2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-2923-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISBI52829.2022.9761676

AMD CLASSIFICATION BASED ON ADVERSARIAL DOMAIN ADAPTATION WITH CENTER LOSS

Shengzhu Yang[1] *Xi Zhang*[2] *He Zhao*[1,2,⋆] *Huiqi Li*[1,2,⋆] *Hanruo Liu*[3] *Ningli Wang*[3]

[1] Institute of Engineering Medicine, Beijing Institute of Technology, China
[2] School of Information and Electronics, Beijing Institute of Technology, China
[3] Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Capital Medical University, China

ABSTRACT

In this paper, we present a deep learning approach for automatic categorization of age-related macular degeneration (AMD). Faced with the deficiency of training data, we propose a solution to combine additional data to effectively assist the classification task. During training process, the retinal fundus images from two datasets are mapped into a common feature space with adversarial domain adaptation to reduce domain discrepancy. Moreover, we introduce center loss to increase the intra-class compactness of the extracted features to further improve the classification performance. Experiments are conducted on three public fundus image datasets: STARE, ODIR and iCHALLENGE-AMD (hereinafter referred to as iAMD). Our method outperforms three state-of-the-art classification models as well as other augmentation approaches. The proposed approach provides a general framework to handle the issue of training samples with domain discrepancy.

Index Terms— AMD classification, adversarial domain adaptation, discriminative features

1. INTRODUCTION

Age-related macular degeneration (AMD) is a common macular disease that is a leading cause of severe vision impairment or even vision loss among people over the age of 50 [1], which can be categorized into two types: dry AMD and wet AMD (neovascular). In recent years, deep learning techniques have been widely employed by research communities to detect early symptoms of AMD [2]. However, these methods all face the challenge of acquiring sufficient data sets with AMD labels. Under such circumstances, it is reasonable to utilize the image data from multiple datasets for training. But the domain discrepancy between them hinders the adaptation of predictive models across domains.

To tackle this issue, we propose a unified framework for AMD classification inspired by domain adaptation. Our approach takes advantage of multiple training datasets by considering the discrepancy across domains. In addition, the label information is also leveraged to constrain the discriminative feature mapping. Our contributions are summarized as follows:

- A new training scheme is proposed for AMD classification, which takes the advantages of existing source datasets in training new target dataset. It helps to solve the problem of classification algorithms performing poorly on a small dataset.
- We propose an efficient loss combination that consists of adversarial loss for reducing cross-domain discrepancy and center loss for generating the discriminative features.
- Our approach is extensively evaluated on three datasets and the results show that our method surpasses the state-of-the-art algorithms for AMD classification.

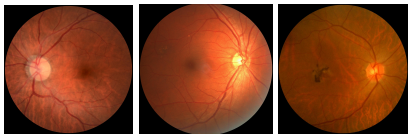
2. RELATED WORK

Deep learning has been proven effective on many public data sets, which has also been investigated in AMD classification. Burlina et al. [3] apply AlexNet to the automatic screening of AMD. Peng et al. [4] also present a deep learning model called DeepSeeNet as a simulation of the human grading of AMD severity. Domain adaptation aims to transfer the representations from a source dataset to a target dataset to minimize their feature distributions. Maximum Mean Discrepancy (MMD) measures the distance of two distributions in the regenerated Hilbert space, and it is applied to the latent representations in Deep Adaptation Network (DAN) [5]. Other methods have taken advantage of adversarial losses for domain adaptation [6, 7, 8].

3. METHOD

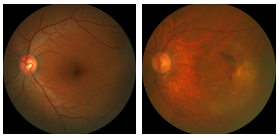
In this paper, we propose a method to classify AMD images based on adversarial networks with center loss constraint. Denote the target dataset as $\mathcal{D}^t = \left\{x_t^i, y_t^i\right\}_i^{N_t}$, where x_t^i is the image and y_t^i is the corresponding label. Our goal is to learn a model with the help of source dataset $\mathcal{D}^s = \left\{x_s^j, y_s^j\right\}_j^{N_s}$ such that the task obtains a higher accuracy when at test time. The

⋆Corresponding authors: He Zhao (zhaohe@bit.edu.cn); Huiqi Li (huiqili@bit.edu.cn).



