Computers in Biology and Medicine 144 (2022) 105341

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Collaborative learning of weakly-supervised domain adaptation for diabetic retinopathy grading on retinal images

Peng Cao [a,b,*], Qingshan Hou [a], Ruoxian Song [a], Haonan Wang [a], Osmar Zaiane [c]

[a] Computer Science and Engineering, Northeastern University, Shenyang, China
[b] Key Laboratory of Intelligent Computing in Medical Image of Ministry of Education, Northeastern University, Shenyang, China
[c] Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Alberta, Canada

ARTICLE INFO

Keywords:
Diabetic retinopathy
Computer-aided diagnosis
Multi-instance learning
Domain adaption
Interpretability

ABSTRACT

Early detection and treatment of diabetic retinopathy (DR) can significantly reduce the risk of vision loss in patients. In essence, we are faced with two challenges: (*i*) how to simultaneously achieve domain adaptation from the different domains and *(ii)* how to build an interpretable multi-instance learning (MIL) on the target domain in an end-to-end framework. In this paper, we address these issues and propose a unified weakly-supervised domain adaptation framework, which consists of three components: domain adaptation, instance progressive discriminator and multi-instance learning with attention. The method models the relationship between the patches and images in the target domain with a multi-instance learning scheme and an attention mechanism. Meanwhile, it incorporates all available information from both source and target domains for a jointly learning strategy. We validate the performance of the proposed framework for DR grading on the Messidor dataset and the large-scale Eyepacs dataset. The experimental results demonstrate that it achieves an average accuracy of 0.949 (95% CI 0.931–0.958)/0.764 (95% CI 0.755–0.772) and an average AUC value of 0.958 (95% CI 0.945–0.962)/0.749 (95% CI 0.732–0.761) for binary-class/multi-class classification tasks on the Messidor dataset. Moreover, the proposed method achieves an accuracy of 0.887 and a quadratic weighted kappa score value of 0.860 on the Eyepacs dataset, outperforming the state-of-the-art approaches. Comprehensive experiments confirm the effectiveness of the approach in terms of both grading performance and interpretability. The source code is available at https://github.com/HouQingshan/WAD-Net.

1. Introduction

Diabetic Retinopathy (DR) is a consequence of retinal microvascular changes triggered by diabetes. It is the most common leading cause of blindness and visual disability in the working-age population worldwide [1]. The diagnosis and grading performance of DR highly depends on the detection of structures such as microaneurysms (MAs) and hemorrhages, which are considered as early signs of DR, as shown in Fig. 1. Thus, the grading of DR severity level is a laborious process that is time-consuming and can sometimes be prone to misdiagnosis. Therefore, an automatic disease diagnosis on retinal fundus images is urgently required for early DR detection and severity level grading for assisting experts. Integrating machine learning algorithms can be established to predict the multi-class labels for the DR severity level [2,3]. Recently, deep learning techniques (e.g., convolutional neural networks) have emerged and made remarkable achievements in DR grading as a fundamental element of automatic disease diagnosis techniques [4,5]. The success of deep learning is mainly attributed to its capability of extracting highly representative features. The procedure is shown in Fig. 2(a). The accurate detection of MAs and hemorrhages is a crucial step for early detection of DR as these are typically the earliest clinically recognizable signs. However, the lack of the pixel-wise lesion annotations hinders the traditional deep learning algorithms from detecting and identifying the suspicious regions. Hence, we aim to develop a weakly-supervised method [6,7], which can leverage the large amount of the image-level annotations to significantly reduce human annotation efforts, which is an important problem in the medical applications.

Multi-instance learning (MIL) is an extension of weakly-supervised learning by treating the whole retinal fundus image as a bag, and each patch as an instance in the bag. In our study, we regard the problem of DR grading as a multi-class multi-instance learning formulation (in Fig. 2(b)). More specifically, the images are divided into a regular grid of

[*] Corresponding author. Computer Science and Engineering, Northeastern University, Shenyang, China.
E-mail address: caopeng@cse.neu.edu.cn (P. Cao).

https://doi.org/10.1016/j.compbiomed.2022.105341
Received 5 October 2021; Received in revised form 20 February 2022; Accepted 20 February 2022
Available online 8 March 2022

small patches, and each patch is regarded as an instance. Through MIL, we can shift the problem of identifying the suspicious MA or hemorrhages regions in supervised learning setting to a weakly supervised learning for the whole fundus images relying only on the global labels, which simplifies data collection tremendously and is more in accordance with clinicians' reasoning. However, MIL still falls short for modeling the relationship between the patches and the global images due to the following challenges:

Challenge 1: how to solve the domain diversities between different domains. To further reduce dependencies on the comprehensive lesion annotations in the target domain, some previous methods leverage the auxiliary datasets with the lesion labels [8,9] to assist the weakly supervised model in the target domain without lesion labels in Fig. 2(c). However, they did not consider the domain shift between the auxiliary domain and target domain since the patient populations and the scanner protocol varied. Simply applying a classifier trained on an auxiliary domain to predict the lesion candidates inevitably performs poorly due to the domain gap [10].

Challenge 2: how to enable the weakly supervised deep learning models to be interpretable in the target domain. The inability to interpret the model prediction is a well-known limitation of most existing computer-aided DR diagnosis methods. Interpretability is essential as it can help in decision-making during diagnosis and treatment planning. From the clinical perspective, identifying task-specific biomarkers provides important insight into the disease to improve the treatment quality of patients. However, the existing deep learning models cannot provide intuitive illustrations for physicians and patients of how the diagnosis is made.

In essence, the question then becomes how to simultaneously achieve domain adaptation by transferring the knowledge from the source domain and build an interpretable multi-instance learning for the target domain with an end-to-end scheme. To address these two problems, we propose an interpretable end-to-end **W**eakly supervised learning network with **A**ttention mechanism and **D**omain adaptation, named WAD-Net for simultaneously diagnosing diabetic retinopathy and highlighting suspicious regions (Fig. 2(d)). More specifically, 1) there exist a large number of irrelevant instances in each bag that hinders the multi-instance learning. To filter out those instances that have a negative influence on the MIL performance, it is desirable to transfer the knowledge from a pre-trained instance classification model in the source

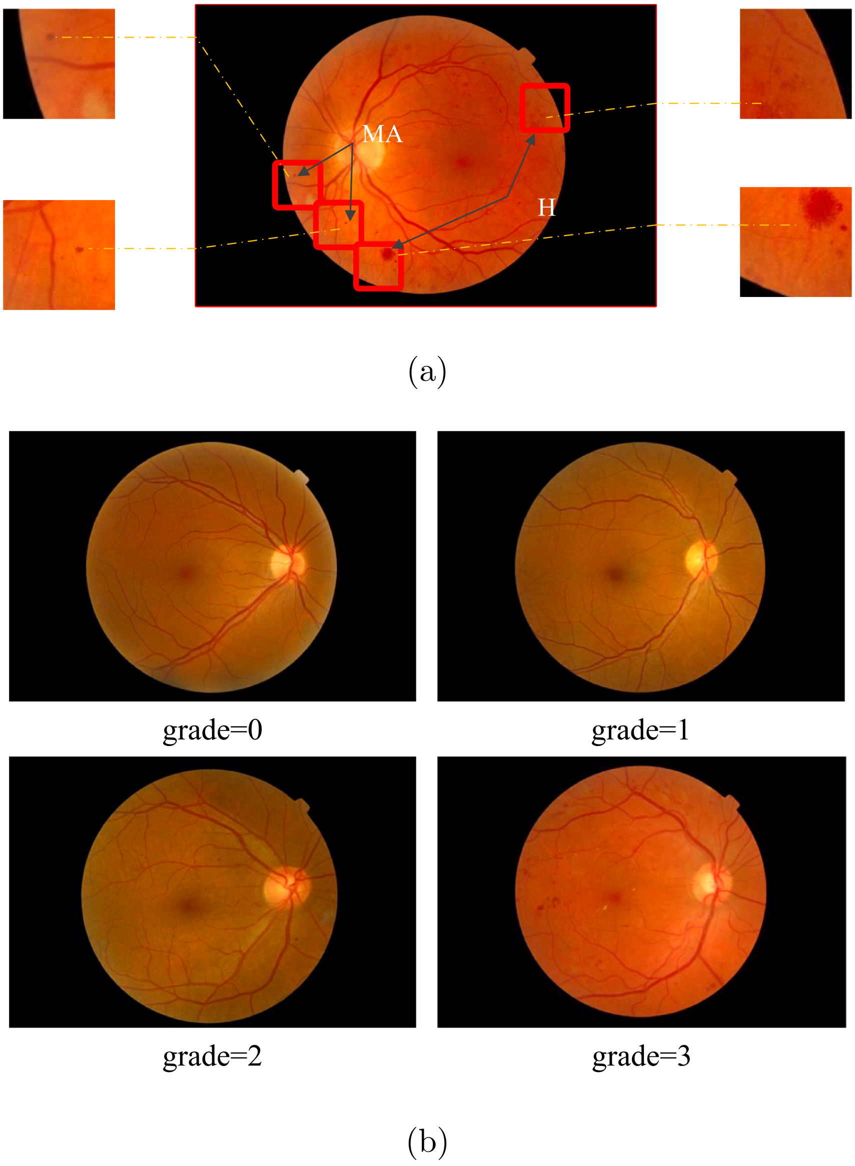


Fig. 1. The red lesions and severity of DR. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

domain to the target domain. Traditionally, the source and target domains generally share the same task but follow different distributions in the traditional domain adaptation. However, it is more challenging since both tasks are different in our work: it is an instance level DR lesion classification task with lesion labels in the source domain while it is an image level classification with only image label in the target domain. To address the domain shifts with respect to the distributions and tasks, the aim of our study is to transfer the knowledge from a source domain with lesion annotations to a target domain by minimizing the difference between domain distributions. On the one hand, we employ cycleGAN [11] to achieve an image-to-image translation, and develop a two-step progressive training scheme for enabling the instance discriminator to be adapted to the target domain. On the other hand, due to the unsupervised learning in cycleGAN, it does not guarantee that the generated samples are be beneficial to the classification performance in the target domain. Hence, we develop a collaborative learning framework to combine cycleGAN, instance progressive discriminator and multi-instance learning into a unified framework. As such, the supervision guided cycleGAN leads to better patch generation with image-level supervision scheme. 2) Meanwhile, we incorporate an attention mechanism into the proposed MIL framework. Through the attention mechanism, attention maps are generated to indicate which pixels play more important roles in making the image-level decision. Experiments on the real-world Messidor [12] and Eyepacs [13] datasets demonstrate that our WAD-Net method not only outperforms the state-of-the-art approaches but also is effective in automatically identifying disease-related lesions in making the image-level decision.

Our contributions can be summarized as follows.

1. In the medical field, it is difficult to obtain pixel-level annotations. We attempt to address the issue of missing lesion labels from a new perspective. To make full use of the existing labeled data, we propose a unified weakly-supervised domain adaptation framework to model the relationship between the instances and bags in the target domain with a multi-instance learning scheme, and to incorporate all available information from both source and target domains with a jointly learning mechanism. Therefore, we do not require a large number of pixel-wise annotated samples anymore.

2. Different from the traditional cycleGAN which is an unsupervised model, our generative model aims to produce more effective samples for the multi-instance learning in the target domain. It better generates patches by a domain adaptation with an end-to-end weak supervision scheme, which is more effective in both the instance generation and classification in the target domain.

