Learn to Segment Retinal Lesions and Beyond

Qijie Wei*†, Xirong Li*†, Weihong Yu‡, Xiao Zhang‡, Yongpeng Zhang§, Bojie Hu¶, Bin Mo§
Di Gong‖, Ning Chen**, Dayong Ding†, Youxin Chen‡
*Key Lab of DEKE, Renmin University of China, Beijing, China
†Vistel AI Lab, Visionary Intelligence Ltd, Beijing, China
‡Peking Union Medical College Hospital, Beijing, China
§Beijing Tongren Hospital, Beijing, China
¶Tianjin Medicial University Eye Hospital, Tianjin, China
‖China-Japanese Riendship Hospital, Beijing, China
**The Affilliated Yantai Yuhuangding Hospital of Qingdao University, Yantai, China
Email: qijie.wei@vistel.cn, xirong@ruc.edu.cn, chenyouxinpumch@163.com

***Abstract*—Towards automated retinal screening, this paper makes an endeavor to simultaneously achieve pixel-level retinal lesion segmentation and image-level disease classification. Such a multi-task approach is crucial for accurate and clinically interpretable disease diagnosis. Prior art is insufficient due to three challenges, *i.e.*, lesions lacking objective boundaries, clinical importance of lesions irrelevant to their size, and the lack of one-to-one correspondence between lesion and disease classes. This paper attacks the three challenges in the context of diabetic retinopathy (DR) grading. We propose *Lesion-Net*, a new variant of fully convolutional networks, with its expansive path re-designed to tackle the first challenge. A dual Dice loss that leverages both semantic segmentation and image classification losses is introduced to resolve the second challenge. Lastly, we build a multi-task network that employs Lesion-Net as a side-attention branch for both DR grading and result interpretation. A set of 12K fundus images is manually segmented by 45 ophthalmologists for 8 DR-related lesions, resulting in 290K manual segments in total. Extensive experiments on this large-scale dataset show that our proposed approach surpasses the prior art for multiple tasks including lesion segmentation, lesion classification and DR grading.**

I. Introduction

Given the increasing demand of retinal screening and the clear shortage of experienced ophthalmologists, fundus image based retinal disease diagnosis is crucial for the well-being of many [1]–[3]. Previous studies on fundus image segmentation concentrate on anatomical structures in retina including optic disc / cup and vessels [4]–[6]. By contrast, this paper aims for *retinal lesions*, which are symptoms of ocular fundus diseases manifested in color fundus images. By answering the question of what lesions are in a fundus image and where in the image they are located, lesion segmentation has a potential to enable clinical interpretability of disease classes predicted at the image level. Attacking lesion segmentation and retinal disease classification in a unified framework is thus valuable.

Note that for natural images as in PASCAL-VOC alike tasks [7]–[9], the semantic segmentation task and the image classification task typically share the same class vocabulary. Consequently, developing a multi-task approach seems to be relatively straightforward, *e.g.*, by converting classes predicted at the pixel-level to the image-level by max or mean pooling.

TABLE I
A summary of the American Academy of Ophthalmology (AAO) preferred practice pattern guidelines for diabetic retinopathy grading. As venous beading and IrMA are very difficult to be recognized even for ophthalmologists and occur rarely, we exclude them from this study. The eight lesions stuided in this work are indicated by ✓.

Grade	Lesion evidence for DR grading	
	Sufficient	*Indirect*
DR1	• Microaneurysm (MA), exclusively ✓	–
DR2	• Intraretinal hemorrhage (iHE) ✓	• Hard exudate (HaEx) ✓
DR3	Any of the following: • Over 20 iHEs in each of 4 quadrants • Venous beading in 2+ quadrants • IrMA in 1+ quadrants	• Cotton-wool spot (CWS) ✓
DR4	Any of the following: • Neovascularization (NV) ✓ • Vitreous hemorrhage (vHE) ✓ • Preretinal hemorrhage (pHE) ✓	• Fibrous proliferation (FiP) ✓

For fundus images, however, lesion labels and disease classes are distinct and lack one-to-one correspondence. See for instance lesions used in the clinical practice guidelines for diabetic retinopathy[1] (DR) grading in Table I. This means lesion segmentation cannot be directly converted to image-level DR grades. Hence, a unified framework that effectively segments lesions and exploits the segmentation for accurate disease classification is in demand.

Given a fundus image, instances of a specific lesion class occupy a specific region or multiple regions with diverse visual appearance, see Fig. 1. With the advent of fully convolutional networks (FCN) [11], exciting progress has been made in semantic segmentation, especially for natural scenes [12]–[16]. However, directly applying the state-of-the-art for retinal lesion segmentation is problematic. Unlike objects in natural images, retinal lesions lack clear boundaries against the background. It is practically impossible for ophthalmologists to segment lesions at the same preciseness, meaning an FCN has to learn from annotations with imprecise boundaries. In the meanwhile, for diagnosis, it is mostly the presence and locality of specific lesions that are involved, see Table I. Extremely precise segmentation is not only difficult to achieve but also

[1]Diabetic retinopathy is a complication of diabetes mellitus caused by damage to blood vessels of the light-sensitive tissue at the retina [10].

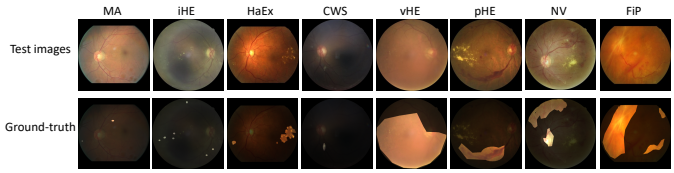

Fig. 1. **Visual examples of 8 DR-related retinal lesions studied in this paper**. For a clear view, we show only one lesion per image.

unnecessary from a clinical view. We thus hypothesize that cutting-edge FCNs, *e.g.*, DeepLabv3+ [14], are over-designed for lesion segmentation.

