Collaborative Learning of Semi-Supervised Segmentation and Classification for Medical Images

Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui and Ling Shao
Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE
{yi.zhou, xiaodong.he, lei.huang, li.liu, fan.zhu, shanshan.cui, ling.shao}@inceptioniai.org

Abstract

Medical image analysis has two important research areas: disease grading and fine-grained lesion segmentation. Although the former problem often relies on the latter, the two are usually studied separately. Disease severity grading can be treated as a classification problem, which only requires image-level annotations, while the lesion segmentation requires stronger pixel-level annotations. However, pixel-wise data annotation for medical images is highly time-consuming and requires domain experts. In this paper, we propose a collaborative learning method to jointly improve the performance of disease grading and lesion segmentation by semi-supervised learning with an attention mechanism. Given a small set of pixel-level annotated data, a multi-lesion mask generation model first performs the traditional semantic segmentation task. Then, based on initially predicted lesion maps for large quantities of image-level annotated data, a lesion attentive disease grading model is designed to improve the severity classification accuracy. Meanwhile, the lesion attention model can refine the lesion maps using class-specific information to fine-tune the segmentation model in a semi-supervised manner. An adversarial architecture is also integrated for training. With extensive experiments on a representative medical problem called diabetic retinopathy (DR), we validate the effectiveness of our method and achieve consistent improvements over state-of-the-art methods on three public datasets.

1. Introduction

In the medical imaging community, automatic disease diagnosis has been widely explored and applied to various practical computer-aided medical systems. Disease grading [28, 6, 50, 40] and pixel-wise lesion segmentation [12, 10, 29] are two main fundamental problems in this area. The goal of disease grading is to predict the classification label for the severity of a disease, while segmentation aims to address more fine-grained, pixel-wise lesion detection. These two tasks are usually studied independently. However, accurate lesion detection can make huge contributions to classifying the disease grades, while class-specific information can also benefit segmentation performance.

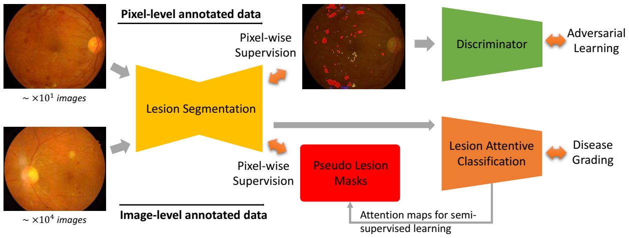

Figure 1. Illustration of the proposed collaborative learning method of semi-supervised multi-lesion segmentation and disease severity classification. Here we conduct studies on the fundus images for diabetic retinopathy.

Labeling medical images is expensive since it requires the very time-consuming dedication of domain experts, especially for pixel-level annotations. Compared with general object segmentation tasks [23, 16, 45, 51, 8], which have large amounts of annotated training data available, employing a fully-supervised architectures [26, 9, 20] to train medical models is impractical. However, purely-unsupervised learning approaches [39, 12, 55] are also not acceptable due to their limited accuracy. As such, we aim to develop a semi-supervised method [34, 35], which can use the limited number of pixel-level annotated images available along with the large quantities of broader, image-level annotations to simultaneously enhance the performance of both the segmentation and classification models.

In this paper, we propose a collaborative learning method for disease grading and lesion segmentation and select a common disease called diabetic retinopathy (DR) for evaluation. DR is an eye disease that results from diabetes mellitus, and can lead to blindness. The severity of DR can be graded into five stages: normal, mild, moderate, severe non-proliferative and proliferative, according to international protocol [19, 3]. The severity grading has strong correlations with different lesion symptoms, such as microaneurysms, haemorrhages, hard exudates and soft exudates appearing on the fundus images. Therefore, multi-lesion segmentation is highly beneficial for analyzing DR gradings. However, since acquiring large quantaties of pixel-level lesion annotation is difficult, a semi-supervised segmentation method is proposed together with image-level severity classification for joint optimization. Fig. 1 illustrates the idea of our proposed method. For images with pixel-level lesion annotations, we pre-train a segmentation model in a fully-supervised manner. Then, a large number of images with only disease grade labels can be passed through the pre-trained segmentation model to generate weak lesion maps. We take the predicted masks with the original images as inputs for learning the lesion attentive classification model. This model both improves the disease grading performance and outputs lesion attention maps as refined pseudo masks that can be used to fine-tune the segmentation model. The main contributions of our method are highlighted as follows:

(1) A multi-lesion mask generator is proposed for the pixel-wise segmentation. Due to extremely limited training data, we carefully design an Xception-module based U-shape network and a joint objective function that incorporates a supervised segmentation loss and an unsupervised adversarial loss for training.

(2) For image-level annotated data, we devise a lesion attention model that can automatically predict lesion maps adopting only weak supervision of class-specific information. The predicted maps can be used to fine-tune the previous segmentation model together with fully-annotated data in a semi-supervised learning manner.

(3) The lesion segmentation and disease grading tasks are optimized in an end-to-end model. The massive amount of class-annotated data can benefit the segmentation performance. Meanwhile, enhanced pixel-wise lesion segmentation can improve grading accuracy. Extensive ablation studies and comparison experiments conducted on the DR data have shown the effectiveness and superiority of our method.

2. Related Work

Disease Grading and Lesion Detection. Recent state-of-the-art disease grading and lesion detection methods in medical imaging tend to adopt general deep learning models. CNN architectures [33, 15, 19] have been proposed to diagnose DR by classifying its severity. Feature selection from the determined features [17] has been introduced for classifying breast cancer malignancy. Moreover, to recognize detailed lesions, bounding-box level detection [41, 49] and pixel-level segmentation [13, 36] have also been studied. However, only a few works [50, 46, 5] have associated the lesion detection and disease severity classification.

Semi-Supervised Semantic Segmentation. For the semantic segmentation task, due to the shortage of pixel-level annotated data, semi-supervised segmentation methods [23, 47, 31] have been explored. An adversarial learning mechanism was used in [22], where the network's discriminator outputs predicted probability maps as the confidence maps for semi-supervised learning. Hong *et al.* [21] proposed a decoupled deep neural network to separate the classification and segmentation networks, and used bridging layers to deliver class-specific information.