(a) Normal, dry and wet AMD images in ODIR dataset.

(b) Normal, dry and wet AMD images in STARE dataset.



(c) Normal and AMD images in iCHALLENGE-AMD dataset.

Fig. 1: Classifications from different datasets.

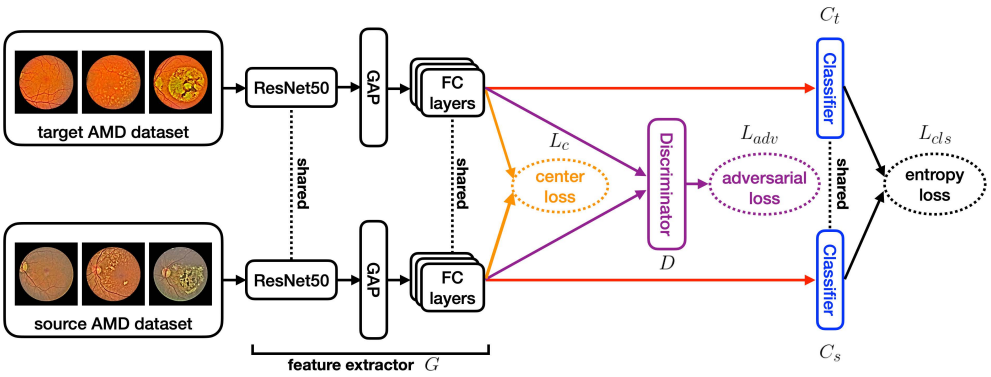


Fig. 2: An overview of the proposed model structure. There are two shared feature extractors G to learn the representations of the target and source dataset respectively. D denotes the discriminator. C_t and C_S with shared weights indicate the classifiers that make a three-class classification with the extracted features from G During testing, only the upper branch of the model with fixed parameters is kept.

network architecture is shown in Fig.2. A siamese-like feature extractor G is proposed to learn the useful information from both $\mathcal{D}^t$ and $\mathcal{D}^s$. A domain discriminator D is utilized to narrow the feature distribution between the two datasets. To further improve the performance, the center loss is also considered to pull features belonging to the same class closer to generate discriminative features. In what follows, we will introduce the details of our model.

3.1. Model Structure

Our Model includes three parts: feature extractor, classifier and discriminator. The feature extractor consists of three components including the backbone network ResNet-50, global average pooling (GAP) layer and fully connected (FC) layers, which finally leads the input image to a low dimension feature vector $\boldsymbol{f}$. The dimension reduction of the features can be beneficial to the follow-up progresses which are used for feature regularization. As to the classifier, it gives out the prediction of the input image with FC layers. The discriminator on the latent space, composed of three fully connected layers, is designed to alleviate the discrepancy of different data domains. It aims to distinguish the feature $\boldsymbol{f_s}$ from source dataset and feature $\boldsymbol{f_t}$ from target dataset. The discriminator updates alternatively with the feature extractor and classifier to make the training data with domain bias can be used to train the same classifier after reducing their feature gap.

3.2. Loss Function

Our loss function consists of three parts: classification loss, adversarial loss for reducing feature discrepancy between domains, and center loss for generating discriminative features.

Adversarial Loss. To minimize the discrepancy of target and source feature representation, the feature extractor and domain discriminator play a minimax game. Based on the feedback of the discriminator D, the feature extractor G is updated with formulation given as:

$$L_{adv,G} = \frac{1}{2}\mathbb{E}_{x_s \sim \mathcal{D}^s} \left(D\left(G\left(x_s\right)\right) - 1\right)^2 \tag{1}$$

where $G\left(x_s\right)$ denotes the latent features of $\mathcal{D}^s$. In this way, the source data can generate a similar feature distribution to the target domain and it can be used to train a robust classifier for the target domain.