Moreover, while the importance of an object in a natural image is largely reflected by its size [17], the importance of a lesion in a fundus image does not count on the amount of pixels it possesses. Diabetic retinopathy, depending on what lesions are presented, is categorized into five levels, from DR0 (*i.e.*, no DR) to DR4 (*i.e.*, proliferative DR). The presence of a preretinal hemorrhage, even though in a relatively small size, means DR4. Such a property cannot be well addressed by current segmentation losses including cross entropy [11], [18], [19], focal loss [20], [21], and Dice [4], [22].

To conquer the aforementioned challenges, this paper makes the first endeavor to simultaneously solve retinal lesion segmentation and disease classification in an end-to-end framework. We choose DR, a leading cause of blindness [10], as our target disease. Our main novelties are

- We study eight lesions including microaneurysm (MA), intraretinal hemorrhage (iHE), hard exudate (HaEx), cotton-wool spot (CWS), vitreous hemorrhage (vHE), preretinal hemorrhage (pHE), neovascularization (NV), and fibrous proliferation (FiP) that support the full range of DR grades. This is a new state-of-the-art in terms of quantity, complexity and clinical usability.
- We propose Lesion-Net for retinal lesion segmentation. While inheriting FCN's classical contracting-and-expansive structure, Lesion-Net has a re-designed expansive path with its length adjustable and its upsampling operation lightweight trainable. We adopt a *dual loss* that combines both semantic segmentation and image classification losses. These two designs enable Lesion-Net to effectively learn from lesion annotations with imprecise boundaries and to substantially reduce false alarms of small-size lesions.
- We propose a multi-task network that effectively harnesses lesion segmentation maps, as side information, for improving DR grading. Such an attention mechanism conceptually differs from prevalent self-attention mechanisms [15], [23], [24]. Once trained, the multi-task network performs three tasks, *i.e.*, lesion segmentation, lesion classification and DR grading, all in one forward pass.
- We conduct extensive experiments on 12K color fundus images collected from Kaggle [25] and local hospitals. With 290K expert-labeled pixel-level lesion segments, the dataset is the largest of its kind. The experiments confirm the superiority of both Lesion-Net and the multi-task network against the prior art including FCN [11], U-Net [18], DANet [15], and DeepLabv3+ [14] for lesion segmentation and classification, Inception-v3 [26] and ABN [24] for DR grading. To promote related research, we have released the Kaggle part of our test data, containing 1,593 images and 34,268 expert-labeled lesion segments[2], substantially larger than the present-day dataset that has only 81 images with manual segmentation for four lesions [27].

II. Related Work

Models for semantic segmentation. Since Long *et al.* [11], FCNs have been the *de facto* standard technique for semantic segmentation. An FCN can be conceptually decomposed into a contracting path and an expansive path. The contracting path progressively extracts and downsamples feature maps from an input image. The expansive path, by transforming and upsampling, produces a full-resolution segmentation map of the same size as the input image. Towards more precise segmentation, novel designs are continuously proposed either in the contracting path, or in the expansive path or in both. For instance, dilated convolutions are introduced in [12], so the contracting path can produce feature maps with higher resolutions to preserve more detailed spatial information. In U-Net [18], the contracting path and the expansive path are carefully designed to be symmetrical. Skip connections from the contracting path to the expansive path are added, again for the purpose of preserving spatial information to generate more accurate segmentation boundaries. In order to capture long-range contextual information in both spatial and channel dimensions, DANet [15] introduces a position attention module and a channel attention module in the expansive path. The state-of-the-art DeepLabv3+ uses both dilated convolutions and spatial pyramid pooling in its contracting path [14]. Its expansive path uses multiple skip connections to exploit features from lower levels. As identifying the precise boundary of a retinal lesion is secondary to the practical use of lesion segmentation, a new FCN is required.

Retinal lesion segmentation. While earlier works for retinal lesion segmentation use traditional image processing techniques [28], [29], current works mostly take a patch-based deep learning approach [30]–[33]. In [30], for instance, a customized CNN is used to segment iHE by patch classification. Similarly in [32], a patch-trained CNN is applied in a sliding window manner, classifying every grid in a test image into five classes, *i.e.*, normal, MA, iHE, HaEx and high-risk lesion. By predicting whether a given patch contains a specific lesion, segmentation maps obtained by the above works tend to be sparse and imprecise. A more fundamental drawback is that the approach lacks a holistic view. Consider MA and iHE for instance. The two lesions are visually close as both are small lesions look like dark dots. However, MA occurs around vessels. Also, an image with no other lesion is more likely to have MA than iHE. For a model looking only at local areas, modeling these kinds of holistic clues is difficult.

Lesion-enhanced DR grading. Initial efforts have been made towards lesion-enhanced DR grading. A two-step

[2]https://github.com/WeiQijie/retinal-lesions

method is developed in [34], where an input image is first converted to a weight map by using a CNN to classify all patches of the image as normal, MA or iHE. The image, multiplied by the weight map, is fed into a DR grading network. A lesion-guided attention mechanism is described in [35] to weigh specific regions in the input image. Three lesions are considered: MA, iHE and HaEx. Neither of these works considers severe lesions such as pHE, vHE, and NV.

Attention-enhanced image classification. The state-of-the-art is Attention Branch Network (ABN) [24], which extends a response-based visual explanation model [36] by introducing an attention branch into a specific CNN. Consequently, ABN not only improves image classification but also produces an attention map to interpret the decision. Note that the attention is self-generated. Our attention mechanism exploits the output of the semantic segmentation network as side information, and thus conceptually differs from ABN.

III. Approach

Given a color fundus image, we aim to perform lesion segmentation, classification and subsequently DR grading in a unified framework. We use $\mathcal{X}$ to denote a specific $s \times s$ image, which contains an array of s^2 pixels $\{x_1, \ldots, x_i, \ldots, x_{s^2}\}$. Let $\mathcal{L} = \{l_1, \ldots, l_m\}$ be m lesions in consideration. Regions of distinct lesions, *e.g.*, HaEx and iHE, often overlap partially, meaning a pixel can be assigned with multiple labels. So the goal of lesion segmentation is to automatically assign to each pixel x_i a m-dimensional probabilistic vector, $\mathbf{p}_i = \{p_{i,1}, \ldots, p_{i,m}\}$, where $p_{i,j} \in [0, 1]$ indicates the probability of the pixel belonging to the j-th lesion. Lesion classification is to predict lesions at the image level. Given the probabilistic segmentation map $\{\mathbf{p}_1, \ldots, \mathbf{p}_{s^2}\}$, the probability of the presence of a specific lesion l_j, denoted as P_j, is naturally obtained by global max pooling on the map, *i.e.*,

$$P_j := \max\{p_{1,j}, \ldots, p_{s^2,j}\}, \quad j = 1, \ldots, m. \tag{1}$$

For both lesion segmentation and classification, hard labels are obtained by thresholding at 0.5. As for DR grading, the goal is to exclusively assign one of the following labels, *i.e.*, {DR0, DR1, DR2, DR3, DR4}, to the given image.