Attention Mechanisms. Visual attention addresses the problem of extracting task-specific salient regions from images, while ignoring irrelevant parts. Attention mechanisms have been studied for many vision tasks such as image classification [38, 27, 53, 52], fine-grained recognition [54, 48] and image captioning [4, 7]. These mechanisms can be categorized into soft and hard attention models, where the former is fully-differentiable to learning attention maps and the latter is not differentiable and involves a stochastic process sampling hidden states with probabilities.

3. Proposed Methods

3.1. Problem Formulation

Given pixel-level annotated images $\mathbf{X}^P$ and image-level annotated images $\mathbf{X}^I$, the final aim of our method is to collaboratively optimize a lesion segmentation model $G(\cdot)$ and a disease grading model $C(\cdot)$, which would work together to improve the precision of one another. To train the segmentation model, we aim to minimize the difference between the predicted lesion maps and the ground-truth masks by the following function:

$$\min_{G} \sum_{l=1}^{L} \mathcal{L}_{Seg}(G(\mathbf{X}^P), G(\mathbf{X}^I), \mathbf{s}_l^P, \widetilde{\mathbf{s}}_l^I), \tag{1}$$

where $\mathbf{s}_l^P$ denotes the ground-truth of pixel-level annotated images and $\widetilde{\mathbf{s}}_l^I$ is the pseudo masks of image-level annotated images learned by the lesion attentive grading model. L is the total number of lesion varieties related to a particular disease. The optimization function for the disease grading model is defined as:

$$\min_{C,att} \mathcal{L}_{Cls}(C(\mathbf{X}^I) \cdot att(G(\mathbf{X}^I)), \mathbf{y}^I), \tag{2}$$

where $att(\cdot)$ indicates the lesion attention model and $\mathbf{y}^I$ is the disease severity classification label for image-level annotated data. Note that $\widetilde{\mathbf{s}}_l^I$ in Eq. 1 is equal to $att(G(\mathbf{X}^I))$. The detailed definitions of $\mathcal{L}_{Seg}$ and $\mathcal{L}_{Cls}$ are explained in Sec. 3.2 and 3.3, respectively. Therefore, to collaboratively learn the two tasks, the most important factor to consider is how to design and optimize $G(\cdot)$, $C(\cdot)$ and $att(\cdot)$.

An overview of the proposed network architecture, which consists of two parts, is illustrated in Fig. 2. For the first part, we take the few $\mathbf{X}^P$ as inputs to train a multi-lesion mask generator in a fully-supervised manner. Once it is pre-trained, the remaining large-scale $\mathbf{X}^I$ are also passed

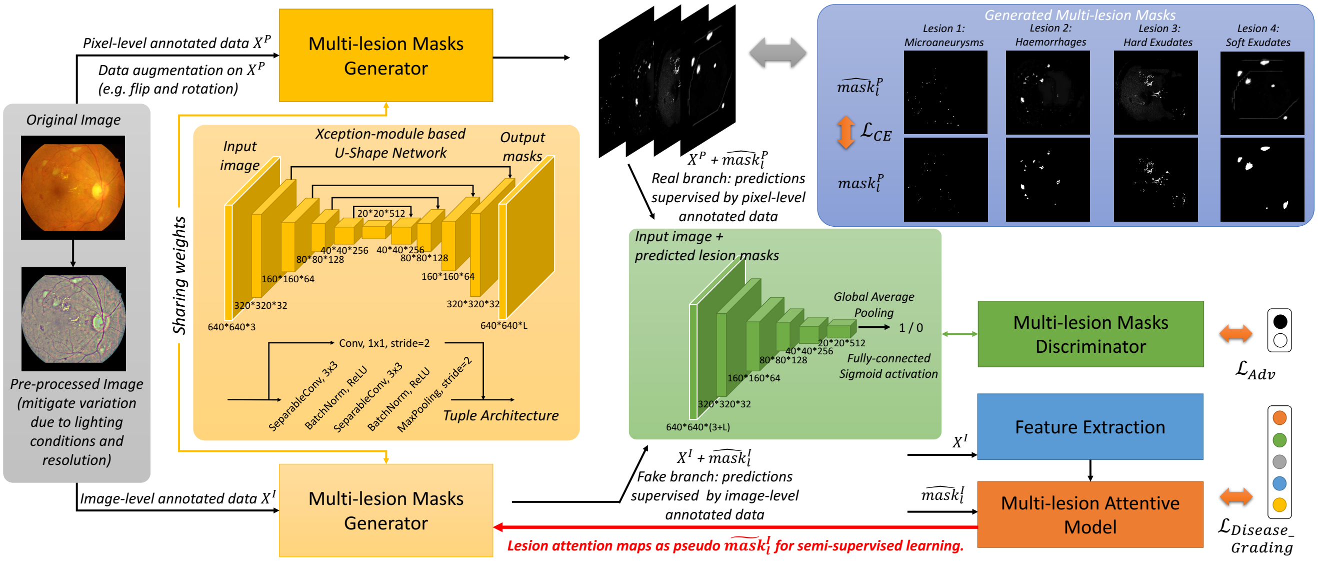
Pixel-level annotated data X^P
Data augmentation on X^P (e.g. flip and rotation)
Multi-lesion Masks Generator
Original Image
Pre-processed Image (mitigate variation due to lighting conditions and resolution)
Sharing weights
Input image
Xception-module based U-Shape Network
Output masks
20*20*512
40*40*256
40*40*256
80*80*128
80*80*128
160*160*64
160*160*64
320*320*32
320*320*32
640*640*3
640*640*L
Conv, 1x1, stride=2
SeparableConv, 3x3
BatchNorm, ReLU
SeparableConv, 3x3
BatchNorm, ReLU
MaxPooling, stride=2
Tuple Architecture
Image-level annotated data X^I
Multi-lesion Masks Generator
Generated Multi-lesion Masks
Lesion 1: Microaneurysms
Lesion 2: Haemorrhages
Lesion 3: Hard Exudates
Lesion 4: Soft Exudates
$\widehat{mask}_l^P$
$\mathcal{L}_{CE}$
$mask_l^P$
$X^P + \widehat{mask}_l^P$
Real branch: predictions supervised by pixel-level annotated data
Input image + predicted lesion masks
Global Average Pooling
1 / 0
20*20*512
40*40*256
80*80*128
160*160*64
320*320*32
640*640*(3+L)
Fully-connected Sigmoid activation
Multi-lesion Masks Discriminator
$\mathcal{L}_{Adv}$
X^I
Feature Extraction
$\widehat{mask}_l^I$
Multi-lesion Attentive Model
$\mathcal{L}_{Disease\_Grading}$
$X^I + \widehat{mask}_l^I$
Fake branch: predictions supervised by image-level annotated data
Lesion attention maps as pseudo $\widetilde{mask}_l^I$ for semi-supervised learning.