Classification Loss. The classification loss L_{cls} appears in the form of cross-entropy loss which is effective for multi-classification tasks. Since the weights for target and source classifiers are tied, their classification losses are formulated in the same way which are expressed as:

$$L_{cls} = -\sum_{k=1}^{C} \mathbb{I}(\mathrm{y} == \mathrm{k}) \log(p) \tag{2}$$

where C denotes the number of categories, p is the probability that a sample belongs to category k predicted by the

model, and $\mathbb{I}$ is the indicator function with value of one when the condition is satisfied. Considering both target and source datasets, the overall classification loss is expressed as:

$$L_{cls} = L_{cls,t} + \gamma L_{cls,s} \tag{3}$$

where L_{cls_t}, and L_{cls_s} stand for the loss for target domain and source domain, respectively; γ is the weight for source domain, while target domain weight is set to one by default.

Center Loss. The center loss L_c minimizes the intra-class distances of the deep features, which clusters the features from same class. Considering the center loss on target and source domain separately, the formulations can be written as:

$$L_{c,t} = \frac{1}{2}\sum_{i=1}^{\mathrm{N_t}} \left\| \boldsymbol{f}_{t,i} - \boldsymbol{c}_{y_{t,i}} \right\|_2^2, L_{c,s} = \frac{1}{2}\sum_{j=1}^{\mathrm{N_s}} \left\| \boldsymbol{f}_{s,j} - \boldsymbol{c}_{y_{s,j}} \right\|_2^2 \tag{4}$$

where $\boldsymbol{f_{t,i}}$ and $\boldsymbol{f_{s,i}}$ represents the features of i-th sample extracted by CNN, $\boldsymbol{c}_{y_{t,i}}$ and $\boldsymbol{c}_{y_{s,j}}$ denotes the y_i-th and the y_j-th class center of the deep feature, N denotes the number of samples. With the variety of features in the training process, $\boldsymbol{c}_y$ is updated following [9]. And the final center loss is given as follows:

$$L_c = \beta_t L_{c,t} + \beta_s L_{c,s} \tag{5}$$

where β_t and β_s are weights for target and source domain respectively. The overall objective function is summarized as:

$$L = \alpha L_{adv,G} + L_c + L_{cls} \tag{6}$$

where α is the weight for adversarial loss. Specifically, the center loss and the classification loss are computed in both the target and source domain.

4. EXPERIMENTS

4.1. Experimental Setup

We validate our proposed adversarial adaptation method among STARE [10] and ODIR [11] datasets with AMD labels of three categories, which are dry AMD, wet AMD and normal, and iAMD [12] dataset with two categories which are AMD and normal. As a result, 55/55 images are used for training and test on STARE, while 74/58 images are used on ODIR dataset, 79/78 images are used on iAMD dataset.

The fundus images are preprocessed with normalization, surrounding back area removal and resizing to an image with size of 224×224. The backbone network of the feature extractor is ResNet-50 pre-trained on ImageNet. The results are the mean accuracy of five-time experiments.

4.2. Experimental Results

We evaluate our AMD classification model across three domain shifts, where STARE, ODIR and iAMD act as target dataset in turn. Classification accuracy and kappa coefficient are used to measure the performance of different approaches. Two training scenarios are considered here, which are training models on single dataset and on mixed dataset, respectively. The results are shown in Table 1 (three-category classification) and Table 2 (two-category classification). Note that other methods (e.g., ResNet-50) are pre-trained on ImageNet and trained with two datasets together without any adaptation. By simply mixing the training data of different datasets, the performance sometimes will decrease or hardly improve. On the contrary, our approach achieves the best accuracy by using source data in an efficient way. The kappa coefficient, as an indicator for checking the consistency of the classification model, has the same trend with the accuracy. It demonstrates that the predictions are consistent with the actual classification results. The above results justify the effectiveness of our algorithm.

4.3. Ablation Experiments

Here we will demonstrate the contribution of the adversarial domain adaptation and center loss. The influence of the weights on loss function will be testified. The influence of the weights on loss function will also be testified. Note that the STARE dataset is treated as target domain and ODIR is used as the source in the three-category classification experiments.

