A Domain Adaptation Multi-instance Learning for Diabetic Retinopathy Grading on Retinal Images

Ruoxian Song
College of Computer Science and Engineering
Northeastern University
Shenyang, China
songruoxian@stumail.neu.edu.cn

Peng Cao*
College of Computer Science and Engineering, Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education
Northeastern University
Shenyang, China
caopeng@cse.neu.edu.cn

Jinzhu Yang
College of Computer Science and Engineering
Northeastern University
Shenyang, China
yangjinzhu@cse.neu.edu.cn

Dazhe Zhao
College of Computer Science and Engineering
Northeastern University
Shenyang, China
zhaodz@neusoft.com

Osmar R. Zaiane
Alberta Machine Intelligence Institute
University of Alberta
Edmonton, Canada
zaiane@ualberta.ca

***Abstract*—Diabetic retinopathy (DR) is one of the most concerning, common and serious diseases in the ophthalmology community. Early detection and treatment of DR can significantly reduce the risk of vision loss in patients. Traditional DR automatic classification algorithms rely on the precise detection of microaneurysms (MA) and hemorrhage (H) lesions. Such lesion annotation is an expensive and time-consuming process, hence it is expected to develop automatic grading methods with only image-level annotations. The lack of the position of MA and H hinders the traditional supervised algorithms for the accurate identification. In our work, we formulate the weakly supervised DR grading as a multi-instance learning problem, and propose a domain adaptation multi-instance learning with attention mechanism for DR grading. Specifically, labeled instances are generated by cross-domain to filter irrelevant instances in the target domain. To model the relationship between the suspicious instances and bag label, a multi-instance learning with attention mechanism is developed to acquire the location information of highly suspected lesions and predict the grade of DR. We evaluate our proposed algorithm on the Messidor dataset, and the experimental results demonstrate that it achieves an average accuracy of 0.764 and an AUC value of 0.749 respectively, outperforming state-of-the-art approaches.**

***Index Terms*—Diabetic retinopathy, Severity level grading, Multi-instance learning, Domain adaptation, Attention**

I. Introduction

Diabetic retinopathy (DR) is one of the diabetes-derived diseases and is the main cause of blindness in both developed and developing countries. The International Diabetes Federation (IDF) has reported that the global prevalence of diabetes in 2019 is estimated to be 9.3% (463 million people), will rise to 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045 [1]. Early screening and treatment of DR can reduce the risk of severe vision loss by more than 98% [2]. Moreover, grading of severity level is a vital activity in DR screening to diagnose retinal diseases. It is an intensive procedure that requires a trained workforce to identify and confirm cases as having pathology abnormality or not as well as differentiate the levels of pathology. Therefore, a computer-aided diagnosis (CAD) system based on retinal fundus images would be an efficient and effective method for early DR diagnosis and assisting experts.

The task of DR graded diagnosis can be regarded as a DR multi-classification problem. Microaneurysms (MA) are regarded as early signs of DR, and as the degree of DR advances hemorrhages (H) become evident in Fig. 1. The identification of HMAs (MA and H) is an important measure of progression of retinopathy in the early stage and may serve as a surrogate end point for severe change in some clinical trials. Therefore, the diagnosis and grading performance of DR highly depend on automatic detection models trained with sufficient labeled HMA instances [3]. However, manual labeling of the images is an expensive and/or time-consuming process, resulting in the lacking of labeled information of suspicious lesions. The issue hinders the traditional supervised algorithms to identify the true positive HMA. That is a weakly supervised learning problem [4]. The lack of labeled data motivates approaches that go beyond traditional supervised learning by incorporating other data and labels that might be available. Multiple instance learning (MIL) is an extension of supervised learning that can train classifiers using such a weakly supervised learning problem. MIL provides a learning framework that allows weak supervision with only image-level labels and does not require detailed HMA location information. Specifically, a whole image that is uniformly split into many patches as instances, called bag, and a single class label is specified for all instances in a bag [5].

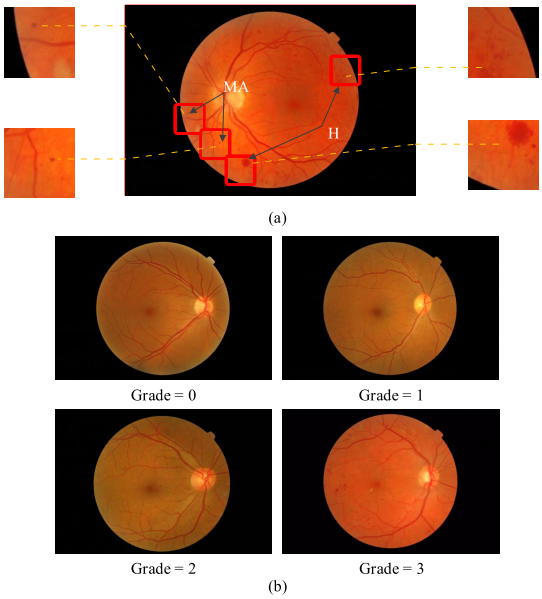


Fig. 1. (a) Early signs of DR, e.g., microaneurysms and hemorrhages in fundus images. (b) The increasing severity of DR.

In recent years, deep learning techniques, especially convolutional neural networks (CNNs), have made remarkable achievements in image classification [6]. Identifying suspicious region for medical images is of significant importance since it provides intuitive illustrations for physicians and patients of how the diagnosis is made [7]. However, MIL with deep learning lacks intuitive explanation and does not give any insight into the pathology. The inability to interpret the prediction model semantically is a well-known shortcoming of most existing computer-aided diagnosis methods.