Figure 2. Pipeline of the proposed method. The input data consists of a very small set of pixel-level annotated lesion images $\mathbf{X}^P$ and a large set of images $\mathbf{X}^I$ with only image-level labels showing the disease severity. A multi-lesion masks generator is proposed for learning the lesion segmentation task in a semi-supervised manner, where $\mathbf{X}^P$ has real ground-truth masks and $\mathbf{X}^I$ uses the pseudo masks learned from the lesion attentive disease grading model. An adversarial architecture is also proposed to benefit the training. Moreover, the segmented lesion masks are adopted to generate attentive features for improving the final disease grading performance. The two tasks are jointly optimized in an end-to-end network.

through the generator. A discriminator, optimized by an adversarial training loss, is designed to distinguish these two types of data. For the second part, the $\mathbf{X}^I$ and its initially predicted lesion maps are adopted to learn a lesion attention model, which only employs disease grading labels. The lesion attentive grading model improves the classification accuracy. Moreover, the generated lesion attention maps can be used as pseudo masks to refine the lesion mask generator for large unannotated data in a semi-supervised manner.

3.2. Adversarial Multi-Lesion Masks Generator

Training a semantic segmentation model usually requires large quantities of pixel-level annotated data. However, for medical imaging where annotation cost is extremely high, we have to find a way of effectively training a model using the limited annotated data available. In our method, we propose a multi-lesion mask generator, derived from a U-shape network and embedded with an Xception module [11] for this task. The U-shape network [36] was first introduced for the segmentation of neuron structures in electron microscopic stacks. It deploys an encoder-decoder structure built with a fully convolutional network. The skip connections concatenate the feature maps of contracting and expansive parts of the same spatial size. This design can best preserve the edge and texture details in the decoding process of the input images and speed up the convergence.

We first extend the U-shape network with a built-in Xception module and modify it to be a multi-lesion mask generator. The Xception module essentially inherits its idea from the Inception module [42], with the difference being that a separable convolution performs the spatial convolution over each channel and the 1×1 convolution projects new channels independently. We incorporate the Xception module for lesion segmentation since the spatial correlations over each channel of feature maps and the cross-channel correlations have less inner relationship and are not expected to jointly learn the mappings. A schematic diagram of the segmentation model is shown in the yellow part of Fig. 2. Together, the encoder and decoder consist of a total of nine feature mapping tuples. Apart from the first tuple of the encoder, which employs normal convolution operations, the remaining tuples are designed with the Xception module. Each tuple is composed of two separable convolutions followed by batch normalization, ReLU activation, max-pooling and a shortcut of 1×1 convolution. The spatial convolution kernel size is 3×3 and the padding is set to be the same. In the decoder part, up-sampling and a skip connection are employed before each tuple. At the end, we add L convolution layers with Sigmoid activation to generate L different lesion masks. Other hyper-parameter settings are based on [36].

To optimize the lesion mask generator, we use both the pixel-level annotated data and the image-level annotated data. With pixel-level annotated lesion masks, a binary cross-entropy loss $\mathcal{L}_{CE}$ is used to minimize distances between the predictions and the ground-truths. Based on the lesion attention model introduced in Sec. 3.3, we also obtain pseudo mask ground-truths for the image-level annotated data to optimize $\mathcal{L}_{CE}$. Moreover, to generate better lesion masks by exploiting data without pixel-level annotations, we add a multi-lesion discriminator D, which contributes to the training through a generative adversarial network (GAN [18]) architecture. Traditional GANs consist of a generative net and discriminative net playing a competitive min-max game. A latent random vector z from a uniform or Gaussian distribution is usually used as the input for the generator to synthesize samples. The discriminator then aims to distinguish the real data x from the generated samples. The essential goal is to converge $p_z(z)$ to a target real data distribution $p_{data}(x)$. In this paper, rather than generating samples from random noise, we take the lesion maps predicted by the generator from the pixel-level annotated data as the real data branch and those from the image-level annotated data as the fake sample branch. The total loss for optimizing the lesion segmentation task can be defined as:

$$\begin{aligned}&\mathcal{L}_{Seg} = \mathcal{L}_{Adv} + \lambda\mathcal{L}_{CE} \qquad (3)\\ &= \mathbb{E}[\log(D(\mathbf{X}^P, G(\mathbf{X}^P))] + \mathbb{E}[\log(1 - D(\mathbf{X}^I, G(\mathbf{X}^I))]\\ &+ \lambda\mathbb{E}[-\mathbf{s}\cdot\log G(\mathbf{X}^{(P,I)} - (1-\mathbf{s})\cdot\log(1 - G(\mathbf{X}^{(P,I)}))],\end{aligned}$$

where $\mathbf{s}$ is a brief expression of $\mathbf{s}_l^P$ and $\widetilde{\mathbf{s}}_l^I$ for the ground-truths of pixel-level and image-level annotated data, respectively. λ is the balance weight of two objective functions.

The predicted multi-lesion masks are concatenated with the input images and then taken as inputs for the discriminator D which has five convolution mapping tuples. Each tuple consists of two convolutional layers with kernel size of 3 and one max-pooling layer with a stride of 2 to progressively encode contextual information for an increasing receptive field. For each tuple, we also adopt ReLU activation and batch normalization. A global average pooling is employed at the end of D, followed by a dense connection and Sigmoid activation that outputs if the predicted lesion maps are supervised by real or pseudo mask ground-truths.

3.3. Lesion Attentive Disease Grading

To grade the severity of a disease, human experts usually make a diagnosis by observing detailed lesion signs characteristic of the disease. Adopting a classic deep classification model can achieve basic performance for this, but with limited accuracy. Visual attention models address recognition tasks in a human-like manner, automatically extracting task-specific regions and neglecting irrelevant information to improve their performance. However, most conventional attention models are proposed for general object images, and only need to predict coarse attention maps. The attention mechanism is usually designed using high-level features. For medical images, where the lesion regions are very small and are expected to be attended in a pixel-wise manner, in our model we also adopt low-level feature maps with high resolutions to guide the learning of the attention model. Moreover, for those images with only image-level disease grade annotations, our lesion attentive model can generate pixel-level attention maps, which are then used as the pseudo masks for semi-supervised learning in the lesion segmentation model.