3. We propose a MIL based DR framework with an attention mechanism. This mechanism benefits the performance from two aspects: 1) the obtained attention map can highlight the suspicious regions for

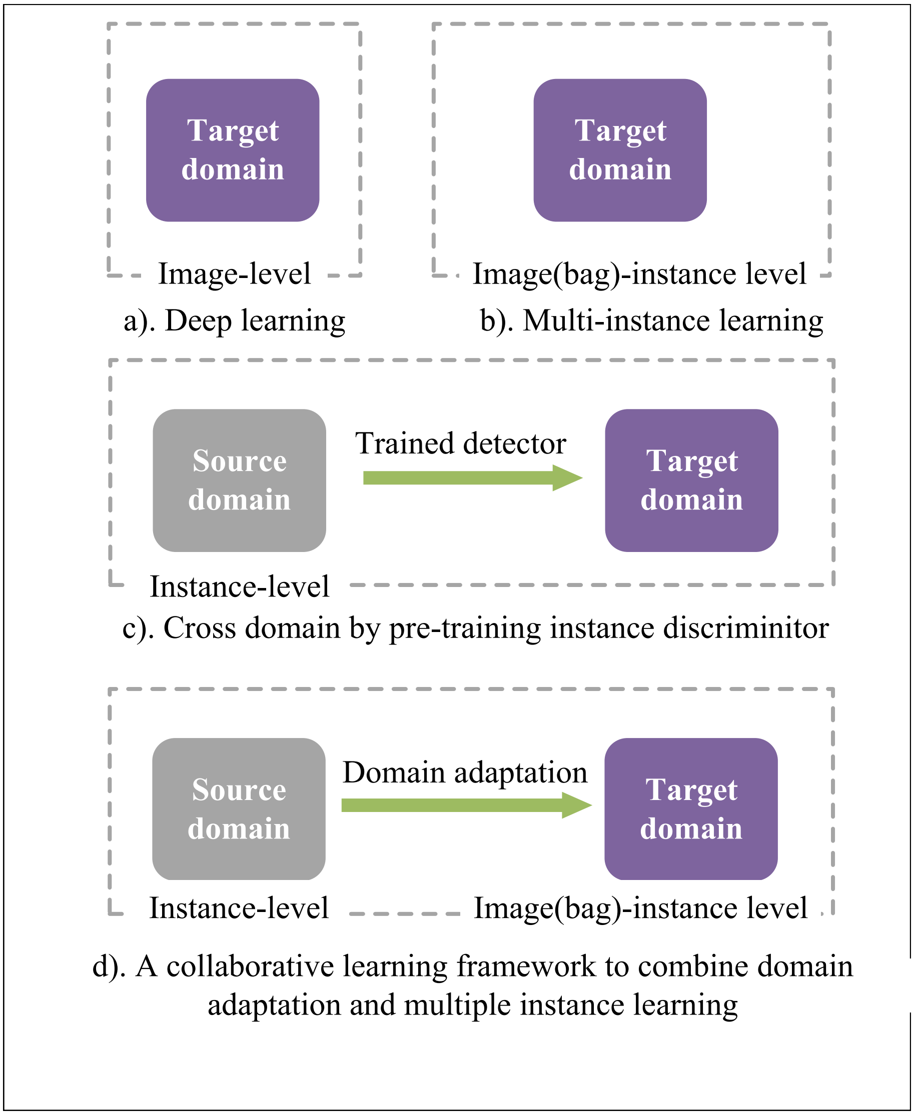

Target domain
Image-level
a). Deep learning
Target domain
Image(bag)-instance level
b). Multi-instance learning
Source domain
Trained detector
Target domain
Instance-level
c). Cross domain by pre-training instance discriminitor
Source domain
Domain adaptation
Target domain
Instance-level
Image(bag)-instance level
d). A collaborative learning framework to combine domain adaptation and multiple instance learning

Fig. 2. The comparison among the traditional DR classification methods and the proposed weakly-supervised domain adaptation method.

providing the decision-making during diagnosis; 2) the attention weights can enhance the contribution of the relevant instances for improved classification performance.

4. Our method outperforms the state-of-the-art methods on the two independent DR datasets (Messidor and Eyepaces) with a binary class (diagnosis) or a multi-class (grading) classification task, respectively. The quantitative and qualitative results on the two different datasets confirm that the proposed unified framework could boost the performance in the target domain by jointly training with the pixel-wise annotated lesions from the source domain and the images with grading labels from the target domain. Moreover, our WAD-Net can identify the localization of the suspicious lesion regions.

2. Related work

2.1. DR diagnosis with deep learning

Recent years have witnessed the growing interest in the automated DR severity grading. Existing DR grading methods can be divided into two categories. The first category is to determine DR grading by identifying the location information of the DR related lesions, e.g., microaneurysms, hemorrhage. The accurate detection of microaneurysms (MA) and hemorrhage is a crucial step for early detection of DR as these are typically the earliest clinically recognizable signs. Recent works show that deep learning can produce promising results in lesion segmentation or detection for early DR diagnosis [14]. Chudzik et al. employed convolutional neural network architecture to detect microaneurysms from fundus images [15]. Adem developed a CNN-based exudate detection system with Circular Hough Transformation to automatically detect the exudates in the retinal image [16]. Yan et al. detected DR red lesions by integrating the handcrafted features and the learned features by pretrained LeNet using a Random Forest classifier [17]. Van et al. proposed a method to improve and speed up the CNN training for hemorrhage detection by dynamically selecting misclassified negative samples [18]. Yang et al. proposed an automatic DR analysis algorithm with a two-stage deep learning algorithm [19]. It can identify the location as well as the type of lesions, and produce the severity level of DR by integrating both local and global networks to learn more complete and specific features for DR analysis. Lin proposed a new attention-based network for unifying lesion detection and DR identification [20]. Tahira et al. proposed a Fast Region-based Convolutional Neural Network (FRCNN) algorithm with fuzzy k-means (FKM) clustering for automated localization and recognition of diabetes-based eye diseases [21]. Although these deep learning based methods have dramatically improved the performance for the DR lesions, it requires large annotated sets of these lesions. It is expensive to annotate the lesions on the medical images in a pixel-wise manner.

Another category is to train a deep learning model, such as ResNet, Inception and DenseNet, for distinguishing the disease severity with only image-level grading label supervision [22]. These methods aim to adopt image-level labels to train DR grading models, saving ophthalmologists' labors for costly pixel-level annotations. For example, Zhou et al. developed a jointly learning framework for simultaneously DR grading and lesion segmentation with an attention mechanism [23]. By exploring the cross-disease relationship, Zhu et al. presented a cross-disease attention network for jointly grading DR and DME [24]. Jiang et al. proposed a deep learning-based multi-label classification model with Gradient-weighted Class Activation Mapping (Grad-CAM) for DR classification and lesions lactation [25]. Wang proposed a hierarchical multi-task deep learning framework for the diagnosis of DR severity and DR related features at the same time [26]. Quellec et al. applied L2 regularization over the best performed DCNN in the KAGGLE competition for DR detection named o-O [27]. Wang et al. proposed a multi-channel based generative adversarial network (MGAN) with semisupervision to make full use of both labeled data and unlabeled data [28]. Li et al. proposed a cross-disease attention network (CANet) to jointly grade DR and DME by exploring the internal relationship between the two relevant diseases with only image-level supervision [24].

It is expensive to annotate the lesion labels on the medical images in a pixel-wise manner, hence we choose the second category to construct a disease grading model in our study. However, these deep learning methods with image-level supervision suffer from two main limitations. First, deep learning is considered a black box, hence the identification of the suspicious regions is not a straightforward process. It is the major problem that hinders the deep learning methods on the clinical application. Second, they did not appropriately make full use of all the valuable information from the auxiliary domains because of the domain gap. Hence, we aim to develop a weakly-supervised domain adaptation paradigm for the DR grading, which leverages the limited number of pixel-level annotated images available along with a large number of image-level annotations to enhance the performance of the DR severity prediction.

2.2. Multi-instance learning

Multi-instance learning (MIL) has been successfully applied to various problems including object detection and computer-aided diagnosis. In a MIL problem, only the labels of the bags are known whereas the individual labels of the instances contained in the bags are not provided. This is different from the supervised classification approach, where the label of each instance is known. The learning process is weakly supervised due to the ambiguous instance labels. In the DR grading study, only a few works formulate weakly supervised DR grading as a MIL problem where each image is represented by a bag (labeled as healthy or abnormal), and the unlabeled lesion candidates in the images are considered as instances. Cao et al. proposed a multi-kernel multi-instance learning method to solve the multi-class DR grading problem [10]. Zhou et al. proposed a deep MIL method by jointly feature learning and classifier training for an improvement on detecting DR images [29].

3. Methods

3.1. Formulation

The task in our study is described as follows:

(i) Source domain $D_S = \left\{ \left(\boldsymbol{x^s}_1, y_1^s \right), \ldots, \left(\boldsymbol{x^s}_{N^s}, y_{N^s}^s \right) \right\}$: the instance-level (lesion-wise) labels are available.

(ii) Target domain $D_T = \left\{ \left(\boldsymbol{X^t}_1, Y_1^t \right), \ldots, \left(\boldsymbol{X^t}_{N^t}, Y_{N^t}^t \right) \right\}$: only image-level labels are observed. Each $\boldsymbol{X^t}$ contains a series of patches. We denote the instance set in the target domain as $\boldsymbol{x}^t = \{ \boldsymbol{x^t}_1, \ldots, \boldsymbol{x^t}_{N^t \times B} \}$, where B represents the patch size in each target image.

(iii) There exists a domain gap between the two domains.

In this work, we propose an end-to-end Weakly-supervised network with Attention mechanism and Domain adaptation (WAD-Net). As illustrated in Fig. 3, the architecture of WAD-Net consists of three components: domain adaptation, instance progressive discriminator and multi-instance learning with attention. Given the domains of D_S and D_T, the aim of WAD-net is to jointly optimize a generative model $G_{\theta_G}(\cdot)$ for domain adaptation, an instance progressive discriminator model $C_{\theta_I}(\cdot)$ from the instance level perspective, and a multi-instance learning model $C_{\theta_M}(\cdot)$ from the image level perspective.