In this paper, we present an unified network, namely **D**omain **A**daptation **M**ulti-**I**nstance **L**earning with **A**ttention mechanism (ADAMIL) for grading the level of disease. We have access to images with instance-level annotations in a source domain and images with image-level annotations in a target domain. Though generating artificial instances cross domains, an instance discriminator is trained to filter irrelevant instances of bags in the target domain. On the other hand, we incorporate an attention mechanism into a deep learning based MIL network to identify the highly suspicious instances and improve the overall classification performance. This model is able to predict the severity level of disease and also to score the importance of each patch in the input image in the final classification decision. The experimental results show that the proposed method achieves the best performance compared with several latest methods. Moreover, the proposed model has an ability to indicate the location of highly suspicious lesions in making the image-level decision.

In summary, our contributions are three folds:

- We propose a domain adaptation multi-instance learning framework for DR grading by exploiting weaker bag labels and instance labels from an auxiliary domain.
- We incorporate an attention mechanism into the proposed multi-instance learning framework to generate attention maps and discover meaningful patches which contain potential lesions in diabetic retinopathy.
- We conduct sufficient experiments to study our proposed method on the public Messidor dataset. We have also conducted extensive ablation studies, which can highlight the contribution of each key component of our proposed framework.

The rest of the paper is organized as follows. Section 2 describes our proposed method in details. Section 3 presents a comparison with state-of-the-art approaches using the Messidor dataset, investigates the influences of parameters, and discusses the experimental results. Section 4 presents the limitations and future directions of our work. At last, this paper is concluded in Section 5.

II. Proposed Approach

In this section, we describe the proposed ADAMIL algorithm for DR graded diagnosis in detail. The proposed algorithm takes color fundus images as input and outputs simultaneous classification of DR level and localization of highly suspicious lesion. Fig. 2 illustrates the proposed framework, which involves three components: domain adaptation, suspicious instance discriminator for filtering, multi-class multi-instance learning with attention mechanism.

Firstly, preprocessing operations such as slicing and data augmentation are performed. Then, domain adaptation is conducted by generating new instances across domains. With both the labeled source instances and the generated instances, an instance discriminator is trained by pre-training and fine-tuning to filter the irrelevant instances. Finally, the proposed multi-class multi-instance learning with an attention mechanism models the relationship between the embedding of the remained instances and the bag label. The details for each component are introduced in the following subsections.

A. Domain Adaptation

In practice, the lesions may be collected from multiple domains with different distributions. In order to mitigate the domain gap, we achieve a progressive domain adaptation by fine-tuning the instance discriminator with the artificially generated samples by CycleGAN [8]. The source domain is represented by $D_S = \{(\boldsymbol{x}_1^{\boldsymbol{s}}, y_1^s), \ldots, (\boldsymbol{x}_{N^s}^{\boldsymbol{s}}, y_{N^s}^s)\}$, the target domain is represented by $D_T = \left\{(\boldsymbol{X}_1^{\boldsymbol{t}}, y_1^t), \ldots, (\boldsymbol{X}_{N^t}^{\boldsymbol{t}}, y_{N^t}^t)\right\}$. In our work, the Messidor dataset is chosen as our target domain and the IDRiD dataset is chosen as our source domain. Owing to the different image sizes, the IDRiD and Messidor fundus images are resized to 1072×712 and 1440×960 without causing geometric distortion to reduce the influence of the image size on transfer learning (TL). Fig. 3 shows the structure and transfer process of the CycleGAN model. The model is composed of two generators G and F, two discriminators Q_s and Q_t, and provides two mapping functions $G : S \rightarrow T$ and $F : T \rightarrow S$. G and F complete the image translation from $D_S \rightarrow D_T$ and $D_T \rightarrow D_S$, respectively. The advantage of CycleGAN is that it allows each input image to

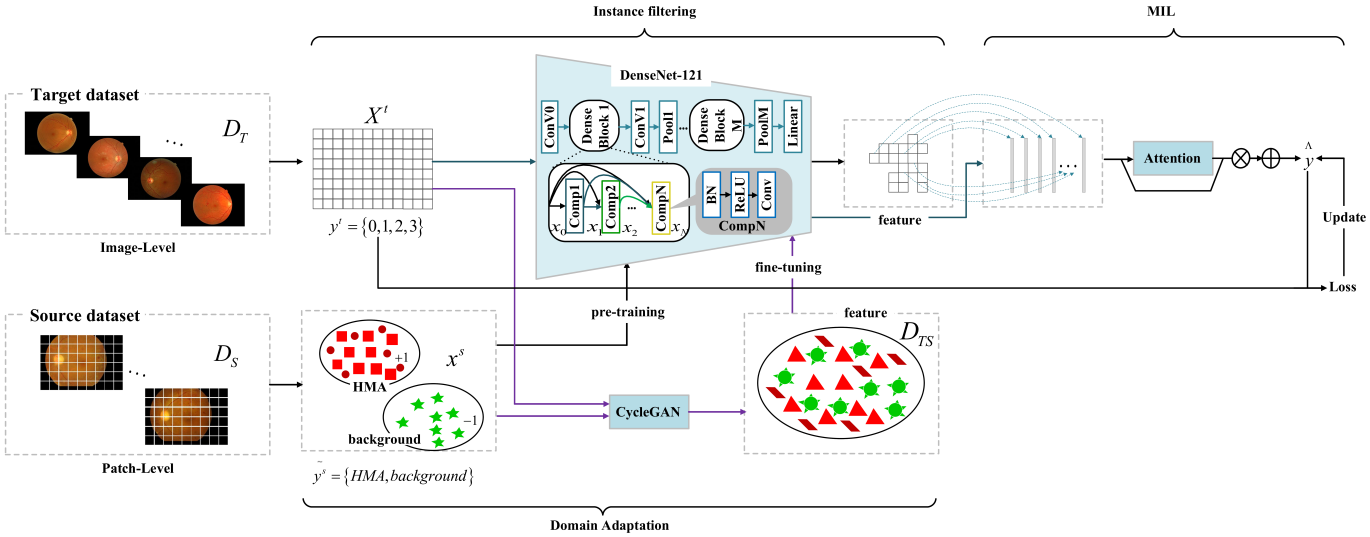


Fig. 2. An outline of our deep ADAMIL algorithm for DR multi-classification, where the DensNet-121 network structure contains $M = 4$ dense blocks, $N = \{12, 24, 48, 32\}$ composite layers and $k = 32$ growth rate.