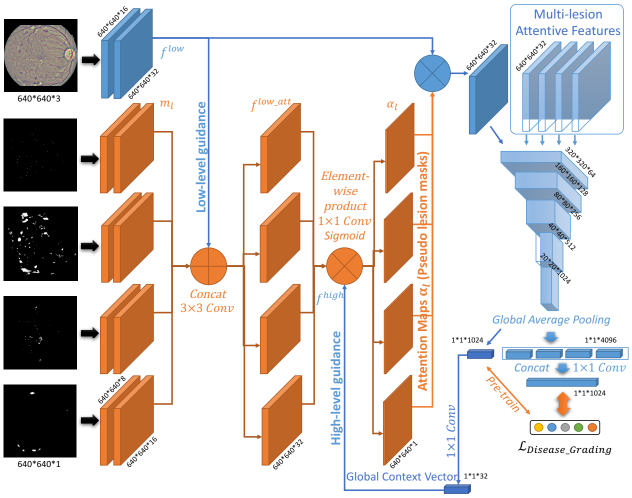

Figure 3. The details of the lesion attentive disease grading. The blue part is the classification model for disease grading and the orange part is the attention model for learning refined lesion maps.

The lesion attentive disease grading model, as shown in Fig. 3, is composed of a main branch for feature extraction and classification of the input disease images, and L branches for learning the attention models of the L lesions. We do not use the lesion masks initially predicted by the segmentation model to directly attend the classification model because the number of pixel-level annotated medical images is usually very small and thus the initially predicted masks are too weak to use. Moreover, the image-level grading labels can be exploited to deliver discriminative localization information to refine the lesion attention maps.

The disease grading model $C(\cdot)$ and lesion attention model $att(\cdot)$ in our method are tightly integrated. We first take a disease classification model with a basic convolutional neural network to learn grading using only input images. Once it is pre-trained, $\mathbf{f}^{low}$ and $\mathbf{f}^{high}$, which denote the low-level and high-level feature representations, respectively, can be extracted as pixel-wise and category-wise guidance for learning the attention model. Moreover, we also encode the initially predicted lesion maps, denoted by $\mathbf{m}_{l=1}^L$, as inputs to the attention model. The overall expression is defined by the following equation:

$$\alpha_{l=1}^L = att(\mathbf{f}^{low}, \mathbf{f}^{high}, \mathbf{m}_{l=1}^L), \qquad (4)$$

where the outputs $\alpha_{l=1}^L$ are the attention maps that give high responses to different lesion regions that characterize the disease. The proposed attention mechanism consists of two steps. The first step is to exploit pixel-wise lesion features by fusing the encoded low-level embeddings from both the input images and the initially predicted lesion masks. For the l-th lesion, we can obtain an intermediate state for an attentive feature by the equation:

$$\mathbf{f}_l^{low\_att} = \text{ReLU}(\mathbf{W}_l^{low}\text{concat}(\mathbf{m}_l, \mathbf{f}^{low}) + \mathbf{b}_l^{low}), \quad (5)$$

where concat$(\cdot)$ indicates the channel-wise concatenation. For the second step, we use a global context vector to correlate with the low-level attentive features and further generate the lesion maps as:

$$\alpha_l = \text{Sigmoid}(\mathbf{W}_l^{high}[\mathbf{f}_l^{low\_att} \odot \mathbf{f}^{high}] + \mathbf{b}_l^{high}), \quad (6)$$

where $\odot$ denotes an element-wise multiplication. The global context vector $\mathbf{f}^{high}$ has the same channel dimension as $\mathbf{f}_l^{low\_att}$, which is computed through a 1×1 convolution over the top layer feature from the basic pre-trained classification model. This high-level guidance contains abundant category information to weight low-level features and refine precise lesion details. Note that $\mathbf{W}_l^{low}$, $\mathbf{W}_l^{high}$ and bias terms are learnable parameters for the l-th lesion.

Based on the L lesion attention maps, we conduct an element-wise multiplication with the low-level image features f^{low} separately and use these attentive features to fine-tune the pre-trained disease classification model. All the lesion attentive features share the same weights as the grading model and the output feature vectors are concatenated for learning a final representation. The objective function $\mathcal{L}_{Cls}$ for disease grading adopts the focal loss [24] due to the imbalanced data problem. Meanwhile, the refined multi-lesion attention maps are used as pseudo masks to co-train the segmentation model in a semi-supervised manner.

3.4. Implementation Details

The training scheme for our model consists of two stages. In the first step, we pre-train the multi-lesion segmentation model using the pixel-level annotated data by $\mathcal{L}_{CE}$, and the basic disease severity classification model using the image-level annotated data by $\mathcal{L}_{Cls}$. Both are trained in a fully-supervised manner. The ADAM optimizer is adopted with the learning rate of 0.0002 and momentum of 0.5. The mini-batch size is set to 32 for pre-training the segmentation model over 60 epochs, while the grading model is pre-trained over 30 epochs with batch size of 128.

Once the pre-training is complete, the initially predicted lesion masks, along with the low-level and high-level feature representations of the input images, can be obtained to simultaneously train the lesion attention model for semi-supervised segmentation and further improve the grading performance. In this stage, we add the $\mathcal{L}_{Adv}$ for semi-supervised learning and the lesion attention module for disease grading. The whole model is fine-tuned in an end-to-end manner. λ in Eq. 3 is set to 10, which yields the best performance. The batch size is set to 16 for fine-tuning over 50 epochs. All experiments are run on an Nvidia DGX-1.