Different Composition of Training Sets

Data augmentation and data mixture are other ways to enlarge training dataset. In this part, we will compare our approach with these two kinds of methods. Four different training sets are used for comparison, which are 1) original dataset, 2) the mixture of two datasets, 3) augmented original dataset, 4) Our approach. The models are trained on them separately and tested on the same data. The results are shown in Table. 3. It demonstrate the operations of augmentation and mixture are able to improve the performance by increasing the datasets, while our approach obtains a large improvement by the discrepancy reduction among datasets.

Efficacy of Adversarial Adaptation and Center Loss

To evaluate the efficiencies of the proposed losses, two sets of ablation studies are conducted. The weight for classification loss $L_{cls,t}$ is fixed to one in the following experiments. We evaluate the adversarial loss weight α with a value range of $[0.05, 1]$. The classification accuracy can achieve 92.12% when the adversarial loss weight is set to 0.1, which is much higher than the baseline. Based on the effect of adversarial loss with weight of 0.1, the center loss on target dataset $L_{c,t}$ is then considered with a weight range from 0.5 to 10. The highest accuracy is 93.33% when $\gamma = 10$, which is approximately 1.2% percent larger compared with the situation where only $L_{adv,G}$ exists. The experimental results presented above verify that the adversarial domain adaptation and center loss both contribute to improve the classification performance.

Table 1: Three-category classification performance of different methods evaluated by accuracy and kappa.

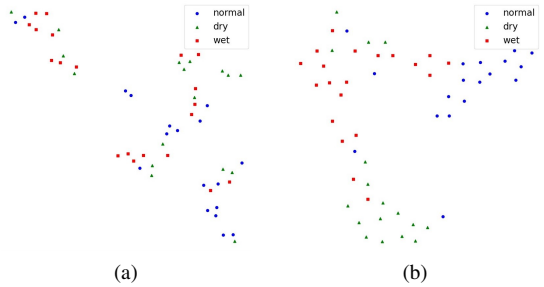
	STARE				ODIR			
	Trained on STARE		Trained on STARE + ODIR		Trained on ODIR		Trained on ODIR + STARE	
Model	Acc	κ	Acc	κ	Acc	κ	Acc	κ
ResNet-50	86.67%	0.80	94.55%	0.93	89.09%	0.85	89.66%	0.84
VGG-19	76.36%	0.64	70.30%	0.55	72.99%	0.57	68.39%	0.50
EfficientNet-b5	87.88%	0.82	95.09%	0.93	86.21%	0.77	89.08%	0.83
Our approach	-	-	**96.97%**	**0.95**	-	-	**92.53%**	**0.88**

Table 2: Two-category classification performance of different methods evaluated by accuracy and kappa.

	STARE					
	Trained on STARE		Trained on STARE + ODIR		Trained on STARE + iAMD	
Model	Acc	κ	Acc	κ	Acc	κ
ResNet-50	98.18%	0.96	97.45%	0.94	98.55%	0.97
VGG-19	98.55%	0.97	97.09%	0.94	92.73%	0.83
EfficientNet-b5	96.73%	0.93	89.09%	0.73	97.09%	0.93
Our approach	-	-	**99.64%**	**0.99**	**98.91%**	**0.97**
	ODIR					
	Trained on ODIR		Trained on ODIR + STARE		Trained on ODIR + iAMD	
Model	Acc	κ	Acc	κ	Acc	κ
ResNet-50	94.14%	0.88	78.62%	0.55	91.03%	0.82
VGG-19	89.31%	0.78	93.79%	0.66	85.52%	0.71
EfficientNet-b5	90.00%	0.79	83.10%	0.65	86.90%	0.73
Our approach	-	-	**95.52%**	**0.91**	**95.17%**	**0.90**
	iAMD					
	Trained on iAMD		Trained on iAMD + ODIR		Trained on iAMD + STARE	
Model	Acc	κ	Acc	κ	Acc	κ
ResNet-50	80.00%	0.60	79.23%	0.58	80.00%	0.60
VGG-19	79.23%	0.58	79.23%	0.58	79.49%	0.59
EfficientNet-b5	80.51%	0.61	78.72%	0.57	78.21%	0.56
Our approach	-	-	**82.05%**	**0.64**	**85.90%**	**0.72**

Table 3: Three-category classification performance of baseline network ResNet-50 trained with different composition of training sets compared to our approach. Ori.: Original dataset; Aug.: Augmentation of the original dataset; Mix.: Mixture of two datasets.