To train the domain adaptation model $G_{\theta_G}(.)$, a mapping function is need to be optimized to achieve a transformation from source domain D_S to target domain D_T, as follows:

$$\min_{\theta_G, \theta_{G'}} \max_{\theta_Q} \sum\nolimits_{n=1}^{N_S} L_G \left(G_{\theta_G}^{S \to T}(x_n^S), G_{\theta_{G'}}^{T \to S}(\widehat{x}_n^S), Q_{\theta_Q}^T(x_n^S, \boldsymbol{x}^t) \right) \tag{1}$$

Where $G_{\theta_G}^{S \to T}(\cdot)$ and $G_{\theta_G'}^{T \to S}(\cdot)$ denote two generators for domain adaptation, $\widehat{x}_n^S$ is the transformed instance by $G_{\theta_G}^{S \to T}(\cdot)$, $Q_{\theta_Q}^T$ is a discriminator to distinguish whether instances are translated from another domain. The

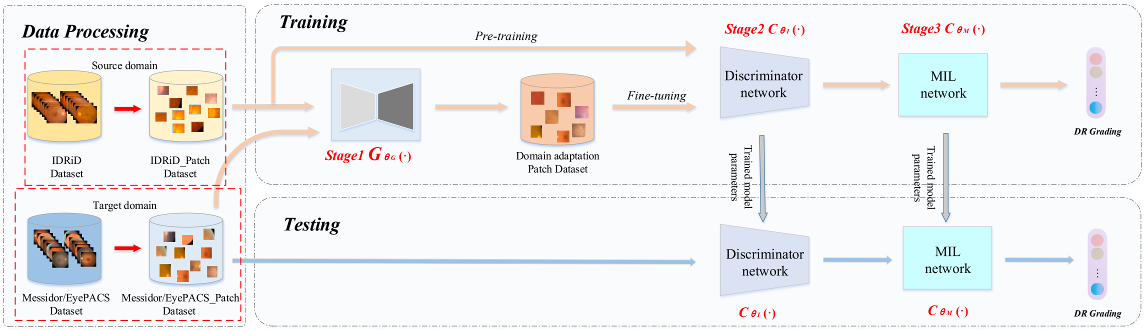


Fig. 3. The overview of our proposed WAD-Net for DR grading diagnosis.

optimization function for the instance progressive discriminator model $C_{\theta_I}(\cdot)$ is defined as:

$$Pre-training:\ \min_{\theta_I}\frac{1}{N_S}\sum\nolimits_{n=1}^{N_S}L_{C_I}\left(C_{\theta_I}\left(x_n^S\right),y_n^S\right)\tag{2}$$

$$Fine-tuning:\ \min_{\theta_I}\frac{1}{N_S}\sum\nolimits_{n=1}^{N_S}L_{C_I}\left(C_{\theta_I}\left(\widehat{x}_n^S\right),y_n^S\right)\tag{3}$$

Where $C_{\theta_I}(\cdot)$ denotes an instance progressive discriminator model, N_S is the total number of instance level data and y^S is the patch-level label ($y^S=1$ indicates lesion and $y^S=-1$ indicates background). The multi-instance learning model $C_{\theta_M}(\cdot)$ can be formulated as:

$$\min_{\theta_M,\alpha}\frac{1}{N_T}\sum\nolimits_{n=1}^{N_T}L_{C_M}\left(C_{\theta_M}\left(X_n^T\right),Y_n^T,\boldsymbol{\alpha},\boldsymbol{S}_n\right)\tag{4}$$

Where $C_{\theta_M}(\cdot)$ represents a multi-instance learning model, $\boldsymbol{\alpha}$ is the attention weights for patches, $\boldsymbol{S}_n$ is a score map for suspected lesion patches obtained by $C_{\theta_I}(\cdot)$, N_T is the total number of image level data and Y^T is the DR severity classification label for the image-level annotated data. Therefore, in order to jointly optimize the three components, the most important consideration is how to design and optimize $G_{\theta_G}(\cdot)$, $C_{\theta_I}(\cdot)$ and $C_{\theta_M}(\cdot)$.

In the target domain, no pixel-wise labeled data is available for training. Our idea is to leverage information from auxiliary pixel-wise labeled images in a source domain. Hence, two domain data are leveraged during the training stage. During testing, given an unseen image from the target domain, the outputs are a predicted DR severity level and a corresponding lesion attention map.

3.2. Overview

In this paper, we propose WAD-Net that jointly does multi-instance learning and domain adaptation with CycleGAN for the weakly supervised DR classification. As shown in Fig. 4, our proposed method contains four steps:

(1) Instance discriminator pre-training(source domain): First, we pre-train an instance discriminator with the instance level annotations in the source domain.

(2) Domain adaptation with CycleGAN(cross two domains): the aim is to learn a mapping function for the cross-domain with unpaired examples. In our study, it acts as an image-image translation by constructing a mapping function between two domains. With CycleGAN, we train a generative model that translates the pixel-wise annotated lesion information from the source domain D_S to the target domain D_T. Then, we can obtain the domain-transferred instances accompanied by the instance-level annotations.

(3) Fine-tuning the instance discriminator with the generated instances: with the pre-trained discriminator (the first step) and the domain-transferred instances (the second step), we fine-tune the discriminator with the generated instances to achieve domain

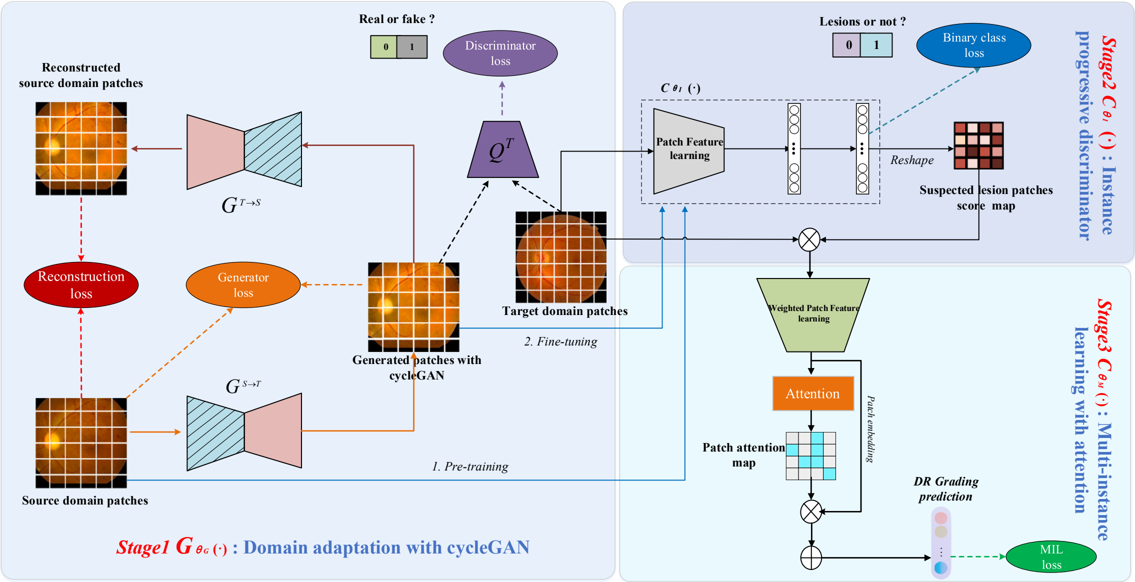

Real or fake ?
0
1
Discriminator loss
Reconstructed source domain patches
$G^{T\to S}$
Q^T
Reconstruction loss
Generator loss
Target domain patches
Generated patches with cycleGAN
2. Fine-tuning
$G^{S\to T}$
1. Pre-training
Source domain patches
Stage1 $G_{\theta_G}(\cdot)$: Domain adaptation with cycleGAN
Lesions or not ?
0
1
Binary class loss
$C_{\theta_I}(\cdot)$
Patch Feature learning
Reshape
Suspected lesion patches score map
Stage2 $C_{\theta_I}(\cdot)$: Instance progressive discriminator
Weighted Patch Feature learning
Attention
Patch embedding
Patch attention map
DR Grading prediction
MIL loss
Stage3 $C_{\theta_M}(\cdot)$: Multi-instance learning with attention

Fig. 4. The detail pipeline of our proposed WAD-Net for the DR grading prediction.

adaptation. The aim is to learn an instance discriminator that can perform well in the target domain with a domain shift. Actually, we develop a two-step progressive training strategy for the instance discriminator. Finally, the trained instance discriminator produces a score for each patch in the target. The score indicates the suspicious lesion probability of each patch.

(4) MIL with attention mechanism(target domain). With the initial score map of each image produced by the instance discriminator (the third step), the proposed multi-class multi-instance learning framework is able to predict the DR level and locate the highly suspicious lesions at the same time.

Noting that all modules are simultaneously trained in an end-to-end manner to achieve the highest performance.

3.3. Domain adaptation by cycleGAN

A common approach to solve this problem is to directly apply a pre-trained model of a source domain to the target domain. However, there exists a large domain gap between the source and target domains due to different illumination or camera versions. The aim of the task is to learn a model $C_{\theta_M}(\cdot)$ that correctly predicts the DR grading output and identifies the suspicious region as accurately as possible in the target domain under the conditions that sufficient instance-level annotations in the source domain and image-level annotations in the target domains are available, respectively.

To mitigate the effects of domain shift, it is desirable to learn a mapping function between the D_S domain and the D_T domain. We aim to learn the mapping function by CycleGAN, of which the advantage is that it allows each input image of the source domain to be converted into a target domain by image-to-image translation, and then be reconstructed to the source domain. The CycleGAN consists of a generator $G^{S\to T}$ is trained to produce convincing target samples that fool an adversarial discriminator Q^T, and a discriminator Q^T which attempts to discriminate the real target data from the generated target data by $G^{S\to T}$.

These correspond to the loss function as follows:

$$\mathrm{L}_{adv} = \left(\mathbb{E}_{\boldsymbol{x}^t\sim P_{data(\boldsymbol{x}^t)}}\left[\log\left(Q^T(\boldsymbol{x}^t)\right)\right]\right) + \mathbb{E}_{\boldsymbol{x}^s\sim P_{data(\boldsymbol{x}^s)}}\left[\log\left(1 - Q^T\left(G^{S\to T}(\boldsymbol{x}^s)\right)\right)\right] \tag{5}$$

The overall loss function in Eq. (5) involves two parts:

$$\mathrm{L}_G = \mathbb{E}_{\boldsymbol{x}^s\sim P(\boldsymbol{x}^s)}\left[\log\left(1 - Q^T\left(G^{S\to T}(\boldsymbol{x}^s)\right)\right)\right] \tag{6}$$

$$\begin{aligned}\mathrm{L}_Q \quad &= \mathbb{E}_{\boldsymbol{x}^t\sim P(\boldsymbol{x}^t)}\left[\log\left(Q^T(\boldsymbol{x}^t)\right)\right] \\ &+\mathbb{E}_{\boldsymbol{x}^s\sim P(\boldsymbol{x}^s)}\left[\log\left(1 - Q^T\left(G^{S\to T}(\boldsymbol{x}^s)\right)\right)\right]\end{aligned} \tag{7}$$

where the data distribution is denoted as $\boldsymbol{x}^s \sim P(\boldsymbol{x}^s)$ and $\boldsymbol{x}^t \sim P(\boldsymbol{x}^t)$.