4. Experimental Results

4.1. Datasets and Evaluation Metrics

IDRID Dataset [32] is the only DR dataset providing pixel-level multi-lesion annotations, to the best of our knowledge. It contains 81 color fundus images with symptoms of DR and is split into 54 images for training and 27 images for testing. The lesions, including microaneurysms, haemorrhages, hard exudates and soft exudates are annotated by medical experts with binary masks. IDRID also has an image-level annotated set containing 413 training images and 103 testing images, which only have severity grading labels. We use the lesion segmentation set to train the multi-lesion mask generator in a fully-supervised manner. Then, the grading set is used to learn the lesion attentive model for classification and semi-supervised segmentation. **EyePACS Dataset** [2] consists of 35,126 training images and 53,576 testing images. The grading protocol is the same as the IDRID dataset, with five DR categories. However, the images collected from this dataset are captured by different types of cameras, under various light conditions and weak annotation quality. Since the dataset only has image-level grading labels, we mainly adopt it to train the lesion attentive disease grading model. **Messidor Dataset** [14] contains 1200 eye fundus images but its grading scale is different from that of the previous two datasets, having only 4 levels. Grades 0 and 1 are marked as non-referable, while Grades 2 and 3 are considered referable. All grades other than Grade 0 indicate an abnormal case of DR. Following the evaluation protocol used in [46], we only adopt this dataset for testing the models trained on EyePACS.

Data Pre-Processing and Augmentation. Since the fundus images from different datasets have various illuminations and resolutions, we proposed a data pre-processing method (clarified in the supplementary file) based on [43] to unify the image quality and sharpen the texture details. Moreover, to augment the data, horizontal flips, vertical flips and rotations are conducted, which can also mitigate the imbalance of samples across different classes.

Evaluation Metrics. To quantitively evaluate the performance of the lesion segmentation task, we compute the area under curve (AUC) value for both the receiving operating characteristic (ROC) curve and precision and recall (PR) curve. Moreover, to evaluate the precision of the DR grading model, in addition to the normal classification accuracy, a quadratic weighted kappa metric [2] is introduced.

4.2. Ablation Studies

4.2.1 Qualitative Multi-lesion Segmentation Results

Before evaluating the quantitative lesion segmentation precision and DR grading accuracy, we first qualitatively demonstrate the effectiveness of the lesion attention model for semi-supervised segmentation on the IDRID dataset,

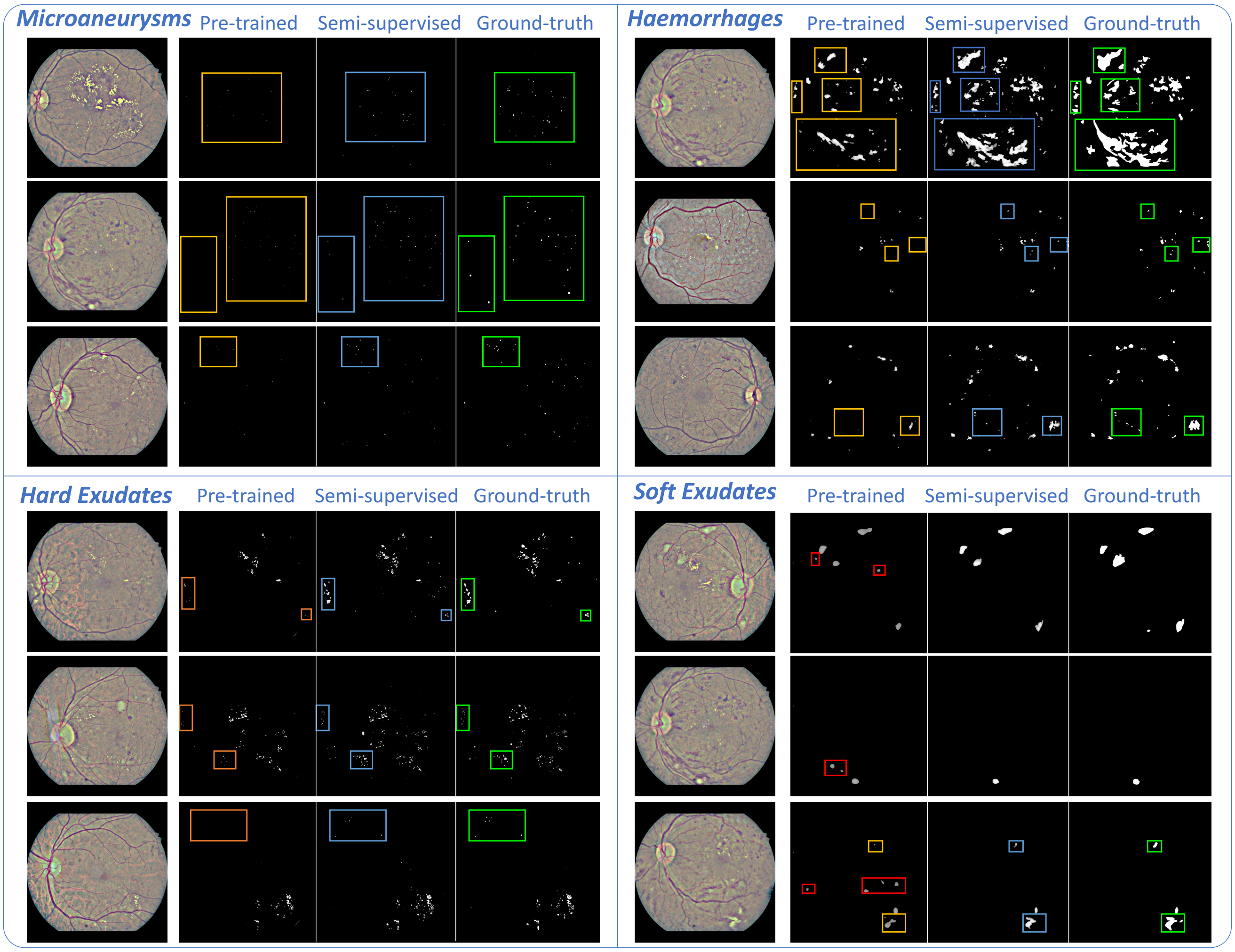

Figure 4. Qualitative multi-lesion segmentation results. We coarsely mark some regions to compare the initial model pre-trained on the limited data with pixel-level lesion annotations and the semi-supervised model trained using large-scale image-level annotated data. The green boxes denote the ground-truth. The blue boxes show the performance of our semi-supervised method, while the yellow and red boxes highlight the miss detections and false alarms, respectively. (Best viewed zoomed in.)

which has the segmentation ground-truth. Fig. 4 compares the segmentation results of four different lesions for the pre-trained model adopting only the limited pixel-level annotated data and the final model semi-supervised trained with large-scale image-level annotated data. For the pre-trained model, the failure case is usually the miss detection of the lesion patterns (false negative). False alarms (false positive) also occur in some small regions. With the help of image-level annotated data for semi-supervised segmentation, the results are obviously improved over all lesions. The effectiveness of the lesion segmentation for improving DR grading is evaluated by the ablation study in Sec. 4.2.2.