Next, we depict the proposed lesion segmentation network, followed by the multi-task network.

A. Lesion-Net for Retinal Lesion Segmentation

Network architecture. For the contracting path of Lesion-Net, we use convolutional blocks of Inception-v3 [26] for its outstanding feature extraction ability. Note that other state-of-the-art CNNs [37]–[39] can, in principle, be used.

Our task-specific design lies in the expansive path, where we leverage the effectiveness of U-Net [18] for re-using information from the contracting path and the flexibility of the original FCN [11] in cutting off the expansive path for preventing over-precise segmentation.

In concrete, in order to re-use feature maps from the contracting path, we adopt U-Net's copy-and-merge strategy instead of adding operations in the FCN, see Fig. 2. For upsampling, we replace U-Net's deconvolution by a 1×1 convolution to adjust the number of feature maps and subsequently a parameter-free bilinear interpolation to enlarge the feature maps. Such a tactic not only reduces the number of parameters. By applying an element-wise sigmoid activation, the output of the 1×1 convolution is naturally transformed to m probabilistic maps with respect to the m lesions.

The fact that retinal lesions lack accurate boundaries makes it unnecessary to seek for very precise segmentation. While the symmetry between the contracting and expansive paths in U-Net is useful in its original context of cell segmentation, we argue that such a constraint is unnecessary for the current task. In fact, extra parameters introduced by the symmetry into the expansive path increases the difficulty of training the network. Therefore, we let the length of Lesion-Net's expansive path adjustable. If the expansive path is cut at an early stage with feature maps of size 28×28, the maps need to be upsampled by a factor of 32 to produce the final segmentation maps. Following the convention of [11], we term this variant Lesion-Net-32s. By contrast, Lesion-Net-2s exploits all the intermediate feature maps. The models that fall in between are Lesion-Net-16s, Lesion-Net-8s and Lesion-Net-4s. Fig. 3 shows Lesion-Net with distinct expansive paths.

Loss function. Training Lesion-Net is nontrivial due to the following two issues. First, while the area of a specific lesion varies, the importance of the lesion does not depend on its size. This property cannot be well reflected in a pixel-wise loss, to which a smaller blob contributes less. Misclassifying a small blob does not lead to a significant increase in the segmentation loss, and thus difficult to be corrected during training. Such a small misclassification, even though ignorable from the viewpoint of semantic segmentation, can be crucial for proper diagnosis of related diseases. Second, the data is extremely imbalanced, making commonly used loss functions such as cross entropy ineffective. Our study on a set of 12k expert-labeled fundus images shows that pixels of lesions account for less than 1%. By contrast, for PASCAL VOC2012 [40], a popular benchmark set for natural image segmentation, the proportion of pixels corresponding to objects is about 25%. We find in preliminary experiments that with the cross-entropy loss, the lesion segmentation model easily got trapped in a local optimum, predicting all pixels as negative, albeit a very low training loss.

To jointly address the two issues, we introduce a new *dual loss* that combines a semantic segmentation loss $loss_{seg}$ and an image classification loss $loss_{clf}$, *i.e.*,

$$loss_{dual} = \lambda \cdot loss_{seg} + (1 - \lambda) \cdot loss_{clf}, \tag{2}$$

where $\lambda \in [0, 1]$ is a hyper parameter to strike a balance between the two sub-losses. We instantiate both $loss_{seg}$ and $loss_{clf}$ using the Dice loss, previously used for segmenting prostate MRI [22] and optic disc / cup [4]. Our ablation study in Section IV-B shows that Dice loss is more effective than Weighed Cross Entropy [41] and Focal Loss [20]. The weight λ is empirically set to 0.8 based on a held-out validation set, a common practice for selecting hyper parameters.