We also apply a cycle consistency loss to ensure the content is well-preserved during the image translation process. The cross-cycle consistency loss is defined as:

$$\mathrm{L}_{cyc} = \mathbb{E}_{\boldsymbol{x}^s\sim P_{data}(\boldsymbol{x}^s)}\left[\|\|\|G^{T\to S}\left(G^{S\to T}(\boldsymbol{x}^s)\right) - \boldsymbol{x}^s\|\|\|_1\right] \tag{8}$$

where $G^{S\to T}$ and $G^{T\to S}$ are two mapping functions. Our generator model $G^{S\to T}$ architecture has an overall architecture consisting of an encoder and a decoder symmetrically on the two sides. The encoding phase is

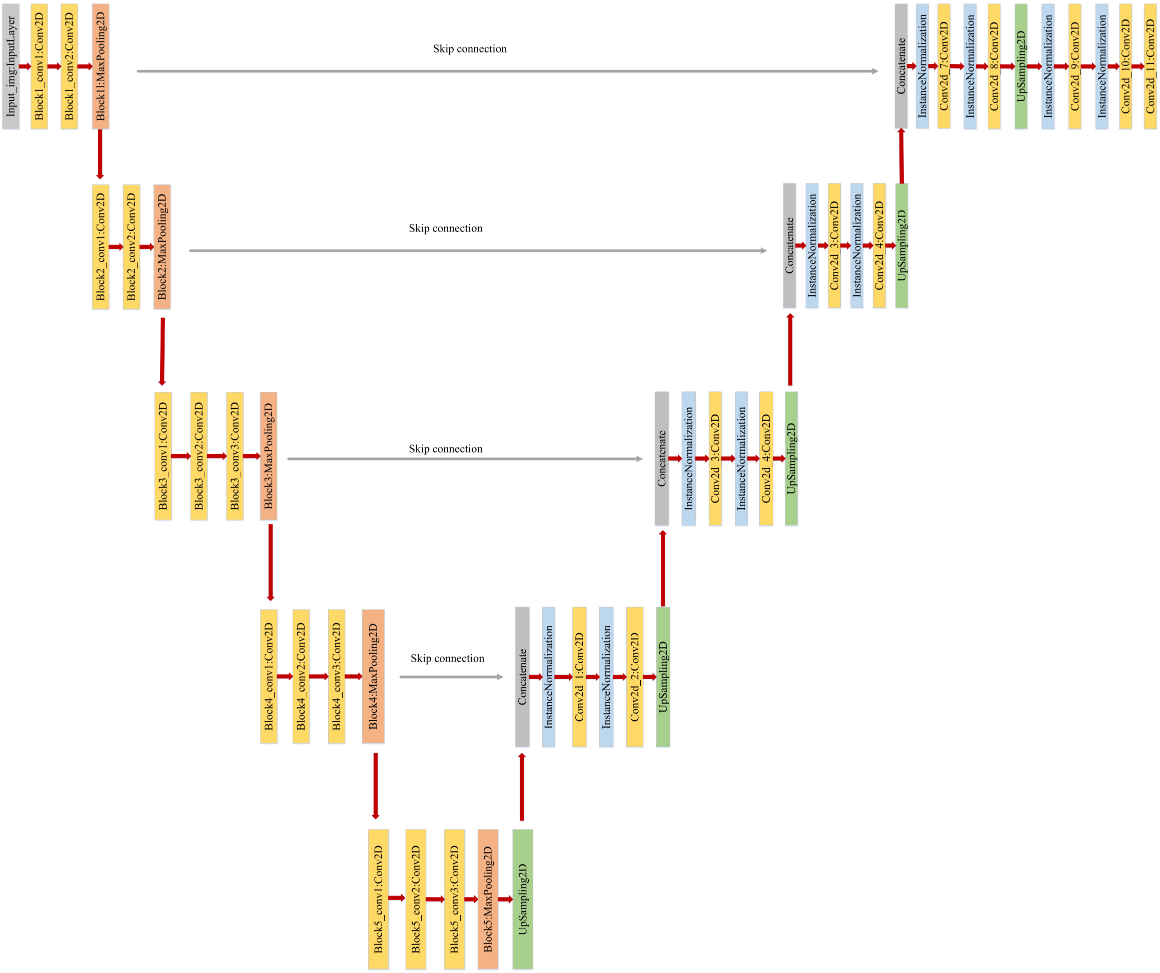


Fig. 5. Network architectures of generator.

used to encode input images in a lower dimensionality with richer filters, while the decoding phase is designed to do the inverse process of encoding by upsampling and merging low dimensional feature maps. The encoder consists of 13 convolutional layers combined with 5 downsampling layers (Fig. 5). The decoder module consists of a set of layers that upsamples the feature map of the encoder to recover spatial information. Besides, skip connection can help propagate the spatial information that gets lost during the pooling operation to help recover the full spatial resolution. Our discriminator involves 5 convolutional layers combined with 3 batch normalization layers (Fig. 6).

3.4. *The two-step progressive Instance Discriminator Training*

The training scheme for our instance discriminator consists of two stages. In the first step, we pre-train the instance discriminator model using the pixel-level annotated data in the source domain, and then fine-tune it with the generated target data. Both are trained in a fully-supervised manner. The aim is to learn a generalized classifier in the presence of a shift between source and target domain distributions.

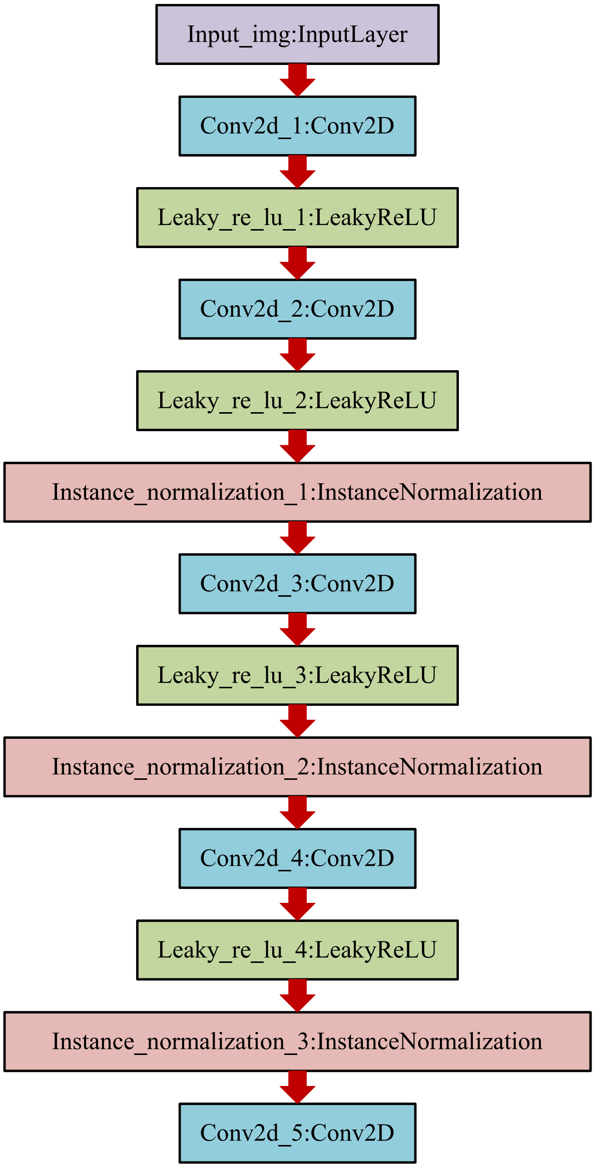


Fig. 6. Network architectures of discriminator.

Source domain: supervised Instance Discriminator Training.

Before training the entire retinal image, it is desirable to remove the irrelevant instances for achieving a better MIL learning performance. To perform lesion identification, our model aims to learn image features with discriminative property to distinguish between the lesions and background. Thus, we define the classification loss with a cross-entropy loss in Eq. (9).

$$\mathrm{L}_{C_I} = -\frac{1}{N_S}\sum\nolimits_{i=0}^{N_S-1}\left(y_i^s * \log\left(\widehat{y_i^s}\right) + \left(1-y_i^s\right) * \log\left(1-\widehat{y_i^s}\right)\right) \tag{9}$$

where $\widehat{y_i^s}$ and y_i^s are a prediction label and a ground-truth label, respectively.

Concretely, we adopt different backbone networks to extract the image features from the source domain D_S. Noting that VGG16 is used as the backbone for the Messidor dataset whereas Resnet50 is chosen as the backbone for the Eyepacs dataset. However, the above supervised model $C_{\theta_I}(\cdot)$ cannot be directly applied to alternative domains due to the domain shift. Thus, we further consider the adaption technique to generalize the discriminative ability to target domain.

Target domain: Fine-tuning the instance discriminator with pseudo labels.

In the target domain, if we use $C_{\theta_I}(\cdot)$ that is trained only in the source domain, it fails to discriminate the true lesions from the backgrounds due to the domain gap. We will later verify it in the experiment. To mitigate the domain gap, we further fine-tune the instance discriminator with the generated samples by CycleGAN to achieve a progressive domain adaptation.

3.5. *Attention based multi-instance learning for weakly supervised learning*

MIL aims to alleviate the labeling and segmentation burden on the ophthalmologist by learning the mapping between a bag of instances and the bag-level label. We firstly train a MIL model on the target domain and predict the bag labels of unseen images. The major challenge of this MIL lies in building the multi-instance learning mechanism to model the relation between the instances and bag. To solve it, we propose a multi-class multi-instance learning model with an attention mechanism, which can learn a better image representation through a local-global scheme for improved DR graded diagnosis. In the MIL framework, the relationships among instances are very important for learning the mapping between a bag of instances and the bag-level label. Therefore, to avoid that all the patches are treated equivalently when generating bag representations using the instance embeddings, we propose an attention mechanism to selectively learn useful instances and appropriately represent the image embedding with a form of weighted-sum pooling scheme. The interpretation of model decisions can be achieved through the attention mechanism, which will be discussed in Sec. 4.4. More specifically, with the instance embedding $\boldsymbol{h}_i^{(j)} \in \mathbb{R}^D$ learned by the feature learning module in $C_{\theta_M}(\cdot)$, the weight of an individual instance can be calculated through an attention mechanism as follows:

$$\boldsymbol{\alpha}_i^{(j)} = \frac{\exp\left\{\boldsymbol{W}^T\left(\tanh\left(\boldsymbol{V}\left(\boldsymbol{h}_i^{(j)}\right)^T\right)\odot sigm\left(\boldsymbol{U}\left(\boldsymbol{h}_i^{(j)}\right)^T\right)\right)\right\}}{\sum_{k=1}^{B}\exp\left\{\boldsymbol{W}^T\left(\tanh\left(\boldsymbol{V}\left(\boldsymbol{h}_i^{(k)}\right)^T\right)\odot sigm\left(\boldsymbol{U}\left(\boldsymbol{h}_i^{(k)}\right)^T\right)\right)\right\}} \tag{10}$$

where $\boldsymbol{\alpha}_i^{(j)} \in \mathbb{R}^C$ denotes the attention weights for the j-th instances in the i-th image, $\boldsymbol{W} \in \mathbb{R}^{M\times C}$, $\boldsymbol{U} \in \mathbb{R}^{M\times D}$ and $\boldsymbol{V} \in \mathbb{R}^{M\times D}$ are learned parameters, $B = b \times b$ and C denote the patch size in each bag and the class number.

In Eq. (10), the tangent and sigmoid functions are introduced to increase the element-wise non-linearity from two aspects: 1) the tangent function can involve both negative and positive values for appropriate gradient flow; 2) the sigmoid function further removes the troublesome linearity in the tangent function.

With the optimized attention weights, each instance can be expressed by multiple weighted embeddings in Eq. (11), each of which is obtained through the original embedding multiplied by the specific attention weights of the corresponding classes (disease levels).

$$\widehat{\boldsymbol{h}}_i^{(j)} = \boldsymbol{\alpha}_i^{(j)} \boldsymbol{h}_i^{(j)} \tag{11}$$

where $\widehat{\boldsymbol{h}}_i^{(j)} \in \mathbb{R}^{C \times D}$, D represents instance embedding dimensionality.

With the instance weights, we proposed a global bag-level image representation $\widehat{\boldsymbol{H}}_i$ by concatenating the instance embeddings with the attention weights as follows:

$$\widehat{\boldsymbol{H}}_i = \textbf{concat}\left[\widehat{\boldsymbol{h}}_i^{(1)}, \ldots, \widehat{\boldsymbol{h}}_i^{(B)}\right] \tag{12}$$

The contribution of each instance is different from the classification, thus automatically identifying the task-specific regions and neglecting irrelevant regions enables to improve their performance. Finally, the multi-instance learning prediction is obtained according to the process in Fig. 7.

Moreover, we develop a multi-class cross-entropy loss as follows:

$$\mathrm{L}_{C_M} = -\frac{1}{N_T} \sum\nolimits_{i=0}^{N_T} \sum\nolimits_{c=0}^{C} \left(y_{i,c}^t * \log\left(\hat{y}{}_{i,c}^t \right) \right) \tag{13}$$

where N_T is the total number of image level data, C indicates the number of DR grading level, $y_{i,c}^t$ is the DR severity classification label for the image-level annotated data and $\hat{y}{}_{i,c}^t$ indicates the predicted label.