4.2.2 Effect of Lesion Attentive Disease Grading

To evaluate the effectiveness of lesion segmentation for DR grading and the improvement for semi-supervised learning by the attention model, we compare three baselines with our final proposed model. **Ori:** We first study if the lesion segmentation model can enhance the DR grading accuracy. In this baseline, we do not use the lesion attentive features but directly train the grading model on the pre-processed fundus images. **Lesion (Pretrained):** A baseline model pre-trained only on the limited pixel-level lesion annotated data is tested as well. The initially generated multi-lesion masks are weighted on image feature maps to train the grading model without the lesion attention model. **Lesion (Semi):** We also explore the improvement of semi-supervised learning by the lesion-attention model, using large-scale image-level grading annotated data. In this baseline, we only adopt the cross-entropy loss for learning the lesion segmentation model. **Lesion (Semi + Adv):** The adversarial training architecture is integrated into the lesion segmentation objective function as our final method.

Table 1. Evaluation of the effectiveness of the lesion attentive disease grading on the IDRID and EyePACS dataset.

Datasets	IDRID		EyePACS	
Methods	Acc.	Kappa	Acc.	Kappa
Ori	0.8458	0.7926	0.8541	0.8351
Lesion(Pretrained)	0.8725	0.8306	0.8598	0.8445
Lesion(Semi)	0.9016	0.8892	0.8792	0.8617
Lesion(Semi+Adv)	**0.9134**	**0.9047**	**0.8912**	**0.8720**

Table 1 shows the classification accuracy and kappa score of different methods. On the IDRID dataset, compared with the basic classification model that doesn't use the lesion mask information, the initial segmentation model pre-trained on the pixel-level annotated data can increase the accuracy of grading by 2.67% and the kappa score by 3.8%. With the semi-supervised learning using the image-level annotated data, an even more significant improvement can be achieved. In particular, the huge gain in the kappa score of 5.86% proves the proposed lesion attention model can effectively refine the lesion maps and thus improve the grading results. Moreover, the adversarial training architecture can also benefit the final result with a further gain of 1.18% for classification accuracy and 1.55% for kappa score. Since the EyePACS dataset only has image-level annotations, we adopt the fully-supervised model pre-trained on the IDRID dataset. A similar comparison can be made for the performance results obtained on the EyePACS dataset and those as those produced on the IDRID dataset, where each component of our model has a positive contribution to the grading task, compared to the other methods.

Table 2. Performance comparisons of two binary classification tasks on the Messidor dataset.

Settings	Referral		Normal	
Methods	AUC	Acc.	AUC	Acc.
Ori	0.934	0.902	0.889	0.878
Lesion(Pretrained)	0.953	0.909	0.919	0.901
Lesion(Semi)	0.971	0.930	0.937	0.918
Lesion(Semi+Adv)	**0.976**	**0.939**	**0.943**	**0.922**

To further evaluate our model, we also conduct experiments on the Messidor dataset. Following the evaluation method and protocol in [46], the AUC of ROC and the accuracy for normal and referral classification are compared in Table 2. For both experimental settings, the proposed method with the lesion attentive model, semi-supervised segmentation and adversarial training architecture achieves the highest performance. Since the image quality of the Messidor dataset is close to that of IDRID, even the pre-trained lesion based model can obtain a substantial gain compared with the basic holistic classification model.

4.2.3 Effect of Semi-Supervised Lesion Segmentation

In addition to the improvement of disease grading performance, we also investigate the effectiveness of semi-supervised segmentation based on the lesion pseudo masks by the lesion attention model. We evaluate the segmentation performance on the IDRID dataset with the pixel-level ground-truths. Four different lesions, including microaneurysms, haemorrhages, hard exudates and soft exudates, which are the main signs of DR, are assessed by the ROC, PR curves and the corresponding AUC values. We explore each proposed component of the final model with three baselines: the pre-trained segmentation model using the normal convolution tuple, the Xception-module based model and the semi-supervised learning component without an adversarial training architecture.

The ROC and PR curves are illustrated in Fig. 5 and detailed AUC values are listed in Table 3. As shown in the upper part of the table, the Xception-module based lesion segmentation model consistently outperforms the normal convolution-based version, over four different lesions. The AUC of the ROC and PR curves increases on average by 1.02% and 1.92%, respectively, proving that separable spatial and channel-wise convolution can indeed benefit the segmentation results. With the lesion attention model design, which exploits more image-level annotated data to generate pseudo masks for semi-supervised segmentation, a clear improvement is observed, with an average gain of 2.16% for the AUC of the PR curve. Besides, the adversarial training architecture for semi-supervised learning can slightly further increase the segmentation precision.

The bottom part of Table 3 shows the overall top three places with AUC scores for the PR curves of different lesions in the challenge [1], as well as the performance of the two semi-supervised segmentation methods AdvSeg [22] and ASDNet [30], transferred from other vision tasks. Although our method shows a slightly lower (0.57%) performance than the current top model for microaneurysms detection, moderate improvements are obtained for the other three lesions. A particularly large improvement of 4.12% is achieved for the soft exudate lesion. Moreover, our model outperforms the AdvSeg and ASDNet by an average increase of 6.89% and 5.22% on AUC of PR, respectively.

4.3. Comparisons with State-of-the-art Models

To make our method more convincing, we compare it with state-of-the-art DR grading models. The combined kernels with multiple losses network (CKML) [44] and VGGNet with extra kernels (VNXK) [44] aims to adopt multiple filter sizes to learn fine-grained discriminant features. Zoom-in-Net [46] was proposed with a gated attention model and combines three sub-networks to classify the holistic image, high-resolution crops and gated regions. The attention fusion network (AFN) [25] has a similar idea of