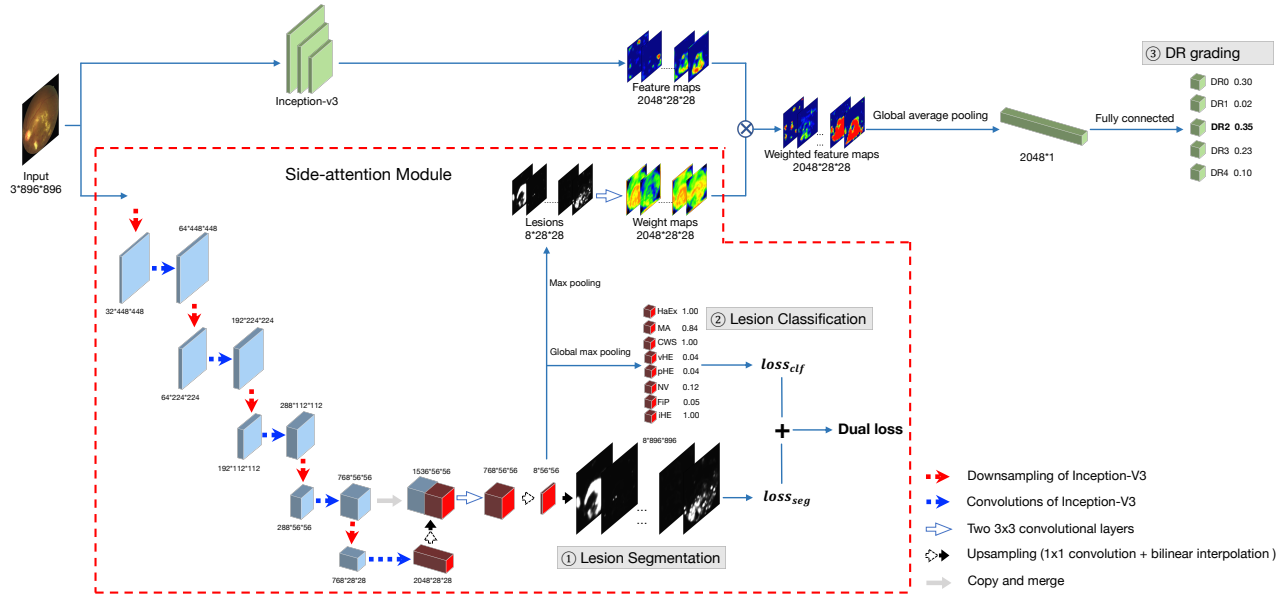

Fig. 2. **Proposed multi-task network for (1) lesion segmentation, (2) lesion classification and (3) DR grading**. Given a color fundus image, Lesion-Net-16s (the lower branch) generates probabilistic segmentation maps for eight lesions. Lesion classification is accomplished by global max pooling on the maps. For lesion-enhanced DR grading, a side-attention branch is used to fuse the segmentation maps with an array of 2,048 feature maps from Inception-v3 in the upper branch. Compared with directly weighing the feature maps with the segmentation maps, the trainable side-attention is more effective.

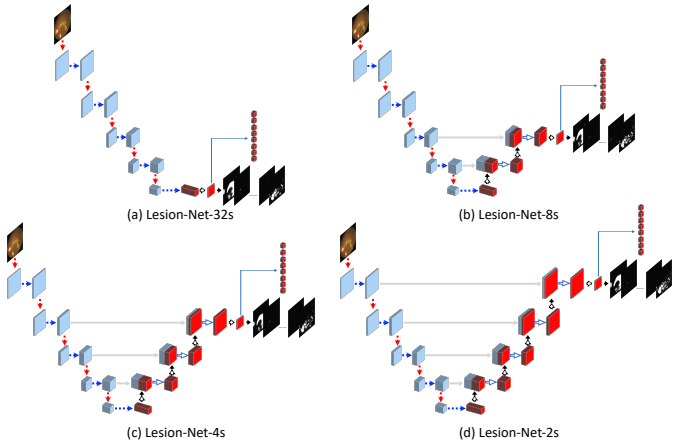

Fig. 3. **Variants of Lesion-Net**. The variable-length expansive path enables learning from lesion annotations with imprecise boundaries.

Note that the dual Dice loss is conceptually different from the multi-scale Dice [4], which combines pixel-level Dice losses computed from images of varied scales and thus remains insensitive to small-sized errors. The motivation of our dual loss also differs from the combined cross-entropy loss described in [42], where the image classification loss is used as a regularization term to reduce overfitting.

Given a mini-batch of n images, we compute the Dice version of $loss_{seg}$ as

$$loss_{seg} = 1 - \frac{2 \cdot \sum_{i=1}^{n \cdot s^2} \sum_{j=1}^{m} p_{i,j} \cdot t_{i,j}}{\sum_{i=1}^{n \cdot s^2} \sum_{j=1}^{m} p_{i,j}^2 + \sum_{i=1}^{n \cdot s^2} \sum_{j=1}^{m} t_{i,j}^2}, \quad (3)$$

where $t_{i,j} \in \{0, 1\}$ is ground truth of the i-th pixel with respect to the j-th lesion. In the extreme case where all pixels are predicted as negative, the dice loss is close to 1.

We compute the Dice version of $loss_{clf}$ as

$$loss_{clf} = 1 - \frac{2 \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} P_{i,j} \cdot T_{i,j}}{\sum_{i=1}^{n} \sum_{j=1}^{m} P_{i,j}^2 + \sum_{i=1}^{n} \sum_{j=1}^{m} T_{i,j}^2}, \quad (4)$$

where $T_{i,j} \in \{0, 1\}$ is the ground-truth label indicating whether the j-th lesion is present in the i-th image in the given batch. Recall that both $P_{i,j}$ and $T_{i,j}$ are obtained by global max pooling on the pixel-level labels, so $P_{i,j}$, $T_{i,j}$ and accordingly $loss_{clf}$ are all invariant to the lesion size.

B. Multi-task Network for Lesion-enhanced DR Grading

To predict DR grades, we choose Inception-v3 [26] as our baseline model. This model has established the state-of-the-art for predicting referable DR [1], age-related eye diseases [43] and other retinal abnormalities [2]. In fact, for all models evaluated in this work, we use Inception-v3 as their backbones for fair comparison. To produce a probabilistic score per DR grade, we modify Inception-v3 by adding after the global average pooling (GAP) layer a fully connected layer of size $2,048 \times 5$, followed by a softmax layer. Different from previous works that use a typical resolution of 299×299 [1], we use a much larger resolution of 896×896, making our Inception-v3 a much stronger baseline.

For lesion-enhanced DR grading, we propose a multi-task network, with its overall architecture shown in Fig. 2. The multi-task network consists of two branches. At the top is its main branch, with Inception-v3 as the backbone, that performs DR grading. The side-attention branch, with Lesion-Net as its backbone, is responsible for injecting semantic and spatial information contained in the m lesion segmentation maps into the main branch. In particular, the injection is performed at the last feature maps, denoted as $\{f_1, \ldots, f_k\}$, in the main branch, with $k = 2,048$. To that end, the side-attention branch

shall generate the same number of weight maps, denoted as $\{w_1, \ldots, w_k\}$. Multiplying the feature maps by the weight maps side by side generates new weighted feature maps as

$$\{f_1 \otimes w_1, f_2 \otimes w_2, \ldots, f_k \otimes w_k\}, \quad (5)$$

where $\otimes$ indicates element-wise multiplication. The new feature maps then go through a GAP layer, followed by a classification block. It is worth point out that the weight information is essentially from the side-attention branch rather than generated by the main branch itself. Hence, the multi-task network is conceptually different from self-attention networks [15], [23], [24].