be converted into a target domain and then be reconstructed to the original domain. Our goal is to learn a mapping function between D_S domain with $\{\boldsymbol{x}_i^s\}_{i=1}^{N^s}$ and D_T domain with $\boldsymbol{x}^t$ from $\{\boldsymbol{X}_i^t\}_{i=1}^{N^t}$, where $\boldsymbol{x}_i^s \in \boldsymbol{x}^s$ and $\boldsymbol{x}_j^t \in \boldsymbol{X}^t$. We denote the data distribution as $\boldsymbol{x}^s \sim p_{data}(\boldsymbol{x}^s)$ and $\boldsymbol{x}^t \sim p_{data}(\boldsymbol{x}^t)$, as well as the number of samples as N^s and N^T. Q_s aims to discriminate between image $\boldsymbol{x}^s$ and translated image $\{F(\boldsymbol{x}^t)\}$, and vice versa.

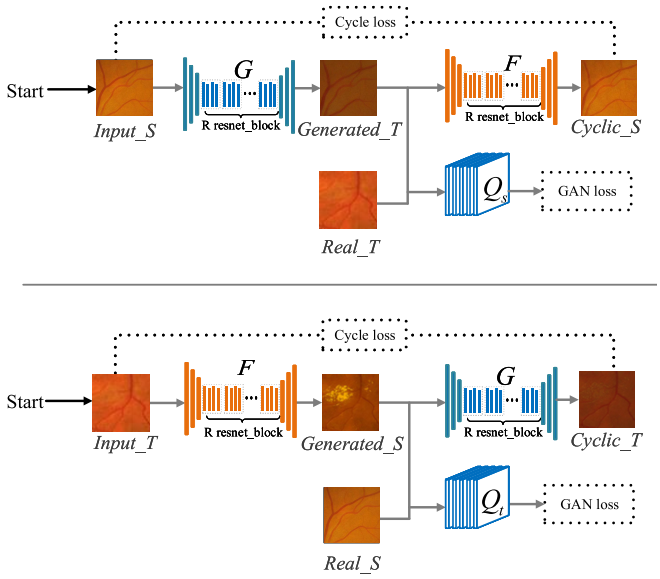


Fig. 3. The structure of CycleGAN model. G and F are different generators with the same network structure whereas Q_s and Q_t are also different discriminators with the same network structure. S and T respectively denote the source domain and the target domain. In our study, $R = 6$ ResNet blocks.

In the process of CycleGAN, there will be two types of losses: adversarial loss and cycle consistency loss. The purpose of adversarial loss is to match the distribution of the generated images with the data distribution in the target domain. The adversarial loss in the direction of $S \rightarrow T$ domain can be expressed as:

$$\begin{aligned}\mathcal{L}_{GAN}(G, Q_t, \boldsymbol{x}^s, \boldsymbol{x}^t) = &\mathbb{E}_{\boldsymbol{x}^t \sim p_{data}(\boldsymbol{x}^t)}\left[\log Q_t\left(\boldsymbol{x}^t\right)\right] \\ &+\mathbb{E}_{\boldsymbol{x}^s \sim p_{data}(\boldsymbol{x}^s)}\left[\log\left(1-Q_t(G(\boldsymbol{x}^s))\right)\right].\end{aligned} \tag{1}$$

The adversarial loss in the direction of $T \rightarrow S$ domain can be expressed as:

$$\begin{aligned}\mathcal{L}_{GAN}(F, Q_s, \boldsymbol{x}^t, \boldsymbol{x}^s) = &\mathbb{E}_{\boldsymbol{x}^s \sim p_{data}(\boldsymbol{x}^s)}\left[\log Q_s\left(\boldsymbol{x}^s\right)\right] \\ &+\mathbb{E}_{\boldsymbol{x}^t \sim p_{data}(\boldsymbol{x}^t)}\left[\log\left(1-Q_s(F(\boldsymbol{x}^t))\right)\right].\end{aligned} \tag{2}$$

When the amount of data is large, the cross-domain data will produce a variety of random permutations and may all achieve the same distribution constraint. Therefore, it is not enough to use the adversarial loss alone, and it cannot ensure that the generated image is unique. For this reason, a cycle consistent loss function is expressed as:

$$\begin{aligned}\mathcal{L}_{cyc}(G, F) = &\mathbb{E}_{\boldsymbol{x}^s \sim p_{data}(\boldsymbol{x}^s)}\left[||F\left(G\left(\boldsymbol{x}^s\right)\right) - \boldsymbol{x}^s||_1\right] \\ &+ \mathbb{E}_{\boldsymbol{x}^t \sim p_{data}(\boldsymbol{x}^t)}\left[||G\left(F\left(\boldsymbol{x}^t\right)\right) - \boldsymbol{x}^t||_1\right].\end{aligned} \tag{3}$$

The full objective of CycleGAN is:

$$\begin{aligned}\mathcal{L}\left(G, F, Q_s, Q_t\right) = &\mathcal{L}_{GAN}\left(G, Q_t, \boldsymbol{x}^s, \boldsymbol{x}^t\right) \\ &+ \mathcal{L}_{GAN}\left(F, Q_s, \boldsymbol{x}^t, \boldsymbol{x}^s\right) \\ &+ \lambda L_{cyc}\left(G, F\right),\end{aligned} \tag{4}$$

where $\lambda = 10$ controls the relative importance of cycle consistent loss and adversarial loss. We aim to solve:

$$G^*, F^* = \arg\min_{G,F}\max_{Q_s,Q_t}\mathcal{L}(G, F, Q_s, Q_t). \tag{5}$$

B. Instance Discriminator Training

Before classifying the entire retinal image, in order to reach better classification performance, we first attempt to filter out the irrelevant instances in the bag, in other words, detect its DR lesions. However, since the Messidor dataset only contains grades but no annotation information of HMA lesion location, we borrow the IDRiD dataset to construct an initial HMA lesion detection model.