3.6. Training

Thanks to the independence of the steps, we design a learning strategy that can perform these steps simultaneously. Taken together, the overall objective function can be formulated as:

$$\min_{\theta_I, \theta_M, \theta_G} \max_{\theta_Q} L_{\text{total}} = \min_{\theta_I, \theta_M, \theta_G} \max_{\theta_Q} \left(L_{adv} + \lambda_c L_{cyc} + \lambda_g L_{C_M} + \lambda_h L_{C_I} \right) \tag{14}$$

During the learning stage, the model learns the parameters θ_G, θ_M and θ_I by minimizing the overall objective function in Eq. (14). This process enforces the three components to generate realistic instance, learn the discriminative property of suspicious regions, and identify the most relevant task-specific regions. The three components of the model can learn their corresponding parameters θ_G, θ_I and θ_M according to the objective function in an end-to-end manner. This optimal solution corresponds to solving the optimization problem:

$$\{\theta_I, \theta_M, \theta_G, \theta_Q\} = \arg \min_{\theta_I, \theta_M, \theta_G} \max_{\theta_Q} L_{\text{total}} \tag{15}$$

$$\begin{cases} \theta_Q \overset{+}{\leftarrow} - \nabla_{\theta_Q}(\mathrm{L}_Q) \\ \theta_I \overset{+}{\leftarrow} - \nabla_{\theta_I}(\lambda_h \mathrm{L}_{C_I}) \\ \theta_M \overset{+}{\leftarrow} - \nabla_{\theta_M}(\lambda_g \mathrm{L}_{C_M}) \\ \theta_G \overset{+}{\leftarrow} - \nabla_{\theta_G}(\mathrm{L}_Q + \lambda_c \mathrm{L}_{cyc} + \lambda_g \mathrm{L}_G) \end{cases} \tag{16}$$

The main iteration steps of WAD-Net are summarized in Algorithm 1.

Algorithm 1. Weakly-supervised network with Attention mechanism and Domain adaptation(WAD-Net).

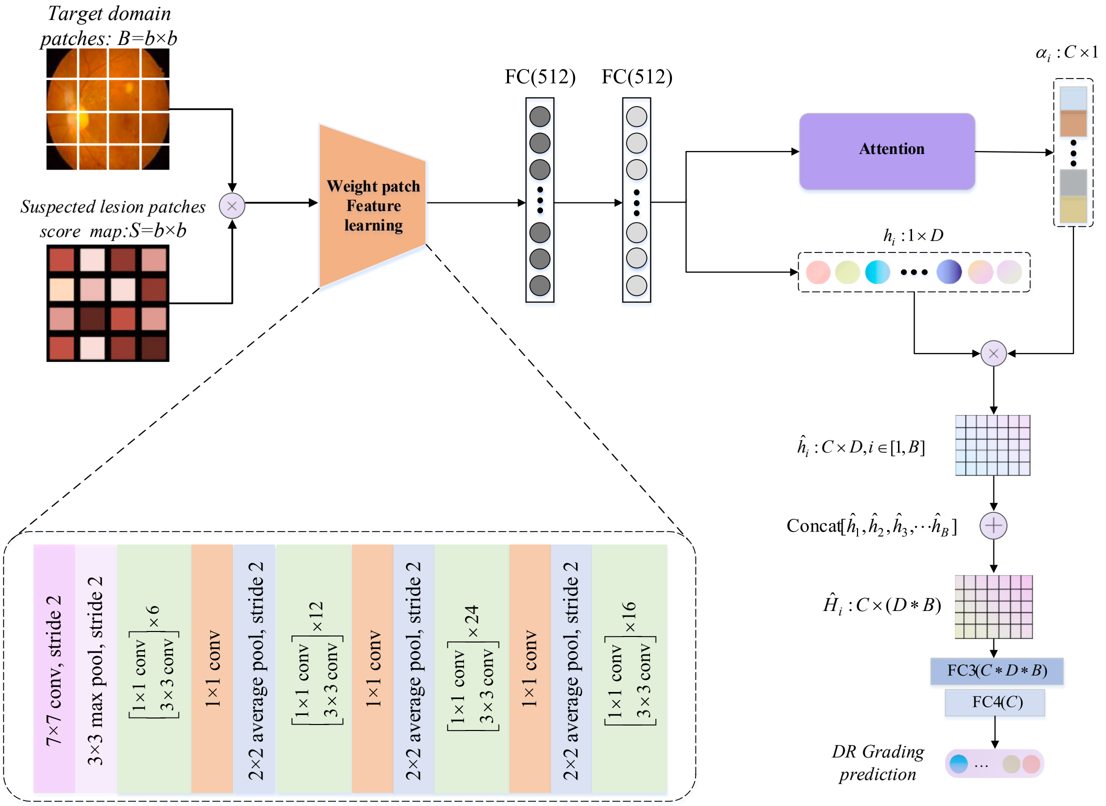

Target domain patches: B=b×b
Suspected lesion patches score map:S=b×b
Weight patch Feature learning
FC(512)
FC(512)
Attention
$\alpha_i : C\times 1$
$h_i : 1\times D$
$\hat{h}_i : C\times D, i\in[1,B]$
Concat$[\hat{h}_1,\hat{h}_2,\hat{h}_3,\cdots\hat{h}_B]$
$\hat{H}_i : C\times(D*B)$
FC3$(C*D*B)$
FC4(C)
DR Grading prediction
7×7 conv, stride 2
3×3 max pool, stride 2
[1×1 conv, 3×3 conv]×6
1×1 conv
2×2 average pool, stride 2
[1×1 conv, 3×3 conv]×12
1×1 conv
2×2 average pool, stride 2
[1×1 conv, 3×3 conv]×24
1×1 conv
2×2 average pool, stride 2
[1×1 conv, 3×3 conv]×16

Fig. 7. Network architectures of the proposed MIL model. The proposed attention mechanism is implemented by an auxiliary layer. At last, a fully connected layer is used to produce the final predictions.

Algorithm 1 Weakly-supervised network with Attention mechanism and Domain adaptation(WAD-Net).

```
Input: Source domain $D_S$ and the training data in the target domain $D_T$
Output: $\theta_I, \theta_G, \theta_M$;
Initialize $\theta_I, \theta_G, \theta_M$.
Stage 1: Pre-training instance discriminator
Update $\theta_I$ with gradient descent according to the loss function in Eq. (8);
Stage 2: Domain adaptation with CycleGAN
Update $\theta_Q$ and $\theta_G$ with gradient descent according to the loss functions in Eq.(4) and Eq.(7);
Stage 3: Fine-tuning with the generated instances
  Update $\theta_I$ with gradient descent according to the loss function in Eq. (8);
Stage 4: MIL with attention mechanism
  Update $\theta_M$ with gradient descent according to the loss function in Eq. (12);
return $\theta_I, \theta_G, \theta_M$
```

4. Experiments

In this section, we conduct several sets of comparative experiments and rigorously analyze our experimental results of DR grading. We use the 10-fold cross-validation to evaluate the proposed method.

4.1. Datasets and performance metrics

In our study, the dataset of Messidor/EyePACS with image grading level annotations are chosen as our target domain and the dataset of IDRiD as our source domain. IDRiD consists of 81 fundus images, with pixel-wise lesion annotations of hemorrhages, microaneurysms, soft exudates, and hard exudates. The Messidor dataset is a public dataset provided by the Messidor program partners [12]. It consists of 1200 retinal images and for each image, two grade information including the DR grade and risk of macular edema are provided. Only retinopathy grades are used in the present work. Kaggle-EyePACS consists of 35 126 training images and 53 576 testing images only containing grading labels [13]. The images are collected from different sources with various lighting conditions and weak annotation quality. Table 1 shows the information of the IDRiD and Messidor/EyePACS datasets.

Table 1
The information of the IDRiD and Messidor/EyePACS datasets.

IDRiD dataset		Messidor dataset		
Lesion	image number	Grade	Description	image number
MA[a]	81	DR-0	$MA = 0$ and $H = 0$	546
H[b]	80	DR-1	$0 < MA \leq 5$ and $H = 0$	153
EX[c]	81	DR-2	$5 < MA < 15$ and $0 < H < 5$	247
SE[d]	40	DR-3	$MA \geq 15$ and $H \geq 5$	254
EyePACS dataset Manifestation	Grade	Train	Test	Image number
No DR	DR-0	25 810	39 533	65 343
Mild	DR-1	2443	3762	6205
Moderate	DR-2	5292	7861	13 153
Severe	DR-3	873	1214	2087
Proliferative DR	DR-4	708	1206	1914

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

We choose the metrics of accuracy, precision, recall, F1-score, kappa and area under the curve (AUC) of ROC to evaluate the performance of our proposed method.

$$ACC = \frac{\sum_{i=0}^{n} TP_i + \sum_{i=0}^{n} TN_i}{\sum_{i=0}^{n} TP_i + \sum_{i=0}^{n} FP_i + \sum_{i=0}^{n} TN_i + \sum_{i=0}^{n} FN_i} \tag{17}$$

$$Precision = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \tag{18}$$

$$Recall = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \tag{19}$$

$$\mathrm{F1} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{20}$$

$$kappa = \frac{P_o - P_e}{1 - P_e}, P_e = \frac{\sum_{i=1}^{C} T_i * P_i}{n^2} \tag{21}$$

where P_o denotes the DR grading accuracy of all fundus images, n is the total number of samples, C indicates the number of DR types, and T_i, P_i denote the number of true and predicted samples for each type of fundus images, respectively.

Table 2
The Comparison between Our Method with the State-of-the-art Methods for DR Grading on Multi-class disease grading classification task.

Methods	Accuracy	Validation	Images
Fractal-based [31]	0.483	5-fold	1200
Expert [33]	0.681	Manual	1200
GLCM/SVM [32]	0.470	5-fold	1200
GLCM/RF [32]	0.459	5-fold	1200
CANet [24]	0.680	10-fold	1200
SKD [30]	0.608	10-fold	1200
WAD-Net(ours)	**0.712**	10-fold	1200

4.2. Comparison with the state-of-the-art methods on the Messidor dataset

4.2.1. Multi-class disease grading

To more comprehensively evaluate our model, we compare WAD-Net with several recent state-of-the-art methods reported on the Messidor dataset in Table 2. The comparison includes two deep learning methods: CANet [24] and SKD [30]. By exploring the internal relationship between the diseases, CANet [24] is proposed to jointly grade DR and DME through a cross-disease attention scheme. SKD [30] is proposed for grading DR with a combination of self-knowledge distillation and CAM-Attention. Both of them are trained with only image-level supervision. Moreover, we compared the traditional machine learning approaches including the Fractal-based feature [31] and GLCM feature combined with different classifiers [32]. Besides, we compare our model with an expert [33].

Experimental results are reported in Table 2 where the best results are boldfaced. Experimental results on the Messidor database demonstrate that the proposed WAD-Net achieves the best performance compared with the other state-of-the-art methods not just in traditional classification methods but also in deep learning methods. It also indicates that our WAD-Net is effective for multi-class DR grading problems. Although the experimental setup in these references is slightly different, it appears that our method performs favorably compared to the previous state-of-the-art.