To convert the lesion segmentation maps $\{s_1, \ldots, s_m\}$ into the weight maps, an intuitive strategy is to let the main branch pay attention to regions with maximal lesion response. This is achieved by channel-wise max pooling (CW-MaxPool) over the segmentation maps[3], *i.e.*

$$w_i := \text{CW-MaxPool}(\{s_1, \ldots, s_m\}), \quad i = 1, \ldots, k. \quad (6)$$

However, a region deemed to be negative with respect to the lesions does not necessarily mean it is useless for DR grading. So we further consider a learning based strategy, using a lightweight convolutional block consisting of two 3×3 convolutional layers, *i.e.*

$$\{w_1, \ldots, w_k\} := \text{Conv}(\{s_1, \ldots, s_m\}). \quad (7)$$

We train the multi-task network with the cross-entropy loss, commonly used for multi-class image classification.

IV. Evaluation

A. Experimental Setup

According to the AAO guidelines [44], there are seven lesions used as sufficient evidence for specific DR grades. Among them, venous beading and IrMA are very difficult to be recognized even for ophthalmologists and occur rarely. So they are excluded from this study. We include three other lesions, *i.e.*, HaEx, CWS, and FiP, indirectly related to DR grading. We compile a final list of eight lesions, see Table I.

Ground-truth construction. Public datasets suited for our purpose does not exist. So we construct a large collection of 12,252 color fundus images with both pixel-level lesion annotations and image-level DR grades as follows. We collected initially 23K color fundus images of posterior pole, consisting of 12k images from our hospital partners and 11k images randomly sampled from the Kaggle DR Detection task [25]. While the images were from patients with diabetes, some of them show other eye diseases such as glaucoma, AMD and RVO. So DR0 does not necessarily mean a healthy eye. Such a characteristic makes the data close to the real scenario and thus challenging.

For expert labeling, a panel of 45 experienced ophthalmologists was formed. We developed a web-based annotation system, where an annotator marks out lesions in a given image using either ellipses or polygons and accordingly grade the image. Lesion annotation and DR grading from a single image are somewhat subjective. So for quality control, each image was assigned to at least three annotators. Images receiving consistent DR grades, *i.e.*, the majority vote for a specific grade, are preserved. Accordingly, per image we cleaned lesion annotations so they are complied to the diagnostic guidelines. Eventually, we obtain 12,252 images with 290K expert-labeled lesion segments. We split the dataset at random for training (70%), validation (10%) and test (20%).

Implementations. An input image is sized to 896×896, as small lesions can not be seen well in lower resolution. We use SGD with a weight decay factor of 0.0001 and a momentum of 0.95. The initial learning rate is 0.001. Validation occurs every 1K batches. If the validation performance does not improve in 4 consecutive validations, the learning rate will be divided by 10. Early stop occurs once the performance does not improve in 10 consecutive validations. For training Lesion-Net, we start with $loss_{seg}$. Once the learning rate is reduced, $loss_{seg}$ is replaced by $loss_{dual}$. For DR grading, a pre-trained Lesion-Net is used for the multi-task network. We tried to train both branches, but found no improvement in DR grading yet an absolute decrease of 0.01 in the segmentation performance. So we did not go further in that direction. For the varied models assessed in this paper, we use Inception-v3 pre-trained on ImageNet [45] as their backbones. Random rotation, crop, flip and random changes in brightness, saturation and contrast are used for data augmentation. Training was performed using PyTorch on two NVIDA Tesla P40 GPUs.

Evaluation criteria. For lesion segmentation, we report pixel-wise F1 score, the harmonic mean of precision and recall. For lesion classification, we report image-wise F1. As lesions predicted at the pixel level are propagated to the image level via global max pooling, the two criteria complements each other, providing a more comprehensive assessment of a specific segmentation model. For DR grading, We report the quadratic weighted *kappa*, which measures inter-annotator agreement and used by the Kaggle DR Detection task [25].

B. Experiment 1. Lesion Segmentation

Baselines. Our criteria for choosing baselines are two-fold: state-of-the-art in related tasks and open-source, allowing us to run them with the same preciseness as intended by their developers. Four prior arts, *i.e.*, FCN [11], U-Net [18], DeepLabv3+ [14], and DANet [15], are compared. For all networks, we use Inception-v3 as the backbone of their contracting paths. In addition, as the majority of the existing works utilize a patch-based sliding window approach to detect retinal lesions, we include patch-based FCN-32s. To train the patch-based model, we uniformly divide each image into 4×4 patches, each sized to 224×224. Given a test image, the model is run with a window size of 224×224 and a stride of 112. Scores from overlapped areas are averaged. All the baselines are trained with the Dice loss.

Comparing Lesion-Net with distinct settings. As Table II shows, the overall performance of Lesion-Net increases first,

[3]The segmentation maps here are already down-sampled to the same size as the feature maps in the main branch.

TABLE II
LESION SEGMENTATION AND CLASSIFICATION BY DIFFERENT MODELS.