	STARE		ODIR	
Training set	Acc	κ	Acc	κ
Ori.	86.67%	0.80	89.09%	0.85
Aug.	92.12%	0.88	91.38%	0.87
Mix.	94.55%	0.92	89.66%	0.86
Our approach	**96.97%**	**0.95**	**92.53%**	**0.88**



(a) (b)

Fig. 3: **Visualization of the extracted features via t-SNE.** (a) Visualization of features generated by baseline model (ResNet-50). (b) Visualization of features generated by our model.

4.4. Visualizations of Feature Distributions

In this section, we aim to justify that the feature extractor trained by our proposed approach is more efficacious than the baseline. The distribution of the extracted features from test sets is visualized through t-SNE [13]. Fig.3(a) and Fig.3(b) exhibit the representations of STARE test images in the latent space when the network is trained by ResNet-50 and our method respectively. The classification accuracies reach 87.273% and 96.364% in the two cases, individually. It can be discovered from the comparison that the features of the same class tend to be more clustered while features belonging to different classes are more divergent for our approach. The aforementioned experiment demonstrates that our method learns a more effective representation of AMD data in feature space than baseline model.

5. CONCLUSION

We propose a novel deep learning framework for the AMD classification task based on adversarial domain adaptation and center loss. To solve the dilemma of insufficient training data, we introduce an additional dataset as the source for assisting the classification of the target dataset. The images from the multi-datasets are mapped into a shared feature space where adversarial domain discriminator is employed to minimize their domain discrepancy. Center loss is adopted to extract more discriminative features for classification. According to the experimental results, the proposed approach outperforms state-of-the-art classification methods, and it achieves higher classification accuracies by contrast with other data augmentation ways. The proposed framework can be further extended to other classification applications.

6. COMPLIANCE WITH ETHICAL STANDARDS

All data comes from public databases.

7. ACKNOWLEDGMENTS

This work is supported by China Postdoctoral Science Foundation (No. 2020M680387) and National Natural Science Foundation of China (No. 82072007).

8. REFERENCES

[1] Laurence S Lim, Paul Mitchell, Johanna M Seddon, Frank G Holz, and Tien Y Wong, "Age-related macular degeneration," *The Lancet*, vol. 379, no. 9827, pp. 1728–1738, 2012.

[2] Philippe Burlina, Katia D Pacheco, Neil Joshi, David E Freund, and Neil M Bressler, "Comparing humans and deep learning performance for grading amd: a study in using universal deep features and transfer learning for automated amd analysis," *Computers in biology and medicine*, vol. 82, pp. 80–86, 2017.

[3] Philippe M. Burlina, Neil Joshi, Michael Pekala, Katia D. Pacheco, David E. Freund, and Neil 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, 11 2017.

[4] Yifan Peng, Shazia Dharssi, Qingyu Chen, Tiarnan D. Keenan, Elvira Agrón, Wai T. Wong, Emily Y. Chew, and Zhiyong Lu, "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.

[5] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan, "Learning transferable features with deep adaptation networks," in *Proceedings of the 32nd International Conference on Machine Learning*, Francis Bach and David Blei, Eds., Lille, France, 07-09 Jul 2015, vol. 37 of *Proceedings of Machine Learning Research*, pp. 97–105, PMLR.

[6] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, Eds., Cham, 2016, pp. 597–613, Springer International Publishing.

[7] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] He Zhao, Qingqing Zheng, Kai Ma, Huiqi Li, and Yefeng Zheng, "Deep representation-based domain adaptation for nonstationary eeg classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 535–545, 2020.

[9] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, Eds., Cham, 2016, pp. 499–515, Springer International Publishing.

[10] A.D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical Imaging*, vol. 19, no. 3, pp. 203–210, 2000.

[11] Neha Gour and Pritee Khanna, "Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network," *Biomedical Signal Processing and Control*, vol. 66, pp. 102329, 2021.

[12] Baidu Research open-access dataset, "iChallenge-AMD dataset," http://ai.baidu.com/broad/subordinate?dataset=amd.

[13] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.