4.2.2. Binary-class disease diagnosis

We also compare our model with the traditional features based methods, deep learning based methods and experts on the task of binary-class disease diagnosis. The comparable methods involves:

Dynamic Shape Features [34]: A discriminative feature is proposed to describe the shape evolution during image flooding.

Splat feature [35]: An optimal set of features is extracted and selected from each splat to represent the lesion characteristics from a variety of interactions with neighboring splats, filter bank, and shape and texture features.

VNXK/LGI [36]: It is inspired mainly from VGGNet, additionally it combines some components from GoogLeNet and ResNet.

CKML Net/LGI [36]: It is an extension of GoogLeNet.

Zoom-in-Net [37]: It involves three subnetworks: a main network (M-Net) for DR classification, an Attention Network for generating attention maps, and a Crop-Network (C-Net) for correcting the predictions from M-Net with high resolution patches of highest attention values as input.

In Table 3, we present the binary classification task (normal V.S. abnormal) results of various studies. Results obtained for the Messidor database also demonstrate that the proposed method outperforms state-of-the-art methods as well as two ophthalmologists A and B [38].

Table 3
The Comparison between our method with the state-of-the-art methods for DR diagnosis on binary classification task.

Methods	Accuracy	AUC	Sen	Spec
Expert A [38]	-	0.922	**0.945**	0.500
Expert B [38]	-	0.865	0.912	0.500
Comprehensive CAD [38]	-	0.876	0.922	0.500
Splat feature/kNN [35]	-	0.870	-	-
Dynamic Shape Features/Random Forest (RF) [34]	-	0.899	0.939	0.500
DenseNet-201	0.878	0.959	0.878	0.881
ResNet-50	0.878	0.951	0.878	0.905
CKML Net/LGI [36]	0.858	0.862	0.916	0.803
VNXK/LGI [36]	0.871	0.870	0.882	0.857
Zoom-in-Net [37]	0.905	0.921	-	-
WAD-Net(ours)	**0.949**	**0.958**	0.927	**0.957**

Table 4
The evaluation of the three important components of WAD-Net algorithm.

Component	Accuracy	Precision	Recall	micro-F1	AUC
ResNet50	0.253	0.112	0.241	0.193	0.502
MIL	0.489	0.498	0.525	0.511	0.659
WAD-Net w/o fine-tuning	0.606	0.729	0.630	0.670	0.737
WAD-Net w/o cycleGAN	0.507	0.594	0.548	0.561	0.672
WAD-Net w/o attention	0.547	0.574	0.630	0.597	0.698
WAD-Net w/o cycle consistency	0.669	0.673	0.669	0.671	0.779
WAD-Net-ts	**0.764**	**0.765**	0.616	0.676	0.749
WAD-Net	0.712	0.716	**0.713**	**0.713**	**0.808**

4.3. Ablation study

4.3.1. Image reconstruction quality comparison

In addition to the quantitative improvements, we qualitatively compare the images produced by the generator $G_{\theta_G}(\cdot)$ in our model and other GAN methods on the Messidor dataset. Fig. 8 shows the qualitative results of generated samples by different GAN methods including UNIT [39], DualGAN [40], MUNIT [41], DRIT [42], UGATIT [43], DiscoGAN [44]. From Fig. 8, it can be observed that our supervision guided cycleGAN is able to synthesize graphs whose quality are close to real images compared to the competing GAN models.

4.3.2. The effectiveness of each component in WAD-Net

The WAD-Net algorithm mainly involves three components: domain adaptation, instance progressive discriminator, multi-instance learning with an attention mechanism. To investigate the effectiveness of our WAD-Net, we compare WAD-Net with its several variants, respectively.

ResNet50: An image-level classification model is constructed based on ResNet50 in the target domain;

MIL: Only simple MIL model without attention is developed, which treats all patches as instances;

WAD-Net w/o fine-tuning: The instance discriminator is only pre-trained in the source domain without the further fine-tuning process in the target domain;

WAD-Net w/o cycleGAN: The pre-trained instance discriminator is employed to predict the pseudo-labels for the instances in the target domain rather than the generated instances.

WAD-Net w/o attention: The instance discriminator is fine-tuned with the generated instances by CycleGAN. With the fine-tuned model, the instances are filtered and fed into the MIL model without attention mechanism.

WAD-Net w/o cycle consistency: To evaluate the effectiveness of cycle consistency in our WAD-Net, the cycle consistency is removed in WAD-Net.

WAD-Net-ts: Moreover, we compare two different learning strategies of WAD-Net: jointly training and two-step training. The two-step training means that the cycleGAN and the other modules are conducted independently. The two-step training indicates that the unsupervised cycleGAN is trained at first, then the progressive instance discriminator and the multi-instance learning are trained based on the patch instances generated by cycleGAN.

From Table 4, we can see that the proposed method outperforms the baseline and contender methods. These results reveal several interesting points:

(1) ResNet50 shows the worst performance among the algorithms for almost all datasets/metrics. With the limited size of the training set, it is difficult to train a bag-level classification model for DR diagnosis.

(2) MIL without any instance filtering achieves a poor result, which indicates that the large amount of irrelevant instances negatively affects the multi-instance learning. It can be verified that filtering

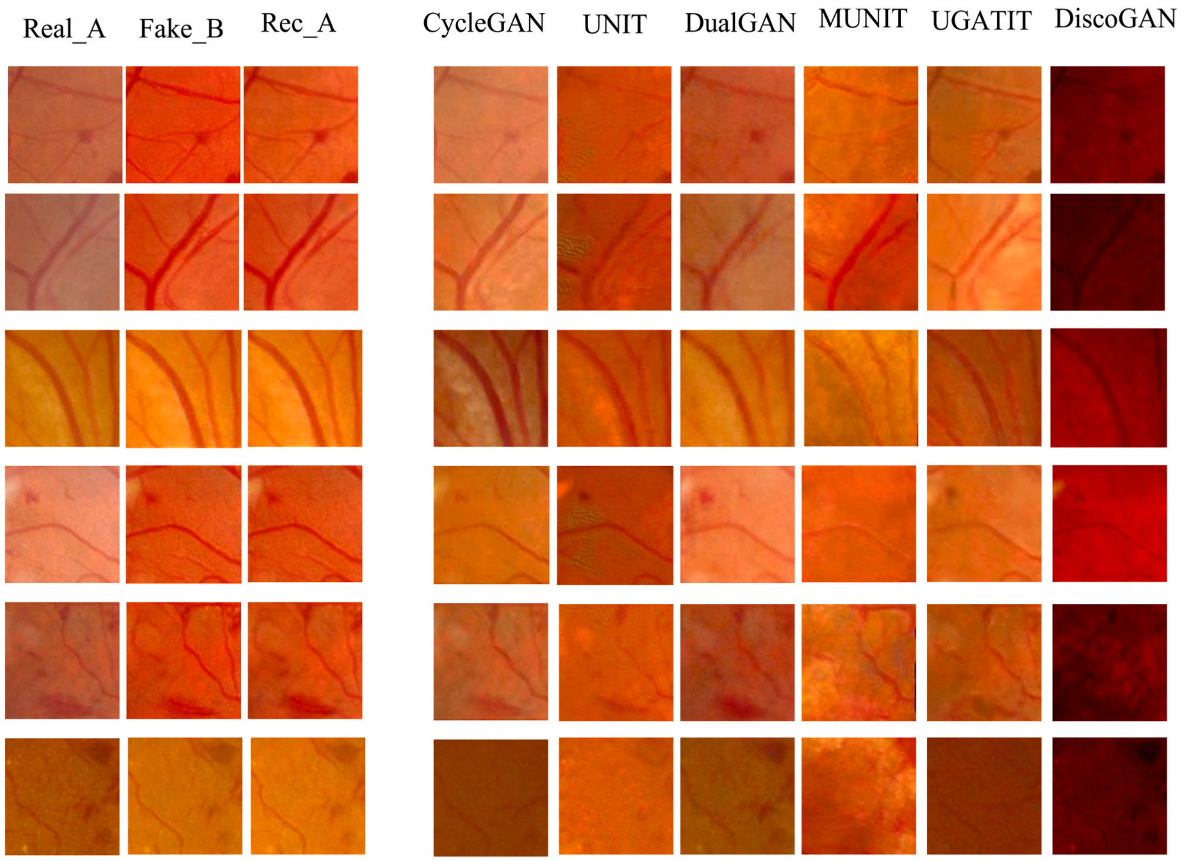


Fig. 8. The reconstruction quality comparison among the multiple GAN models.

Table 5
Influence of the pooling operations on WAD-Net algorithm.

Pooling Method	Accuracy	Precision	Recall	micro-F1	AUC
Attention(ours)	**0**.712	**0**.716	**0**.712	**0**.713	**0**.808
GMP	0.558	0.598	0.586	0.592	0.706
GAP	0.567	0.574	0.620	0.595	0.710
GLP	0.526	0.524	0.552	0.537	0.684

Table 6
Influence of the variation of GAN loss function in the WAD-Net algorithm.

Loss variation	Accuracy	Precision	Recall	micro-F1	AUC
only adv	0.669	0.673	0.669	0.671	0.779
single direction ($S \rightarrow T \rightarrow S$) (ours)	0.712	0.716	**0**.713	**0**.713	**0**.808
two directions ($S \rightarrow T \rightarrow S$) + ($T \rightarrow S \rightarrow T$)	**0**.764	**0**.765	0.616	0.676	0.749

the irrelevant instances contribute to the performance improvements of MIL.

(3) WAD-Net w/o fine-tuning performs worse than the WAD-Net method, which demonstrates that the cross-domain distribution complicates the traditional classification. Thus employing the classifier merely trained on the labeled data from the source domain produces a poor transferring classification performance. Our results seem to yield a piece of solid evidence that imposing a domain adaptation method during the training of the network is a viable method for improving the cross-domain classification performances. Another important conclusion is that WAD-Net w/o cycleGAN may generate wrong labels due to the inconsistent distribution from multi-domain data. It is even worse than WAD-Net w/o fine-tuning. Without considering the cycleGAN for the domain adaptation, the different domain confuses the instance discriminator, which results in lowering the performance of instance filtering.

(4) WAD-Net-ts achieves the best accuracy and precision but a lower recall and AUC, which indicates that the quality of the fake patches generated by the unsupervised cycleGAN is worse than the MIL guided cycleGAN. It verifies that collaboratively learning is critical once again.

(5) Our WAD-Net further improves the performance over WAD-Net w/o attention by incorporating the attention mechanism. Besides the interpretability support provided, it indicates that the contribution of instances is different for the MIL performance.

Besides, we investigate the effectiveness of pooling strategy and the generation module in our framework in Tables 5 and 6. More specifically, we compare our attention method with global max pooling (GMP), global average pooling (GAP) and global log-sum-exp pooling (GLP) [45]. Results in Table 5 show that the proposed pooling strategy with attention provides better performance by emphasizing critical local regions and filtering irrelevant information. Furthermore, we explore the generation module in the cycleGAN based domain adaptation and demonstrate that the role of the consistency loss during the generation is important and only the direction from the source domain to the target domain can sufficiently solve the domain gap.