Model	Lesion segmentation									Lesion classification								
	Mean	MA	iHE	HaEx	CWS	vHE	pHE	NV	FiP	**Mean**	MA	iHE	HaEx	CWS	vHE	pHE	NV	FiP
patch FCN-32s	0.553	0.209	0.583	0.714	0.535	0.622	0.549	0.554	0.659	0.704	0.886	0.849	0.828	0.720	0.634	0.544	0.637	0.535
FCN-32s	0.571	0.327	0.592	0.728	0.528	0.642	0.562	0.530	0.662	0.769	0.900	0.858	0.856	0.771	0.722	0.683	0.694	0.669
FCN-16s	0.587	0.369	0.608	0.737	0.575	0.639	0.515	0.581	0.671	0.787	0.890	0.849	0.847	0.743	0.758	0.726	0.696	0.783
FCN-8s	0.586	0.369	0.609	0.740	0.573	0.640	0.534	**0.583**	0.639	0.778	0.891	0.858	0.854	0.749	0.766	0.711	0.671	0.725
U-Net	0.570	0.384	0.598	0.730	0.565	0.547	0.604	0.538	0.592	0.757	0.888	0.855	0.843	0.755	0.639	0.689	0.653	0.737
DeepLabv3+	0.553	0.367	0.612	0.732	0.558	0.550	0.477	0.498	0.631	0.794	0.899	0.863	0.866	0.764	**0.800**	0.693	0.677	0.792
DANet	0.585	0.351	0.608	0.733	0.560	0.623	0.589	0.543	0.671	0.775	0.900	0.853	0.852	0.772	0.713	0.682	0.715	0.712
Inception-v3	–	–	–	–	–	–	–	–	–	0.716	0.895	0.893	0.865	0.766	0.500	0.540	0.594	0.678
ABN-lesion	–	–	–	–	–	–	–	–	–	0.726	0.900	**0.900**	**0.871**	0.761	0.519	0.552	0.627	0.678
Lesion-Net (Dual Dice loss)																		
Lesion-Net-32s	0.573	0.289	0.590	0.730	0.539	0.632	0.536	0.582	**0.687**	0.792	0.899	0.881	0.857	0.778	0.720	**0.773**	0.669	0.762
Lesion-Net-16s	0.591	0.377	0.612	0.740	0.565	0.645	0.590	0.571	0.623	**0.801**	0.902	0.882	0.866	0.792	0.733	0.726	0.701	**0.807**
Lesion-Net-8s	**0.603**	0.377	**0.617**	0.740	0.575	**0.648**	**0.616**	0.580	0.667	0.780	0.900	0.881	0.861	0.771	0.687	0.693	0.711	0.733
Lesion-Net-4s	0.592	0.394	0.614	0.743	**0.577**	0.633	0.588	0.570	0.616	0.781	**0.904**	0.883	0.862	0.791	0.678	0.660	**0.719**	0.748
Lesion-Net-2s	0.581	0.381	0.614	**0.744**	0.567	0.634	0.569	0.565	0.572	0.787	0.900	0.893	0.868	**0.793**	0.706	0.706	0.667	0.764
Lesion-Net-16s (WCE)	0.364	0.180	0.389	0.543	0.346	0.424	0.245	0.363	0.423	0.534	0.864	0.822	0.780	0.574	0.324	0.274	0.354	0.282
Lesion-Net-16s (Focal)	0.458	0.165	0.479	0.682	0.409	0.471	0.488	0.423	0.548	0.745	0.880	0.869	0.837	0.723	0.650	0.693	0.627	0.683
Lesion-Net-16s (Dice)	0.594	0.362	0.609	0.734	0.570	0.637	0.587	0.573	0.683	0.769	0.899	0.860	0.851	0.754	0.709	0.661	0.694	0.724

from 0.573 (Lesion-Net-32s) to 0.591 (Lesion-Net-16s), and decreases later, from 0.592 (Lesion-Net-4s) to 0.581 (Lesion-Net-2s). The peak is obtained by Lesion-Net-8s, with an F1 of 0.603. The result confirms our hypothesis that when the network parameters keep increasing, the additional layers can have a negative effect on the performance.

Comparing loss functions. As the parameter α in Focal loss [20] is dataset-dependent, we set it to 0.8 according to our validation set, with the parameter γ set to 2 as suggested in the original paper. Dice and the proposed dual loss outperform Focal and WCE [41] with a large margin, see Table II. Correcting small-sized errors cannot be well reflected by the pixel-wise F1 score. This explains the relatively small difference between Dice and the dual loss for lesion segmentation.

Comparing with the baselines. Lesion-Net outperforms the baselines. Patch-based FCN-32s is less effective than its full-resolution counterpart. As noted in Section II, properly recognizing MA requires a holistic view, which is absent for the patch-based model. This explains its lowest performance (F1 of 0.209) on this lesion. Patch-based FCN-32s also has difficulty in segmenting large lesions such as vHE and pHE. Compared to DeepLabv3+, Lesion-Net shows similar performance on MA, iHE, HaEx and CWS while noticeably better for vHE, pHE, NV and FiP. Comparing the two groups of lesions, the latter lack clear boundaries. As shown in Fig. 4, irregular segmentation boundaries produced by DeepLabv3+ implies its attempt to produce precise boundaries, which are however unnecessary for retinal lesions. The results confirm our hypothesis that DeepLabv3+ is over-designed for this task. In the meantime, the viability of the proposed Lesion-Net for retinal lesion segmentation is justified.

As shown in Fig. 5, for small-sized lesions with relatively clear boundaries (MA, iHE, CWS and HaEx), we observe close performance among distinct models. Exceptions are L-Net-32s and FCN-32s, as they do $32\times$ upsampling by parameter-free bilinear interpolation, and thus difficult to accurately locate small lesions. For large lesions yet with imprecise boundaries (NV, pHD and vHE), the simplicity of FCN and L-Net becomes advantageous. The L-Net series produce more smooth segmentation boundaries. The fact that the top performer for FiP is L-Net-32s is due to the relatively clear boundary of this large lesion.

C. Experiment 2. Lesion Classification

Baselines. We re-use the baselines from Experiment 1, with lesion classification obtained by global max pooling on segmentations. We also compare with two segmentation-free models, *i.e.*, Inception-v3 [26] and ABN [24], both trained using image-level lesion annotations and Dice.

Comparing Lesion-Net with distinct settings. As Table II shows, for lesion classification Lesion-Net with a shorter expansive path, *e.g.* Lesion-Net-16s and Lesion-Net-32s, is preferred. From Table II we see that Lesion-Net trained with the dual loss is the best, suggesting small misclassified blobs are reduced.

Comparing with the baselines. Inception-v3 and ABN are less effective than the majority of the segmentation based models. The results suggest the importance of lesions' spatial information even for making image-level predictions. Different from its behavior for lesion segmentation, DeepLabv3+ becomes runner-up for lesion classification. For vHE, this model outperforms the others with a large margin. Note that DeepLabv3+ is specifically designed to capture multi-scale information by its parallel dialated convolutions. This design appears to be good at capturing the major pattern of vHE which often occupies more than half of an image. Overall Lesion-Net-16s is the best.