DenseNet-121 was used in our proposed discriminator architecture, in which each layer was directly connected to every other layer in a feed-forward fashion. As shown in Fig 2, it consists of four dense blocks, three transition layers and a total of 121 layers (117-conv, 3-transition, and 1-classification). As described in the original DenseNet paper [9], each conv layer corresponds to a composite sequence of operations consisting of batch normalization (BN)-Relu-Conv. The Classification subnetwork includes 7×7 global average pooling, 1D fully-connected layer, and softmax. The ℓ^{th} layer receives the feature-maps of all preceding layers, $x_0, ..., x_{\ell-1}$, as input:

$$x_\ell = \boldsymbol{H}_\ell([x_0, x_1, ..., x_{\ell-1}]), \tag{6}$$

where $\boldsymbol{H}_\ell(\cdot)$ denotes the non-linear transformation function of the composite layers, and $[x_0, x_1, ..., x_{\ell-1}]$ denotes the concatenation of feature-maps generated in the $0, ..., \ell-1$ layers. If each function $\boldsymbol{H}_\ell$ generates k feature-maps, the ℓ^{th} layer has $k_0 + k \times (\ell - 1)$ input feature-maps, where k_0 is the number of channels in the initial input layer of each dense block. Specifically, we develop 4 dense blocks in our experiment, each dense block has {12,24,48,32} composite layers in turn, and a growth rate of 32 to predict the presence or absence of lesions.

With DenseNet, we build a two-steps instance discriminator training. First, we pre-train it with the labeled instances in the source domain. Second, we fine-tune it with the generated patches by CycleGAN.

C. Multi-class Multi-instance Learning with Attention Mechanism

We propose a multi-class multi-instance learning model with an attention mechanism, which can learn the local to global feature representation of each fundus image to implement a DR graded diagnosis. The process of multi-class multi-instance learning can be described as follows: Given a training dataset $\{(\boldsymbol{X}_1^t, y_1^t), (\boldsymbol{X}_2^t, y_2^t), ..., (\boldsymbol{X}_{\boldsymbol{N_t}}^t, y_{N_t}^t)\}$, where $\boldsymbol{X}_i^t$ is regarded as a bag, ${y_i}^t$ means bag-level label, and N_t is the number of training samples. A bag consists of multiple instances, namely $\boldsymbol{X}_i = \{\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}, ..., \boldsymbol{x}_i^{(n_i)}\}$, where n_i is the number of instances of $\boldsymbol{X}_i$, and each instance has no label.

With the embedding $\boldsymbol{h}_i^{(j)}$ of each remained instance in $\boldsymbol{X}_i$ learned by the instance discriminator, the attention weight of each instance is calculated as follow:

$$\boldsymbol{a}_i^{(j)} = \frac{\exp\left\{\boldsymbol{w}^T\left(\tanh\left(\boldsymbol{V}\left(\boldsymbol{h}_i^{(j)}\right)^T\right) \odot sigm\left(\boldsymbol{U}\left(\boldsymbol{h}_i^{(j)}\right)^T\right)\right)\right\}}{\sum_{k=1}^{n_i} \exp\left\{\boldsymbol{w}^T\left(\tanh\left(\boldsymbol{V}\left(\boldsymbol{h}_i^{(k)}\right)^T\right) \odot sigm\left(\boldsymbol{U}\left(\boldsymbol{h}_i^{(k)}\right)^T\right)\right)\right\}}, \tag{7}$$

where $\boldsymbol{w} \in \mathbb{R}^{LX1}$, $\boldsymbol{U} \in \mathbb{R}^{LXM}$ and $\boldsymbol{V} \in \mathbb{R}^{LXM}$ are parameters, $\odot$ is an element-wise multiplication and $sigm(\cdot)$ is the sigmoid non-linearity. The $\tanh(\cdot)$ non-linearity may not be effective in learning complex relationships. Therefore, it is proposed to use the gating mechanism [10] together with the $\tanh(\cdot)$ non-linearity to eliminate the troublesome linearity in $\tanh(\cdot)$. With the attention weight, the bag-level mapping relationship composed of weighted instances is expressed by:

$$\boldsymbol{z}_i = \left[\boldsymbol{a}_i^{(1)}\boldsymbol{h}_i{}^{(1)} \quad \boldsymbol{a}_i^{(2)}\boldsymbol{h}_i{}^{(2)} \quad \boldsymbol{a}_i^{(3)}\boldsymbol{h}_i{}^{(3)} \cdots \boldsymbol{a}_i^{(n_i)}\boldsymbol{h}_i^{(n_i)}\right]. \tag{8}$$

The label of each bag is converted to one-hot encoding vector. Let $N_{max} = \max_{i=1...N_t} n_i$ be the largest number of all the training bags. The weighted 2D instance-level is expanded to a tensor bag-level representation $\boldsymbol{z}_i \in \mathbb{R}^{N_{max} \times L \times P}$ by stacking multiple instances embedding, where P is the dimensionality of instance embedding. Finally, the bag-level prediction of the $L \times 1$ dimension is obtained with a tensor bag-level representation $\boldsymbol{z}_i$ by a FC layer combined with a softmax activation function. Fig. 4 shows the process of multi-classification in MIL.

III. Experiments

In this section, we introduce the experimental results of DR grading. We describe the datasets and performance metrics used, further verify the core components of the proposed ADAMIL algorithm, analyze the influence of hyperparameters, compare with other state-of-the-art methods, and verify its interpretability. All models are implemented with Python using the Pytorch framework. A computer with 8 NVIDIA GTX 1080TI and 128GB internal memory is applied to train and test.

