun, Y.: Efficient deep learning-based automated pathology identification in retinal optical coherence tomography images. Algorithms **11**(6), 88 (2018)
9. Ji, Q., Huang, J., He, W., Sun, Y.: Optimized deep convolutional neural networks for identification of macular diseases from optical coherence tomography images. Algorithms **12**(3), 51 (2019)
10. Karri, S.P.K., Chakraborty, D., Chatterjee, J.: Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration. Biomed. Opt. Express **8**(2), 579–592 (2017)
11. Kermany, D., Zhang, K., Goldbaum, M., et al.: Labeled optical coherence tomography (OCT) and chest X-Ray images for classification (2018)
12. Liu, Y.Y., Chen, M., Ishikawa, H., Wollstein, G., Schuman, J.S., Rehg, J.M.: Automated macular pathology diagnosis in retinal OCT images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding. Med. Image Anal. **15**(5), 748–759 (2011)
13. Lu, W., Tong, Y., Yu, Y., Xing, Y., Chen, C., Shen, Y.: Deep learning-based automated classification of multi-categorical abnormalities from optical coherence tomography images. Transl. Vis. Sci. Technol. **7**(6), 41 (2018)
14. Motozawa, N., et al.: Optical coherence tomography-based deep-learning models for classifying normal and age-related macular degeneration and exudative and non-exudative age-related macular degeneration changes. Ophthalmol. Ther. **8**(4), 527–539 (2019)
15. Sankar, S., et al.: Classification of SD-OCT volumes for DME detection: an anomaly detection approach. In: Medical Imaging 2016: Computer-Aided Diagnosis, vol. 9785, pp. 97852O. International Society for Optics and Photonics (2016)
16. Shih, F.Y., Patel, H.: Deep learning classification on optical coherence tomography retina images. Int. J. Pattern Recogn. Artif. Intell. **34**(08), 2052002 (2020)
17. Srinivasan, P.P., et al.: Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. Biomed. Opt. Express **5**(10), 3568–3577 (2014)
18. Sugmk, J., Kiattisin, S., Leelasantitham, A.: Automated classification between age-related macular degeneration and Diabetic macular edema in OCT image using image segmentation. In: The 7th 2014 Biomedical Engineering International Conference, pp. 1–4. IEEE (2014)
19. Sun, Y., Li, S., Sun, Z.: Fully automated macular pathology detection in retina optical coherence tomography images using sparse coding and dictionary learning. J. Biomed. Opt. **22**(1), 016012 (2017)
20. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: International Conference on Learning Representations (2019)
21. Venhuizen, F.G., et al.: Automated age-related macular degeneration classification in OCT using unsupervised feature learning. In: Hadjiiski, L.M., Tourassi, G.D. (eds.) SPIE Medical Imaging, Orlando, Florida, USA, pp. 94141I (2015)
22. Wang, J., et al.: Deep learning for quality assessment of retinal OCT images. Biomed. Opt. Express **10**(12), 6057–6072 (2019)

23. Wang, L., Yoon, K.-J.: Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks. IEEE Trans. Pattern Anal. Mach. Intell., 1 (2021). https://doi.org/10.1109/TPAMI.2021.3055564
24. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: a survey on few-shot learning. ACM Comput. Surv. (CSUR) **53**(3), 1–34 (2020)