D. Experiment 3. Lesions for DR Grading

Baselines. We again compare with Inception-v3 and ABN, both re-trained for DR grading. One might also consider a more straightforward method that enriches the output of the GAP layer by concatenating the m-dimensional lesion vector $(P_1, \ldots, P_m)$. Accordingly, the size of the fully connected layer is adjusted to $(2,048 + m) \times 5$. Note that similar ideas have been exploited in the context of image captioning for obtaining semantically enhanced image features [46]. We term this baseline Lesion-Concat.

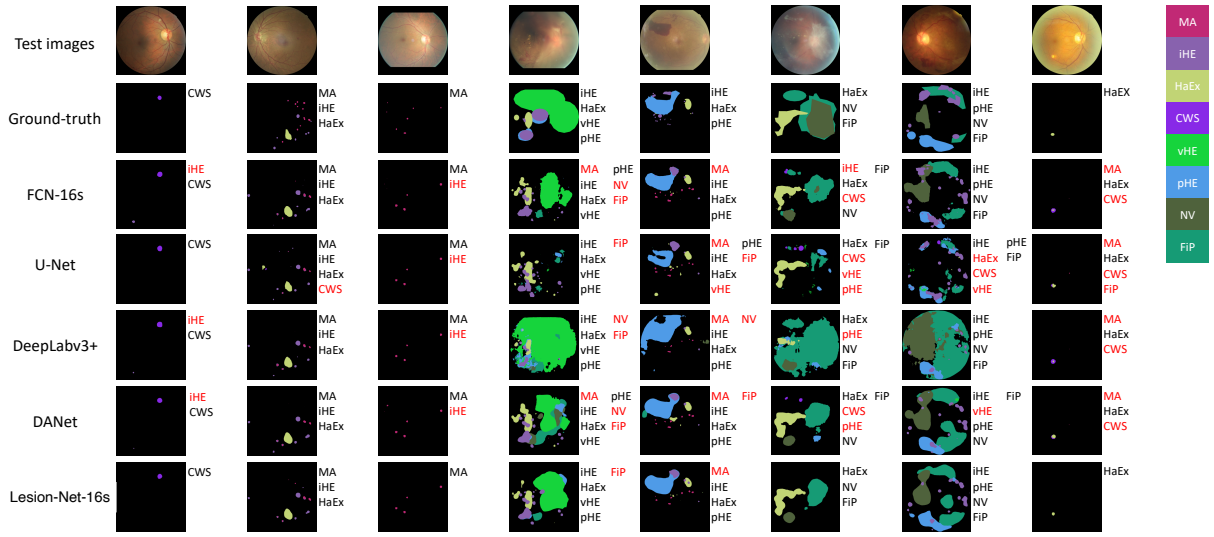

Fig. 4. **Some qualitative results of lesion segmentation and classification**. Red font indicates false alarms. The proposed Lesion-Net-16s produces more smooth segmentation boundaries and less false alarms. Best viewed digitally.

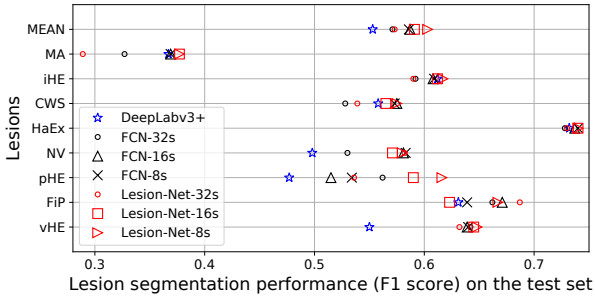

Fig. 5. Illustrating Table II, lesions sorted by their average size.

TABLE III
DR GRADING RESULTS. NUMBERS AFTER INCEPTION-V3 MEANS THE INPUT RESOLUTION.

DR model	Attention weights	Lesion model	Kappa
Inception-v3 (224 × 224)	–	–	0.660
Inception-v3 (488 × 448)	–	–	0.729
Inception-v3	–	–	0.774
Lesion-Concat	–	Inception-v3	0.780
Multi-task network	Conv (Eq. 7)	U-Net	0.780
Multi-task network	CW-MaxPool (Eq. 6)	Lesion-Net-16s	0.781
Multi-task network	Conv (Eq. 7)	FCN-8s	0.787
Multi-task network	Conv (Eq. 7)	DeepLabv3+	0.787
Lesion-Concat	–	Lesion-Net-16s	0.788
ABN-grading	–	–	0.797
Multi-task network	Conv (Eq. 7)	Lesion-Net-16s	**0.803**

Results. We use Lesion-Net-16s in the multi-task network for its best overall performance in the previous experiments. As Table III shows, using a better lesion segmentation model results in more accurate DR grading, with the multi-task network (Lesion-Net-16s) as the top performer. The better performance of learned weights (Eq. 7) compared to CW-MaxPool (Eq. 6) supports our statement that a region deemed to be negative with respect to the lesions does not necessarily mean it is useless for DR grading

We summarize the performance in Table IV. When compared to the best baseline per task, the improvement seems to be not significant. However, for the best overall performance, one has to simultaneously deploy three distinct baselines (FCN-8s, DeepLabv3+ and ABN-grading) with 3.2GB GPU memory at run time, while our multi-task network performs

TABLE IV
OVERALL PERFORMANCE OF DIFFERENT MODELS.

Model	Lesion segmentation	Lesion Classification	DR grading
FCN-8s	0.586	0.778	-
DeepLabv3+	0.553	0.794	-
ABN-grading	-	-	0.797
Our model	**0.591**	**0.801**	**0.803**

better in all three tasks with half GPU memory (1.5GB).

V. CONCLUSIONS

We have developed a multi-task deep learning approach to lesion segmentation, lesion classification and disease classification for color fundus images. Extensive experiments justify the superiority of the proposed approach against the prior art. The proposed Lesion-Net, with its re-designed expansive path and the proposed dual loss, is found to be effective for learning from retinal lesion annotations with imprecise boundaries. Exploiting Lesion-Net as a side-attention branch, the multi-task network simultaneously improves DR grading and interprets the decision with lesion maps.