A. Datasets and Performance Metrics

Experiments involve two publicly available dataset: IDRiD [11][1] and Messidor [12][2]. Ten-fold cross-validation is adopted to evaluate the proposed method, and accuracy, precision, recall, micro-F1, receiver operating characteristic (ROC) and area under the curve (AUC) of ROC are used as performance metrics. The contents of the two datasets are described as follows.

a) IDRiD Dataset: We employ the ISBI 2018 IDRiD subchallenge 1 dataset. This dataset consists of 516 color images in JPG format, captured by a KowaVX-10α digital fundus camera with 50° FOV, and has a resolution of 4288×2848. Among them, there are 81 color fundus images containing pixel-level lesion labeling information. Table I shows the number of images that exist for each type of lesion. The IDRiD dataset is annotated by the mapping relationship between the original image and the matched labeled image. Corresponding to the evaluation criteria of Messdior grade, only two major lesions of MA and H are enough.

[1]https://idrid.grand-challenge.org/Grading/

[2]http://www.adcis.net/en/third-party/messidor/

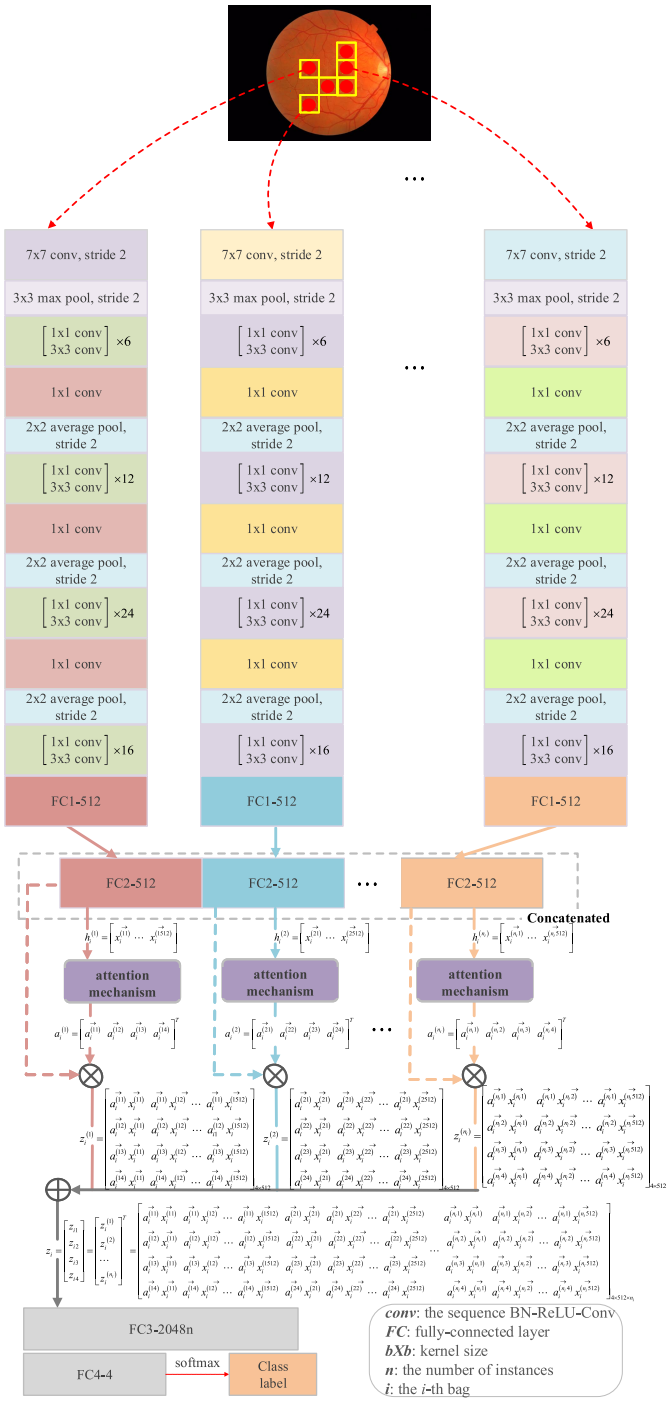


Fig. 4. Deep MIL network structure with attention mechanism

b) Messidor Dataset: This dataset contains 1200 color fundus images in TIF format acquired by three ophthalmology departments between 2005 and 2006. The grading result, ranging from 0 to 3 of each image, is provided. Table I shows the characteristics and number of lesions at different stages, where 0, 1, 2, and 3 grades indicate no DR, mild, moderate and severe DR respectively.

TABLE I
TYPES OF DR LESIONS IN IDRID DATABASE AND CRITERIA OF DR GRADING IN MESSIDOR DATABASE

IDRiD		Messidor		
Lesion	Nb Images	Grade	Description	Nb Images
MA^{a}	81	0	$MA = 0$ and $H = 0$	546
H^{b}	80	1	$0 < MA \leq 5$ and $H = 0$	153
EX^{c}	81	2	$5 < MA < 15$ and $0 < H < 5$	247
SE^{d}	40	3	$MA \geq 15$ and $H \geq 5$	254

[a]Microaneurysms, [b]Hemorrhages, [c]Hard Exudates, [d]Soft Exudates

B. The Comparison with the Baseline Methods

The ADAMIL algorithm mainly involves three components: domain adaptation, suspicious instance discriminator for filtering, and a multi-class multi-instance learning with attention mechanism. We investigate the three components of our ADAMIL at first. We evaluate the generalization performance of all methods using ten-fold cross-validation for all the comparable methods to ensure a fair comparison.