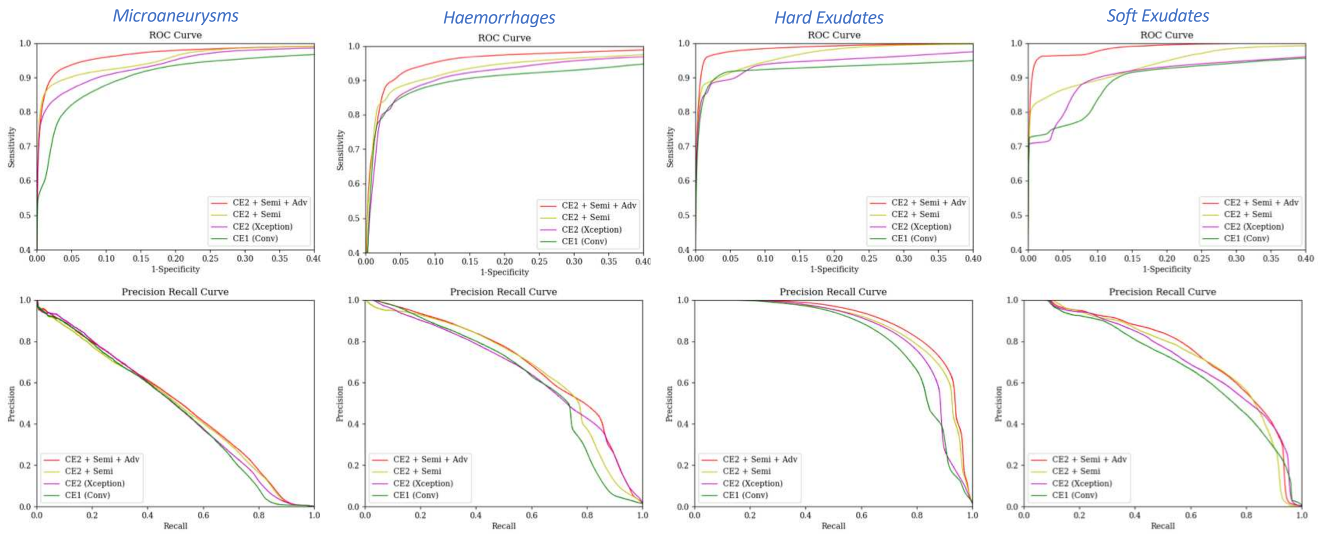

Figure 5. ROC and PR curves for segmentation over four lesions of DR. Four methods are compared to explore the effectiveness of the Xception-module based architecture, the lesion attentive model for semi-supervised segmentation and the adversarial training loss.

Table 3. Performance comparisons of multi-lesion segmentation on the IDRID dataset. CE1 and CE2 denotes the segmentation model adopting the normal convolution and the Xception module, respectively.

Lesion	Microaneurysms		Haemorrhages		Hard Exudates		Soft Exudates	
Methods	AUC_ROC	AUC_PR	AUC_ROC	AUC_PR	AUC_ROC	AUC_PR	AUC_ROC	AUC_PR
CE1(Conv)	0.9503	0.4625	0.9438	0.6456	0.9615	0.8263	0.9443	0.6817
CE2(Xception)	0.9653	0.4733	0.9540	0.6579	0.9675	0.8455	0.9537	0.7161
CE2+Semi	0.9776	0.4886	0.9699	0.6812	0.9886	0.8757	0.9713	0.7337
CE2+Semi+Adv	**0.9828**	0.4960	**0.9779**	**0.6936**	**0.9935**	**0.8872**	**0.9936**	**0.7407**
VRT	-	0.4951 (2)	-	0.6804 (1)	-	0.7127 (11)	-	0.6995 (1)
PATech	-	0.474 (3)	-	0.649 (2)	-	0.885 (1)	-	-
iFLYTEK-MIG	-	**0.5017** (1)	-	0.5588 (3)	-	0.8741 (2)	-	0.6588 (3)
AdvSeg [22]	0.9612	0.4706	0.9256	0.5923	0.9456	0.8032	0.9318	0.6756
ASDNet [30]	0.9692	0.4782	0.9324	0.6285	0.9502	0.8095	0.9489	0.6924

Table 4. Performance comparisons of DR grading on the EyePACS and Messidor datasets.

EyePACS		Messidor				
Test set		Settings	Referral		Normal	
Methods	Kappa	Methods	AUC	Acc.	AUC	Acc.
Min-Pooling	0.849	VNXK [44]	0.887	0.893	0.870	0.871
o_O	0.845	CKML [44]	0.891	0.897	0.862	0.858
RG	0.839	Expert [37]	0.94	-	0.922	-
Zoom-in [46]	0.854	Zoom-in [46]	0.957	0.911	0.921	0.905
AFN [25]	0.859	AFN [25]	0.968	-	0.935	-
Ours	**0.872**	Ours	**0.976**	**0.939**	**0.943**	**0.922**

unifying lesion detection and DR grading. However, the attention model used is only class-driven and cannot learn precise semantic lesion maps. Moreover, human experts [37] are also invited to grade on the Messidor dataset.

Table 4 compares the results of different methods. On the EyePACS dataset, Kappa values of the top three places from the Kaggle competition [2] are shown where the top-1 place can achieve 84.9%. The Zoom-in-Net and AFN slightly improve the performance by introducing attention mechanisms for learning class-driven lesion maps. Our method proposes to collaborate the semantic lesion mask guidance and the class-driven attention guidance to enhance the final model which obtains 1.3% gain over AFN. Moreover, for both the referable/non-referable and normal/abnormal settings of Messidor, our method can obtain the highest AUC scores of ROC and also grading accuracy, compared to other approaches. It is worth mentioning that our method outperforms the human experts by 3.6% and 2.1% on the AUC of referral and normal settings, respectively.

5. Conclusion

In this paper, we proposed a collaborative learning method of semi-supervised lesion segmentation and disease grading for medical imaging. Lesion masks were used to attend the classification model and improve the grading accuracy, while a lesion attentive model exploiting class-specific labels also benefited the segmentation results. Extensive experiments showed that our method achieves improvements on the DR problem.