4.3.3. *Influence of the size of image patches*

In the previous experiments, the patch size is empirically fixed as 128 × 128 for the WAD-Net method. We also investigate the influence of the patch size by comparing the WAD-Net-128 with WAD-Net-256 on

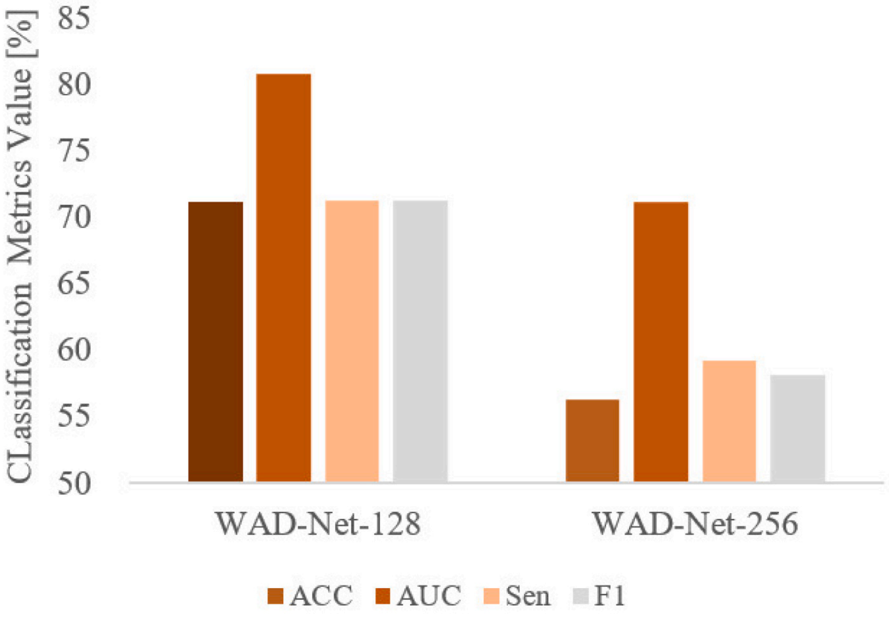


Fig. 9. Influence of the patch size on WAD-Net performance.

the Messidor dataset. From Fig. 9, it can be observed that a better classification performance is obtained with the smaller patch size in terms of all the metrics. It suggests that the patch size is critical for instance representation learning and multi-instance learning performance.

4.4. *Interpretability*

The deep learning-based models can not typically provide interpretability, which lacks the evidence support for doctors. Investigating the performance comparison through ablation studies and quantitative evaluations alone may not be sufficient to fully understand the benefits and behavior of our model. Although the proposed MIL with attention contributes to the performance improvement, it is interesting to investigate the attention mechanism working as expected. Therefore, we improve the interpretability ability of deep learning through the attention mechanism to support the decision-making by producing the location information of highly suspected lesions. Fig. 10 shows the interpretable results of WAD-Net. An attention map is calculated by multiplying the pixel intensity values with the corresponding attention weights of the patches. The lesion regions can be identified through the attention map.

4.5. *Comparison with the state-of-the-art methods on the Eyepacs dataset*

In addition to diagnosing diabetic retinopathy, we also verify the generality of the proposed WAD-Net for DR grading on other datasets. Hence, we also compare our method with other DR grading models reported on the large scale EyePACS dataset in Table 7. We use multiple baselines to evaluate the DR grading performance of our WAD-Net method. The first kinds of baselines adopt a basic classification-only model with different classic backbones, including VGG-16, ResNet-50, Inception v3 and DenseNet-121. The second kind of baselines are ensemble models proposed by the top three places from the Kaggle challenge [46], including Min-pooling, o_O and Reformed Gamblers. Last but not least, we compared the DR grading performance of the WAD-Net method with five other methods on EyePACS datasets. There are two main branches for DR grading: employing auxiliary information [47–49] and multi-task joint learning [50].

Lesion-based CL [48]: Instead of taking entire images as the input, lesion-based CL uses lesion patches to encourage the feature extractor to learn representations that are highly discriminative for DR grading via contrastive learning.

MMCNN [49]: It is proposed to predict the label with both classification and regression to consider the relationships of images with different stages.

DeepMT-DR [50]: It is a hierarchical deep multi-task learning structure that simultaneously processes the low-level task of image super-resolution, the mid-level task of lesion segmentation and the high-level task of DR grading.

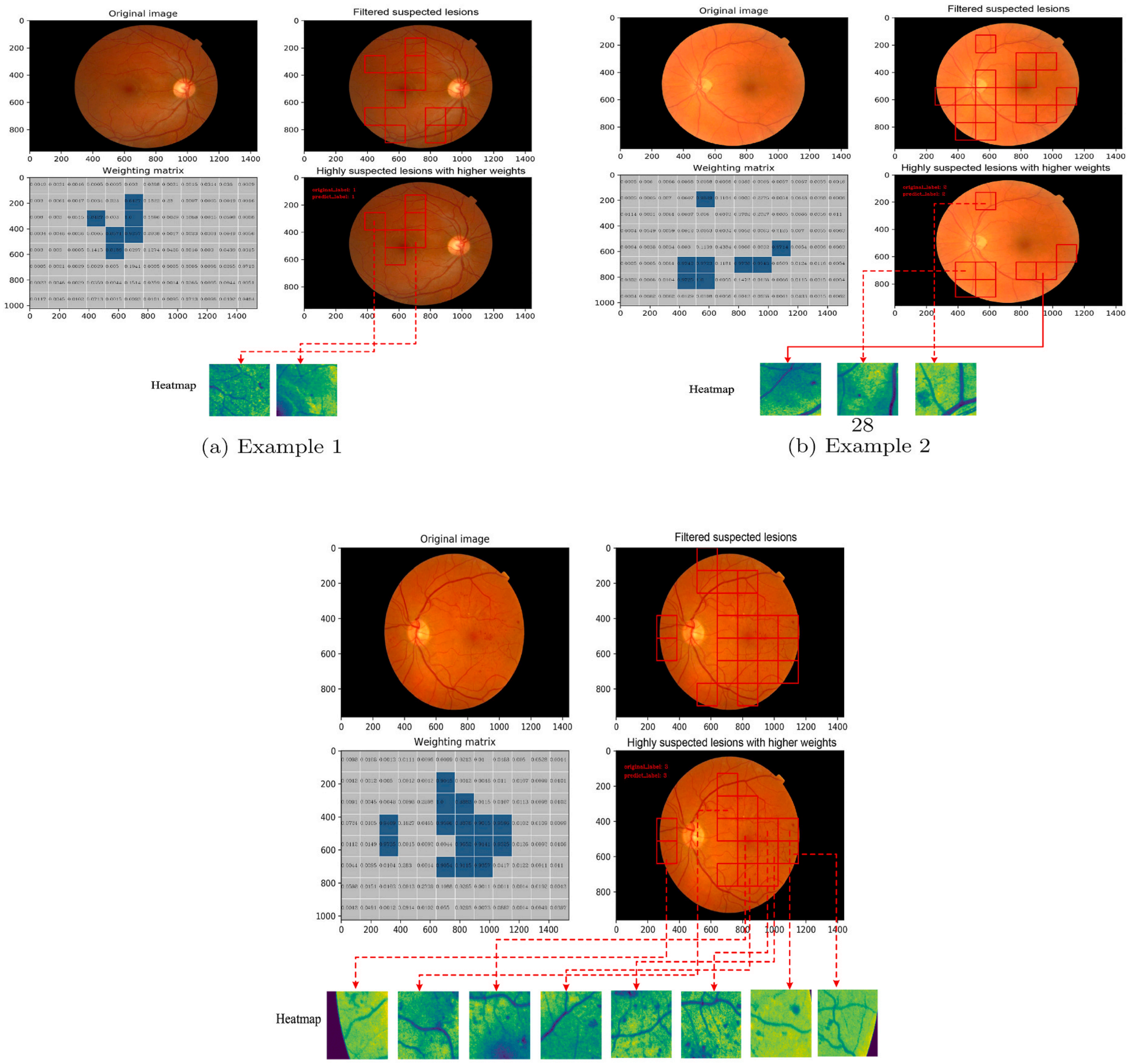
As shown in Table 7, we conduct the comparison experiment on EyePACS dataset to evaluate the performance of our proposed method by quadratic weighted kappa (QWK) which works well for unbalanced datasets and accuracy. From Table 7, we can observe that our model usually achieves competitive performance against the state-of-the-art methods.

5. Discussion

Although state-of-the-art DR grading methods have achieved great success, they did not make full use of all valuable information because of the domain gap and classification performance. We point out four key issues that are previously ignored concerning the severity of DR, and we hope this work can inspire further research on the DR grading prediction.

1. How to leverage the information from another domain?

Existing DR grading methods suffer from a major limitation which is the insufficient available annotations. More and more work leverages information from auxiliary pixel-wise labeled images to improve the performance. However, few works consider the domain gap when borrowing the auxiliary datasets. When the source and target domains are related but from different distributions, simply applying the model trained in one domain into another might negatively affect the learner's performance in the target domain due to the possible shift between training and test samples. In medical application scenarios, the data may come from multiple domains with different distributions. In our study, we regard cross-domain DR grading as a style adaptation with cycle-GAN. It is a weakly-supervised domain adaptation paradigm for DR grading domain adaptation. The aim is to learn a generalized DR grading model in the presence of a shift between source and target domain



(a) Example 1

(b) Example 2

(c) Example 3

Fig. 10. Some examples of attention map of DR grading process.

Table 7
The comparison between our method with the State-of-the-art Methods for DR grading on the Eyepacs dataset.

Methods	Kappa	ACC	Methods	Kappa	ACC
VGG-16	0.819	0.836	Reformed Gamblers	0.839	-
ResNet50	0.823	0.845	Zoom-in-Net (2017) [47]	0.854	-
Inception v3	0.811	0.839	DeepMT-DR(2021) [50]	0.839	0.857
DenseNet-121	0.834	0.853	Lesion-base CL(2021) [48]	0.832	-
Min-pooling	0.849	-	MMCNN(2018) [49]	0.841	-
o_O	0.844	-	WAD-Net(ours)	**0.860**	**0.887**

distributions. This paradigm can be easily extended to other types of medical images with weak label and domain gap characteristics.

2. How to improve the classification model interpretability ?

Another advantage of our model is simultaneously grading DR and highlighting lesion regions. Identifying suspicious regions for medical images is of significant importance since it provides intuitive illustrations for physicians and patients of how the diagnosis is made. Despite the good performance achieved by the state-of-the-art deep learning method, the major limitation is that the networks are trained with only image-level supervision, making it very challenging to find the accurate abnormal signs, such as soft exudates, hard exudates, microaneurysms, and hemorrhage. The identification of lesion regions in fundus images is also very important, since it provides visual clues for ophthalmologists to assist their diagnosis. The previous deep learning based models trained on the Messidor or Eyepacs datasets can only be used to predict a severity grade without providing any interpretability for ophthalmologists. Our key contributions include the attention mechanism in the MIL model to automatically extract the task-specific regions and neglect the irrelevant information to improve their performance. Through the attention mechanism, our model can simultaneously grade DR and highlight lesion regions by generating attention maps which highlight suspicious regions trained with only image-level supervisions. The visual attention mechanism enables our model to act in a clinicians-like manner, and automatically discover the suspicious regions in the image.

3. How to design a unified DR grading framework collaborated with other networks?

Compared with existing deep learning based methods, our method aims at improving the grading performance by combining domain adaptation, irrelevant patch filtering and MIL classification. The common strategies are that treating these components as independent tasks. This study explores a new perspective: could it be simultaneously trained in an end-to-end fashion? Our model attempts to integrate the three components: domain adaptation, lesion (instance) classification and DR grading into an end-to-end training system, and yields higher grading accuracy. Please refer to Figs. 3 and 4. The key contribution of the proposed framework lies in the joint training, which can improve the generalization performance from two aspects: generating more realistic patches and producing more accurate patch scores. More specifically, in the collaborative learning framework, the classification loss on the target domain encourages the domain adaption model to generate the instances which are more beneficial to the classification. Meanwhile, the classification loss also helps the instance classification to produce more accurate scores and focus on the relevant patches for its subsequent multi-instance learning task. Moreover, the improved performance confirms that the three tasks are correlated and they can benefit from each other. Our results suggest that the relevant tasks could be jointly trained to improve DR grading performance.