No MIL: an image-level classification model is constructed based on ResNet50;

MIL: a deep MIL model regards all patches as instances without filtering the irrelevant instances;

MIL+PT: a deep MIL model combined with the instance discriminator pre-trained only on the source domain without fine-tuning;

MIL+PL: the instance discriminator in 'MIL+PT' model is used to classify the instances in the target domain and a pseudo-label of each instance is obtained. With the pseudo-labels of target instances, the 'MIL+PT' model is further fine-tuned.

MIL+PT+FT: the instance discriminator in 'MIL+PT' model is fine-tuned with the generated instances by Cycle-GAN. With the fine-tuned model, the instances are filtered and fed into the MIL model without attention mechanism.

From Table II, we can see that the proposed method consistently achieves better classification performance than the competing methods in terms of Accuracy, Precision, micro-F1 and AUC, which demonstrates the effectiveness of our ADAMIL method. Moreover, Fig. 5 shows the ROC curve of the comparable methods. It is apparent that our proposed method achieves higher AUC value than the other contender methods. These results reveal several interesting points: (1) With the limited size of the training set, it cannot sufficiently train a bag-level classification model. (2) MIL without any instance filtering achieves a worse result, which indicates that the large amounts of irrelevant instances negatively affect the multi-instance learning. (3) The performance of MIL+PT

TABLE II
THE COMPARISON OF THE THREE COMPONENTS IN THE ADAMIL ALGORITHM (THE BEST RESULTS ARE HIGHLIGHTED)

Component	Accuracy	Precision	Recall	micro-F1	AUC
No MIL	0.253	0.112	0.241	0.193	0.502
MIL	0.489	0.498	0.525	0.511	0.659
MIL+PT	0.606	0.729	**0.630**	0.670	0.737
MIL+PL	0.507	0.594	0.548	0.561	0.672
MIL+PT+FT	0.547	0.574	**0.630**	0.597	0.698
ADAMIL	**0.764**	**0.765**	0.616	**0.676**	**0.749**

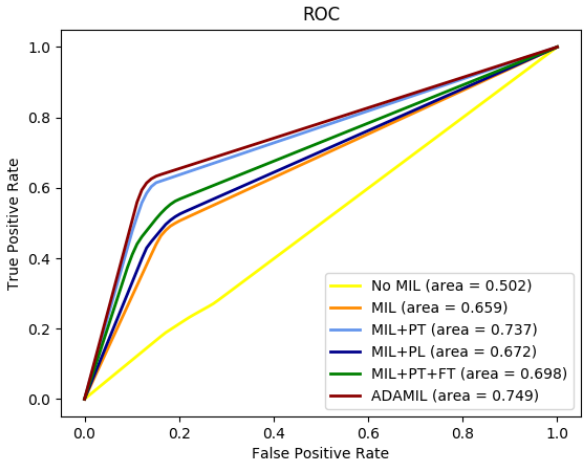


Fig. 5. ROC curve of the comparative methods.

is lower than the ADAMIL method, which indicates that the distribution from multiple data domains is different, thus employing the classifier simply borrowing labeled data from the source domain results in poor instance classification performance. Another important conclusion is that MIL+PL may generate wrong labels due to the inconsistent distribution from multi-domain data. (4) ADAMIL further improves the performance over MIL+PT+FT by incorporating the attention mechanism. It demonstrates that the attention mechanism not only provides the interpretability support, but also improves the MIL performance.

C. Ablation Study

a) Image Reconstruction Quality Comparison: We first visualize the reconstruction process of CycleGAN, as shown in Fig. 6, where the features between $Real\_S$ and $Cyclic\_S$ and between $Real\_T$ and $Cyclic\_T$ are very similar. This phenomenon verifies that CycleGAN has achieved the domain adaptation between IDRiD and Messidor datasets. Recently, many adversarial-based methods for cross-domain adaptation with unpaired data are developed, e.g. UNIT[13], DualGAN[14], MUNIT[15], DRIT[16], UGATIT[17], DiscoGAN[18]. Therefore, we further visually evaluate the transfer performance of CycleGAN and other methods from $D_S \rightarrow D_T$ and $D_T \rightarrow D_S$, as shown in Fig. 6. The evaluation criteria is based on the color, shape, and texture of the lesions and the background without lesions. We can observe that the reconstructed patches generated by CycleGAN model are best.

b) Influence of the Size of Image Patches: In the previous experiments, the patch size in the ADAMIL method is fixed as 128×128. We now study the influence of patch size on the performance of ADAMIL. In Fig. 7, we compare the ADAMIL-128 with ADAMIL-256 on the Messidor dataset. From this figure, it can be seen that ADAMIL with a patch size of 128×128 obtains a better performance. It implies that the small lesions within the larger patches are difficult to obtain a discriminate representation. In addition, large patches bring a huge computation at cost.

c) Influence of MIL Attention Mechanism: The key component in our ADAMIL is the multi-class multi-instance learning with attention, which aggregates instance probability distribution vectors or instance feature vectors into a bag representation. To demonstrate the advantage of the attention mechanism, we compare the attention mechanism with other pooling operators, such as max pooling, mean pooling, and log-sum-exp (LSE) pooling [19]. As shown in Table III, compared with other methods, the attention mechanism we adopt is more preferable.

TABLE III
THE INFLUENCE OF DIFFERENT POOLING ON ADAMIL ALGORITHM

Pooling Method	Accuracy	Precision	Recall	micro-F1	AUC
Attention(ours)	**0.764**	**0.765**	**0.616**	**0.676**	**0.749**
max	0.200	0.190	0.231	0.208	0.467
mean	0.540	0.651	0.554	0.586	0.656
LSE	0.259	0.276	0.141	0.174	0.506

D. The Comparison with the State-of-the-art Approaches for DR Diagnosis

We also compare our method with several recent state-of-the-art methods reported on the Messidor dataset in Table IV. In [20], a method combining fractal dimension with random forest classifier was developed. Additionally, red lesion detection generates a lesion probability map, which combines location, size to express features, and finally uses random forest (RF) to realize DR grading; two ophthalmologists A and B also grade the Messidor dataset respectively [21]. Moreover, texture analysis methods like statistical moments and GLCM are used to extract features, and feed them into classifiers e.g. Support Vector Machine (SVM), RF, AdaBoost, Gradient Boost, Gaussian Naive Bayes (GaussianNB) for DR grading in [22]. From the results, we can find that our algorithm achieved a very competitive performance when compared with the state-of-the-art methods for DR grading on the Messidor dataset. Moreover, it is worth noting that our algorithm outperforms two ophthalmologists by 4.7% and 12.2% with respect to the accuracy.