References

[1] Idrid diabetic retinopathy segmentation challenge. https://idrid.grand-challenge.org/. 7

[2] Kaggle diabetic retinopathy detection competition. https://www.kaggle.com/c/diabetic-retinopathy-detection. 5, 8

[3] International clinical diabetic retinopathy disease severity scale. *American Academy of Ophthalmology*, 2012. 1

[4] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, June 2018. 2

[5] B. Antal, A. Hajdu, et al. An ensemble-based system for microaneurysm detection and diabetic retinopathy grading. *IEEE transactions on biomedical engineering*, 59(6):1720, 2012. 2

[6] A. M. Boers, R. S. Barros, I. G. Jansen, C. H. Slump, D. W. Dippel, A. van der Lugt, W. H. van Zwam, Y. B. Roos, R. J. van Oostenbrugge, C. B. Majoie, et al. Quantitative collateral grading on ct angiography in patients with acute ischemic stroke. In *MICCAI*, pages 176–184. Springer, 2017. 1

[7] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, July 2017. 2

[8] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *CVPR*, June 2018. 1

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2018. 1

[10] X. Chen, J. Hao Liew, W. Xiong, C.-K. Chui, and S.-H. Ong. Focus, segment and erase: An efficient network for multi-label brain tumor segmentation. In *ECCV*, September 2018. 1

[11] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint*, pages 1610–02357, 2017. 3

[12] A. V. Dalca, J. Guttag, and M. R. Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *CVPR*, June 2018. 1

[13] T. de Moor, A. Rodriguez-Ruiz, R. Mann, and J. Teuwen. Automated soft tissue lesion detection and segmentation in digital mammography using a u-net deep learning network. *arXiv preprint arXiv:1802.06865*, 2018. 2

[14] 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 Analysis & Stereology*, 33(3):231–234, 2014. 5

[15] D. Doshi, A. Shenoy, D. Sidhpura, and P. Gharpure. Diabetic retinopathy detection using deep convolutional neural networks. In *Computing, Analytics and Security Trends (CAST), International Conference on*, pages 261–266. IEEE, 2016. 2

[16] R. Fan, Q. Hou, M.-M. Cheng, G. Yu, R. R. Martin, and S.-M. Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In *ECCV*, September 2018. 1

[17] P. Filipczuk, M. Kowal, and A. Marciniak. Feature selection for breast cancer malignancy classification problem. *Journal of Medical Informatics & Technologies*, 15:193–199, 2010. 2

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in NIPS*, pages 2672–2680, 2014. 4

[19] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016. 1, 2

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988. IEEE, 2017. 1

[21] S. Hong, H. Noh, and B. Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *NIPS*, pages 1495–1503, 2015. 2

[22] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018. 2, 7, 8

[23] Q. Li, A. Arnab, and P. H. Torr. Weakly- and semi-supervised panoptic segmentation. In *ECCV*, September 2018. 1, 2

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *TPAMI*, 2018. 5

[25] Z. Lin, R. Guo, Y. Wang, B. Wu, T. Chen, W. Wang, D. Z. Chen, and J. Wu. A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion. In *MICCAI*, pages 74–82. Springer, 2018. 7, 8

[26] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1

[27] X. Long, C. Gan, G. de Melo, J. Wu, X. Liu, and S. Wen. Attention clusters: Purely attention based local feature integration for video classification. In *CVPR*, June 2018. 2

[28] E. Miranda, M. Aryuni, and E. Irwansyah. A survey of medical image classification techniques. In *Information Management and Technology (ICIMTech), International Conference on*, pages 56–61. IEEE, 2016. 1

[29] T. Nair, D. Precup, D. L. Arnold, and T. Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. In *MICCAI*, pages 655–663. Springer, 2018. 1

[30] D. Nie, Y. Gao, L. Wang, and D. Shen. Asdnet: Attention based semi-supervised deep networks for medical image segmentation. In *MICCAI*, pages 370–378. Springer, 2018. 7, 8

[31] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, December 2015. 2

[32] 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*, 3(3):25, 2018. 5

[33] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng. Convolutional neural networks for diabetic retinopathy. *Procedia Computer Science*, 90:200–205, 2016. 2

[34] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille. Deep co-training for semi-supervised image recognition. In *ECCV*, September 2018. 1

[35] T. Robert, N. Thome, and M. Cord. Hybridnet: Classification and reconstruction cooperation for semi-supervised learning. In *ECCV*, September 2018. 1

[36] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2, 3

[37] C. I. Sánchez, M. Niemeijer, A. V. Dumitrescu, M. S. Suttorp-Schulten, M. D. Abramoff, and B. van Ginneken. Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data. *Investigative ophthalmology & visual science*, 52(7):4866–4871, 2011. 8

[38] N. Sarafianos, X. Xu, and I. A. Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *ECCV*, September 2018. 2

[39] F. Sener and A. Yao. Unsupervised learning and segmentation of complex activities from video. In *CVPR*, June 2018. 1

[40] L. Seoud, J. Chelbi, and F. Cheriet. Automatic grading of diabetic retinopathy on a public database. In *MICCAI*. Springer, 2015. 1

[41] L. Seoud, T. Hurtut, J. Chelbi, F. Cheriet, and J. P. Langlois. Red lesion detection using dynamic shape features for diabetic retinopathy screening. *IEEE transactions on medical imaging*, 35(4):1116–1126, 2016. 2

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 3

[43] M. J. van Grinsven, B. van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez. Fast convolutional neural network training using selective data sampling: application to hemorrhage detection in color fundus images. *IEEE transactions on medical imaging*, 35(5):1273–1284, 2016. 5

[44] H. H. Vo and A. Verma. New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space. In *Multimedia (ISM), 2016 IEEE International Symposium on*, pages 209–215. IEEE, 2016. 7, 8

[45] X. Wang, S. You, X. Li, and H. Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, June 2018. 1

[46] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection. In *MICCAI*, pages 267–275. Springer, 2017. 2, 5, 7, 8

[47] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang. Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *CVPR*, June 2018. 2

[48] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, June 2018. 2

[49] K. Yan, X. Wang, L. Lu, and R. M. Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(3):036501, 2018. 2

[50] Y. Yang, T. Li, W. Li, H. Wu, W. Fan, and W. Zhang. Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. In *MICCAI*, pages 533–540. Springer, 2017. 1, 2

[51] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Learning a discriminative feature network for semantic segmentation. In *CVPR*, June 2018. 1

[52] Y. Zhou, L. Liu, and L. Shao. Vehicle re-identification by deep hidden multi-view inference. *IEEE Transactions on Image Processing*, 27(7):3275–3287, 2018. 2

[53] Y. Zhou and L. Shao. Viewpoint-aware attentive multi-view inference for vehicle re-identification. In *CVPR*, June 2018. 2

[54] C. Zhu, X. Tan, F. Zhou, X. Liu, K. Yue, E. Ding, and Y. Ma. Fine-grained video categorization with redundancy reduction attention. In *ECCV*, September 2018. 2

[55] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, September 2018. 1