While working on fundus images, our work reveals good practices for developing a semantic segmentation network given training data with imprecise object boundaries and extremely imbalanced classes, and for converting attributes predicted at pixel-level to categories at a higher level. We believe the lessons learned are beyond the specific domain.

ACKNOWLEDGMENTS

This research was supported in part by the National Natural Science Foundation of China (No. 61672523), Beijing Natural Science Foundation (No. 4202033), Beijing Natural Science Foundation Haidian Original Innovation Joint Fund (No. 19L2062), the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (No. 2018PT32029), CAMS Initiative for Innovative Medicine (CAMS-I2M, 2018-I2M-AI-001), and the Pharmaceutical Collaborative Innovation Research Project of Beijing Science and Technology Commission (No. Z191100007719002). Corresponding authors: Xirong Li and Youxin Chen.

REFERENCES

[1] V. Gulshan, L. Peng, M. Coram, M. 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*, vol. 316, no. 22, pp. 2402–2410, 2016.

[2] X. Wang, L. Ju, X. Zhao, and Z. Ge, "Retinal abnormalities recognition using regional multitask learning," in *MICCAI*, 2019.

[3] W. Wang, Z. Xu, W. Yu, J. Zhao, J. Yang, F. He, Z. Yang, D. Chen, D. Ding, Y. Chen, and X. Li, "Two-stream CNN with loose pair training for multi-modal AMD categorization," in *MICCAI*, 2019.

[4] H. Fu, J. Cheng, Y. Xu, D. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE transactions on medical imaging*, vol. 37, no. 7, pp. 1597–1605, 2018.

[5] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "CE-Net: Context encoder network for 2d medical image segmentation," *IEEE transactions on medical imaging*, 2019.

[6] T. Laibacher, T. Weyde, and S. Jalali, "M2U-Net: Effective and efficient retinal vessel segmentation for real-world applications," in *CVPRW*, 2019.

[7] T. Durand, T. Mordan, N. Thome, and M. Cord, "WILDCAT: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *CVPR*, 2017.

[8] W. Ge, S. Yang, and Y. Yu, "Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning," in *CVPR*, 2018.

[9] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," in *CVPR*, 2018.

[10] S. D. Solomon, E. Chew, E. J. Duh, L. Sobrin, J. K. Sun, B. L. VanderBeek, C. C. Wykoff, and T. W. Gardner, "Diabetic retinopathy: A position statement by the American Diabetes Association," *Diabetes Care*, vol. 40, no. 3, pp. 412–418, 2017.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[12] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *T-PAMI*, vol. 39, no. 12, pp. 2481–2495, 2017.

[14] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[15] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *CVPR*, 2019.

[16] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *ICCV*, 2019.

[17] A. C. Berg, T. L. Berg, H. Daum, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi, "Understanding and predicting importance in images," in *CVPR*, 2012.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[19] R. Hu, P. Dollar, K. He, T. Darrell, and R. Girshick, "Learning to segment every thing," in *CVPR*, 2018.

[20] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017.

[21] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *CVPR*, 2018.

[22] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3DV*, 2016.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[24] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch ntwork: Learning of attention mechanism for visual explanation," in *CVPR*, 2019.

[25] Kaggle, "Diabetic retinopathy detection," https://www.kaggle.com/c/diabetic-retinopathy-detection, 2015.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[27] P. Porwal, S. Pachade, M. Kokare *et al.*, "IDRiD: Diabetic retinopathy segmentation and grading challenge," *MIA*, vol. 59, p. 101561, 2020.

[28] A. Fleming, K. Goatman, S. Philip, G. Williams, G. Prescott, G. Scotland, P. McNamee, G. Leese, W. Wykes, P. Sharp *et al.*, "The role of haemorrhage and exudate detection in automated grading of diabetic retinopathy," *Br. J. Ophthalmology*, vol. 94, no. 6, pp. 706–711, 2010.

[29] M. Niemeijer, B. van Ginneken, S. Russell, M. Suttorp-Schulten, and M. Abramoff, "Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis," *IOVS*, vol. 48, no. 5, pp. 2260–2267, 2007.

[30] M. van Grinsven, B. van Ginneken, C. Hoyng, T. Theelen, and C. Sánchez, "Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1273–1284, 2016.

[31] J. Tan, H. Fujita, S. Sivaprasad, S. Bhandary, K. Rao, K. Chua, and R. Acharya, "Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network," *Inf. Sci.*, vol. 420, pp. 66–76, 2017.

[32] C. Lam, C. Yu, L. Huang, and D. Rubin, "Retinal lesion detection with deep learning using image patches," *IOVS*, vol. 59, no. 1, p. 590, 2018.

[33] C. Playout, R. Duval, and F. Cheriet, "A multitask learning architecture for simultaneous segmentation of bright and red lesions in fundus images," in *MICCAI*, 2018.

[34] 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*, 2017.

[35] Z. Lin, R. Guo, Y. Wang, B. Wu, T. Chen, W. Wang, D. Chen, and J. Wu, "A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion," in *MICCAI*, 2018.

[36] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[39] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019.

[40] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *IJCV*, vol. 111, no. 1, pp. 98–136, 2015.

[41] C. Sudre, W. Li, T. Vercauteren, S. Ourselin, and J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2017.

[42] R. Zhang, H. Zhang, and A. Chung, "A unified mammogram analysis method via hybrid deep supervision," in *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Springer, 2018, pp. 107–115.

[43] F. Grassmann, J. Mengelkamp, C. Brandl, S. Harsch, M. E. Zimmermann, B. Linkohr, A. Peters, I. M. Heid, C. Palm, and B. H. Weber, "A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography," *Ophthalmology*, vol. 125, no. 9, pp. 1410–1420, 2018.

[44] AAO, "Diabetic retinopathy preferred practice pattern," https://www.aao.org/preferred-practice-pattern/diabetic-retinopathy-ppp-updated-2017, 2017.

[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *CVPR*, 2009.

[46] X. Li, C. Xu, X. Wang, W. Lan, Z. Jia, G. Yang, and J. Xu, "COCO-CN for cross-lingual image tagging, captioning and retrieval," *T-MM*, vol. 21, no. 9, pp. 2347–2360, 2019.