E. Interpretability Validation

The deep learning-based models typically lack interpretability, which is a missing evidence to support doctors. Therefore,

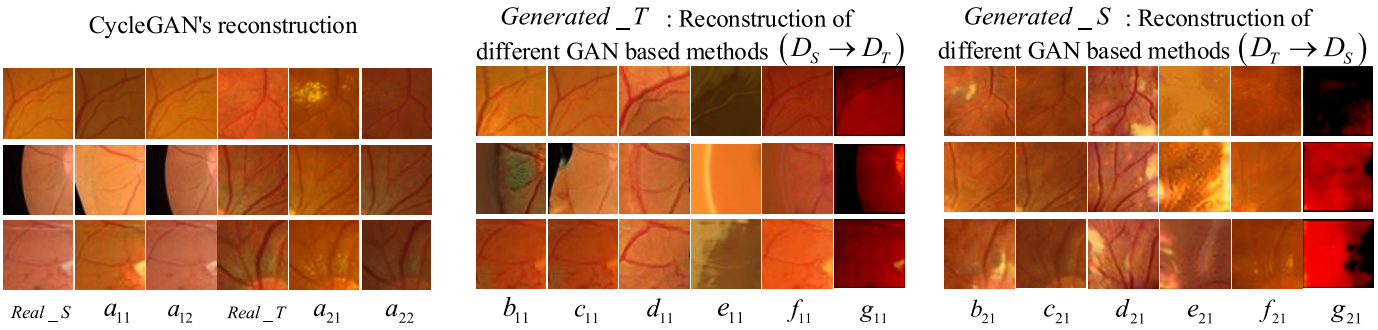


Fig. 6. Reconstruction visualization of CycleGAN model and comparison with different GAN based methods. Original images are defined as $Real\_S$, $Real\_T$. Generated images of $D_S \rightarrow D_T$ and $D_T \rightarrow D_S$ are defined as $Generated\_T$ and $Generated\_S$, i.e., a_{11}, a_{21} respectively. Reconstructed images of $D_S \rightarrow D_T$ and $D_T \rightarrow D_S$ are defined as $Cyclic\_S$ and $Cyclic\_T$, i.e., a_{12}, a_{22} respectively. b, c, d, e, f, g are images generated sequentially by different GAN based methods i.e., UNIT, DualGAN, MUNIT, DRIT, UGATIT, DiscoGAN.

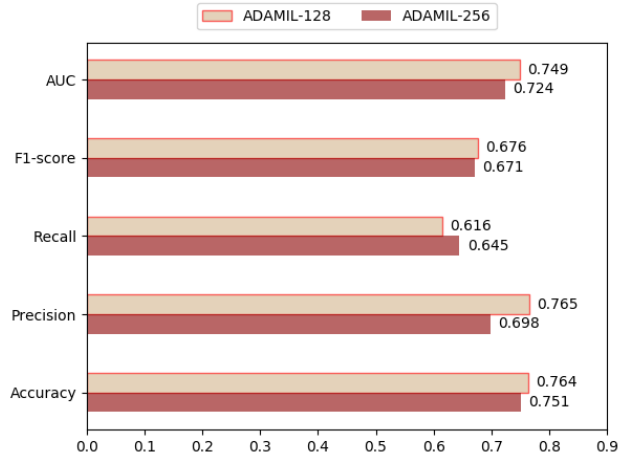
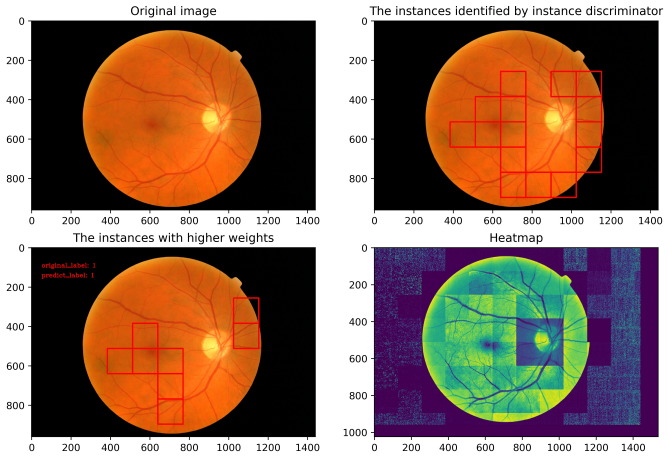


Fig. 7. Influence of the size of image patches on the performance of ADAMIL.