4. Image (global) level or patch (local) level ? How to develop a patch-Aware DR grading framework?

DR diagnosis in clinic highly depends on the detected retinal pathologies, such as microaneurysm. Identifying suspicious regions for medical images is of significant importance since it provides intuitive illustrations for physicians and patients of how the diagnosis decision is obtained. However, the lesions may only occupy a small part of the whole fundus image. Higher-level features with larger receptive fields have an abstract semantic information, tending to ignore the small lesions. To address the aforementioned issues, we develop a lesion-aware framework by focusing on the patches as instances for DR severity grading on fundus image. Instead of using entire fundus images, patches combined with the suspected lesion scores are taken as the input for our MIL model. Moreover, the domain adaptation and instance classifier components are also conducted from the local view, both of which enable our model to be lesion-aware. By focusing on patches, the network is encouraged to learn more discriminative features. Therefore, integrating the lesion-aware components is a effective exploration for DR grading.

6. Conclusion

Focusing on improving the performance of DR diagnosis and grading when the lesion labeling is scarce, we formulated the problem of DR grading as an MIL problem, and propose an interpretable end-to-end MIL network with attention mechanism and domain adaptation for simultaneously diagnosing diabetic retinopathy and highlighting suspicious regions. By combining the strengths of domain adaptation and multi-instance learning, the proposed approach significantly improved the state-of-the-art results in DR grading on benchmark datasets.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (No.62076059) and the Science Project of Liaoning provice (2021-MS-105).

References

[1] E. AbdelMaksoud, S. Barakat, M. Elmogy, A comprehensive diagnosis system for early signs and different diabetic retinopathy grades using fundus retinal images based on pathological changes detection, Comput. Biol. Med. 126 (2020) 104039.

[2] N. Tsiknakis, D. Theodoropoulos, G. Manikis, E. Ktistakis, O. Boutsora, A. Berto, F. Scarpa, A. Scarpa, D.I. Fotiadis, K. Marias, Deep learning for diabetic retinopathy detection and classification based on fundus images: a review, Comput. Biol. Med. (2021) 104599.

[3] A. Garifullin, L. Lensu, H. Uusitalo, Deep bayesian baseline for segmenting diabetic retinopathy lesions: advances and challenges, Comput. Biol. Med. (2021) 104725.

[4] G.T. Zago, R.V. Andreão, B. Dorizzi, E.O.T. Salles, Diabetic retinopathy detection using red lesion localization and convolutional neural networks, Comput. Biol. Med. 116 (2020) 103537.

[5] A. Sugeno, Y. Ishikawa, T. Ohshima, R. Muramatsu, Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning, Comput. Biol. Med. 137 (2021) 104795.

[6] M. Kandemir, F.A. Hamprecht, Computer-aided diagnosis from weak supervision: a benchmarking study, Comput. Med. Imag. Graph. 42 (2015) 44–50.

[7] G. Quellec, G. Cazuguel, B. Cochener, M. Lamard, Multiple-instance learning for medical image and video analysis, IEEE Rev. Biomed. Eng. (99) (2017), 1–1.

[8] B. Antal, A. Hajdu, An ensemble-based system for automatic screening of diabetic retinopathy, Knowl. Base Syst. 60 (2014) 20–27.

[9] S. Roychowdhury, D.D. Koozekanani, K.K. Parhi, Dream: diabetic retinopathy analysis using machine learning, IEEE J. Biomed. Health Inf. 18 (5) (2014) 1717–1728.

[10] P. Cao, F. Ren, C. Wan, J. Yang, O. Zaiane, Efficient multi-kernel multi-instance learning using weakly supervised and imbalanced data for diabetic retinopathy diagnosis, Comput. Med. Imag. Graph. 69 (2018) 112–124.

[11] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[12] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, et al., Feedback on a publicly distributed image database: the messidor database, Image Anal. Stereol. 33 (3) (2014) 231–234.

[13] B. Graham, Kaggle Diabetic Retinopathy Detection Competition Report, University of Warwick, 2015.

[14] V. Mayya, S.S. Kamath, U. Kulkarni, Automated Microaneurysms Detection for Early Diagnosis of Diabetic Retinopathy: A Comprehensive Review, Computer Methods and Programs in Biomedicine Update, 2021, p. 100013.

[15] P. Chudzik, S. Majumdar, F. Calivá, B. Al-Diri, A. Hunter, Microaneurysm detection using fully convolutional neural networks, Comput. Methods Progr. Biomed. 158 (2018) 185–192.

[16] K. Adem, Exudate detection for diabetic retinopathy with circular hough transformation and convolutional neural networks, Expert Syst. Appl. 114 (2018) 289–295.

[17] Y. Yan, J. Gong, Y. Liu, A novel deep learning method for red lesions detection using hybrid feature, in: 2019 Chinese Control and Decision Conference (CCDC), IEEE, 2019, pp. 2287–2292.

[18] M.J. Van Grinsven, B. van Ginneken, C.B. Hoyng, T. Theelen, C.I. Sánchez, Fast convolutional neural network training using selective data sampling: application to hemorrhage detection in color fundus images, IEEE Trans. Med. Imag. 35 (5) (2016) 1273–1284.

[19] Y. Yang, T. Li, W. Li, H. Wu, W. Fan, W. Zhang, Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 533–540.

[20] Z. Lin, R. Guo, Y. Wang, B. Wu, T. Chen, W. Wang, D.Z. Chen, J. Wu, A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 74–82.

[21] T. Nazir, A. Irtaza, A. Javed, H. Malik, D. Hussain, R.A. Naqvi, Retinal image analysis for diabetes-based eye disease detection using deep learning, Appl. Sci. 10 (18) (2020) 6185.

[22] A. Bora, S. Balasubramanian, B. Babenko, S. Virmani, S. Venugopalan, A. Mitani, G. de Oliveira Marinho, J. Cuadros, P. Ruamviboonsuk, G.S. Corrado, et al., Predicting the risk of developing diabetic retinopathy using deep learning, Lancet Digit. Health 3 (1) (2021) e10–e19.

[23] Y. Zhou, X. He, L. Huang, L. Liu, F. Zhu, S. Cui, L. Shao, Collaborative learning of semi-supervised segmentation and classification for medical images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2079–2088.

[24] X. Li, X. Hu, L. Yu, L. Zhu, C.-W. Fu, P.-A. Heng, Canet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading, IEEE Trans. Med. Imag. 39 (5) (2019) 1483–1493.

[25] H. Jiang, J. Xu, R. Shi, K. Yang, D. Zhang, M. Gao, H. Ma, W. Qian, A multi-label deep learning model with interpretable grad-cam for diabetic retinopathy classification, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2020, pp. 1560–1563.

[26] J. Wang, Y. Bai, B. Xia, Simultaneous diagnosis of severity and features of diabetic retinopathy in fundus photography using deep learning, IEEE J. Biomed. Health Inf. 24 (12) (2020) 3397–3407.

[27] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, M. Lamard, Deep image mining for diabetic retinopathy screening, Med. Image Anal. 39 (2017) 178–193.

[28] S. Wang, X. Wang, Y. Hu, Y. Shen, Z. Yang, M. Gan, B. Lei, Diabetic retinopathy diagnosis using multichannel generative adversarial network with semisupervision, IEEE Trans. Autom. Sci. Eng. 18 (2) (2020) 574–585.

[29] L. Zhou, Y. Zhao, J. Yang, Q. Yu, X. Xu, Deep multiple instance learning for automatic detection of diabetic retinopathy in retinal images, IET Image Process. 12 (4) (2018) 563–571.

[30] L. Luo, D. Xue, X. Feng, Automatic diabetic retinopathy grading via self-knowledge distillation, Electronics 9 (9) (2020) 1337.

[31] F. Alzami, R.A. Megantara, A.Z. Fanani, et al., Diabetic retinopathy grade classification based on fractal analysis and random forest, in: 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), IEEE, 2019, pp. 272–276.

[32] J.D. Labhade, L. Chouthmol, S. Deshmukh, Diabetic retinopathy detection using soft computing techniques, in: 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), IEEE, 2016, pp. 175–178.

[33] L. Seoud, J. Chelbi, F. Cheriet, Automatic Grading of Diabetic Retinopathy on a Public Database, 2015, pp. 97–104.

[34] L. Seoud, T. Hurtut, J. Chelbi, F. Cheriet, J.P. Langlois, Red lesion detection using dynamic shape features for diabetic retinopathy screening, IEEE Trans. Med. Imag. 35 (4) (2015) 1116–1126.

[35] L. Tang, M. Niemeijer, J.M. Reinhardt, M.K. Garvin, M.D. Abramoff, Splat feature classification with application to retinal hemorrhage detection in fundus images, IEEE Trans. Med. Imag. 32 (2) (2012) 364–375.

[36] H.H. Vo, A. Verma, New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space, in: 2016 IEEE International Symposium on Multimedia (ISM), IEEE, 2016, pp. 209–215.

[37] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, 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.

[38] C.I. Sánchez, M. Niemeijer, A.V. Dumitrescu, M.S. Suttorp-Schulten, M. D. Abramoff, B. van Ginneken, Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data, Investig. Ophthalmol. Vis. Sci. 52 (7) (2011) 4866–4871.

[39] M.-Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: Advances in Neural Information Processing Systems, 2017, pp. 700–708.

[40] Z. Yi, H. Zhang, P. Tan, M. Gong, Dualgan: unsupervised dual learning for image-to-image translation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2849–2857.

[41] X. Huang, M.-Y. Liu, S. Belongie, J. Kautz, Multimodal unsupervised image-to-image translation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 172–189.

[42] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, M.-H. Yang, Diverse image-to-image translation via disentangled representations, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 35–51.

[43] J. Kim, M. Kim, H. Kang, K. Lee, U-gat-it: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-To-Image Translation, 2019 arXiv preprint arXiv:1907.10830.

[44] T. Kim, M. Cha, H. Kim, J.K. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1857–1865.

[45] X. Wang, Y. Yan, P. Tang, X. Bai, W. Liu, Revisiting multiple instance neural networks, Pattern Recogn. 74 (2018) 15–24.

[46] Kaggle diabetic retinopathy detection competition. https://www.kaggle.com/c/diabetic-retinopathy-detection.

[47] W. Zhe, Y. Yin, J. Shi, W. Fang, H. Li, X. Wang, Zoom-in-net: deep mining lesions for diabetic retinopathy detection, in: Medical Image Computing and Computer Assisted Intervention, MICCAI, 2017, pp. 267–275.

[48] Y. Huang, L. Lin, P. Cheng, J. Lyu, X. Tang, Lesion-based contrastive learning for diabetic retinopathy grading from fundus images, in: Medical Image Computing and Computer Assisted Intervention, MICCAI, 2021, pp. 113–123.

[49] K. Zhou, Z. Gu, W. Liu, W. Luo, J. Cheng, S. Gao, J. Liu, Multi-cell multi-task convolutional neural networks for diabetic retinopathy grading, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, 2018, pp. 2724–2727.

[50] X. Wang, M. Xu, J. Zhang, L. Jiang, L. Li, Deep multi-task learning for diabetic retinopathy grading in fundus images, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, 2021, pp. 2826–2834.