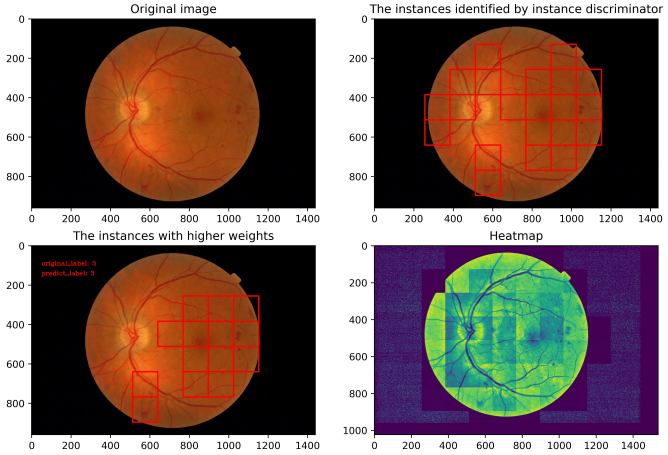
TABLE IV
THE COMPARISON BETWEEN OUR METHOD WITH THE STATE-OF-THE-ART METHODS FOR DR GRADING

Methods	Accuracy	Validation	Images
Fractal-based [20]	0.483	5-fold	1200
Expert A [21]	0.730	Manual	1200
Expert B [21]	0.681	Manual	1200
RF [21]	0.741	leave-one-out	1200
SVM [22]	0.47	5-fold	1200
RF [22]	0.459	5-fold	1200
AdaBoost [22]	0.36	5-fold	1200
Gradient Boost [22]	0.412	5-fold	1200
GaussianNB [22]	0.35	5-fold	1200
ADAMIL(ours)	**0.764**	10-fold	1200

we solve the black box problem of deep learning through HMA lesion discrimination and the attention mechanism, i.e., outputting the severity of DR while giving the location information of highly suspected lesions to support the decision-making. Fig. 8 shows the workflow of the ADAMIL algorithm. A heatmap is obtained by multiplying the pixel values with their attention weights of the corresponding patches. The lesion regions can be identified through the heatmap.



(a) Grade = 1



(b) Grade = 3

Fig. 8. Some examples of instances filtering and heatmap obtained by MIL with attention

CONCLUSION

Traditional automatic classification algorithms are mostly based on the identification of HMA lesions. However, due to lack of lesion annotation data, it is difficult to employ supervised learning methods for HMA detection. To overcome it, we propose a DR grading framework based on a domain adaptation multi-instance learning with attention(ADAMIL), which only requires image-level annotation to achieve both classification of DR and the location of highly suspected lesions. We formulated the problem of DR grading as a multi-class multi-instance learning problem. Under the support of IDRiD dataset with the lesion annotation, we develop an instance discriminator for filtering negative lesions with domain adaptation. Afterwards, an attention mechanism is incorporated into the MIL framework to obtain the important weights of the patches, thus providing medical diagnosis interpretability. The experimental results on the public Messidor dataset indicate that our method achieves a better performance compared to the state-of-the-art approaches. Moreover, the proposed HMA with CycleGAN and attention mechanism can provide interpretable results, which is very important for the potential application of automatic computer-aided diagnosis in the practical clinical workflow. Our experimental results indicate that the proposed ADAMIL is far superior to the current DR grading method, and its performance can compete with human experts. We use embedded technology to achieve local-to-global representation, which can greatly reduce annotation work while maintaining predictive performance at an acceptable level. Furthermore, our method can be extended to other medical fields where data is weakly supervised. In future work, we will evaluate our model on the Kaggle's Diabetic Retinopathy Detection Challenge (EyePACS) and other medical fields, such as histopathology images of cancers.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (No.62076059) and the Fundamental Research Funds for the Central Universities (No. N2016001)

REFERENCES

[1] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. Motala, K. Ogurtsova, J. Shaw, D. Bright, and R. Williams, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas, 9th edition," *Diabetes Research and Clinical Practice*, vol. 157, pp. 107843, Sep 2019.

[2] L. Crossland, D. Askew, R. Ware, P. Cranstoun, P. Mitchell, A. Bryett, and C. Jackson, "Diabetic retinopathy screening and monitoring of early stage disease in australian general practice: Tackling preventable blindness within a chronic care model," *Journal of Diabetes Research*, vol. 2016, no. 4, pp. 1–7, Dec 2016.

[3] W. Cao, N. Czarnek, J. Shan, and L. Li, "Microaneurysm detection in fundus images using small image patches and machine learning methods," *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 325–331, Nov 2017.

[4] V. Cheplygina, M. de Bruijne, and J. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, Apr 2019.

[5] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, no. 1, pp. 329–353, Dec 2018.

[6] A. A. M. Al-Saffar, T. Hai, and M. A. Talab, "Review of deep convolution neural network in image classification," *International Conference on Radar*, pp. 26–31, Jan 2018.

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

[8] J.-Y. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *ICCV*, pp. 2242–2251, Oct 2017.

[9] G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger, "Densely connected convolutional networks," *CVPR*, pp. 2261–2269, Jul 2017.

[10] Y. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *ICML*, pp. 933–941, Dec 2016.

[11] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, pp. 25, Jul 2018.

[12] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, "Feedback on a publicly distributed image database: The messidor database," *Image Analysis & Stereology*, vol. 33, no. 3, pp. 231–234, Jul 2014.

[13] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *NIPS*, pp. 700–708, Mar 2017.

[14] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," *ICCV*, pp. 2868–2876, Oct 2017.

[15] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," *ECCV*, vol. 11207, pp. 179–196, Apr 2018.

[16] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," *ECCV*, vol. 11205, pp. 36–52, Aug 2018.

[17] J. Kim, M. Kim, H.-W. Kang, and K. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," *ICLR*, Jul 2019.

[18] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," *Proc. Int. Conf. Machine Learn. (ICML)*, vol. 34, pp. 1857–1865, Mar 2017.

[19] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, Oct 2016.

[20] F. Alzami, Abdussalam, R. Megantara, A. Fanani, and P. Purwanto, "Diabetic retinopathy grade classification based on fractal analysis and random forest," *iSemantic*, pp. 272–276, Sep 2019.

[21] L. Seoud, J. Chelbi, and F. Cheriet, "Automatic grading of diabetic retinopathy on a public database," *Proceedings of the Ophthalmic Medical Image Analysis International Workshop*, pp. 97–104, Oct 2015.

[22] J. Labhade, L. Chouthmol, and S. Deshmukh, "Diabetic retinopathy detection using soft computing techniques," *International Conference on Automatic Control and Dynamic Optimization Techniques*, pp. 175–178, Sep 2016.