Lesion-Aware Transformers for Diabetic Retinopathy Grading

Rui Sun[1*], Yihao Li[1*], Tianzhu Zhang[1†], Zhendong Mao[1], Feng Wu[1], Yongdong Zhang[1]
[1]University of Science and Technology of China
{issunrui, luoheliyihao}@mail.ustc.edu.cn, {tzzhang, zdmao, fengwu, zhyd73}@ustc.edu.cn

Abstract

Diabetic retinopathy (DR) is the leading cause of permanent blindness in the working-age population. And automatic DR diagnosis can assist ophthalmologists to design tailored treatments for patients, including DR grading and lesion discovery. However, most of existing methods treat DR grading and lesion discovery as two independent tasks, which require lesion annotations as a learning guidance and limits the actual deployment. To alleviate this problem, we propose a novel lesion-aware transformer (LAT) for DR grading and lesion discovery jointly in a unified deep model via an encoder-decoder structure including a pixel relation based encoder and a lesion filter based decoder. The proposed LAT enjoys several merits. First, to the best of our knowledge, this is the first work to formulate lesion discovery as a weakly supervised lesion localization problem via a transformer decoder. Second, to learn lesion filters well with only image-level labels, we design two effective mechanisms including lesion region importance and lesion region diversity for identifying diverse lesion regions. Extensive experimental results on three challenging benchmarks including Messidor-1, Messidor-2 and EyePACS demonstrate that the proposed LAT performs favorably against state-of-the-art DR grading and lesion discovery methods.

1. Introduction

Diabetic retinopathy (DR) is one of the most severe complications of blood vessel damage triggered by diabetes, which can lead to vision impairment and even irreversible blindness [13, 6, 25]. Usually, as shown in Figure 1 (a), ophthalmologists identify DR severity based on the type and number of associated lesion symptoms, such as microaneurysms, haemorrhages, soft exudates and hard exudates [46, 22]. According to the international protocol [17, 35], the severity of DR can be divided into five grades, including normal, mild, moderate, severe non-proliferative, and proliferative. These five grades can also be fused into binary classification, *i.e. no DR* (normal) versus *DR* (abnormal), or *non-referable* (normal and mild DR) versus *referable* (moderate and worse DR) [17, 34, 41].

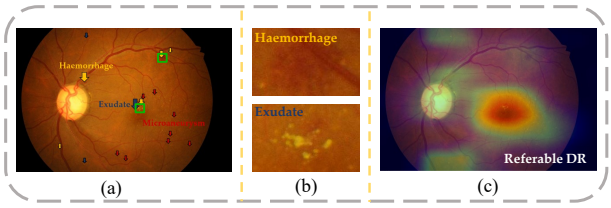


Figure 1: (a) A sample fundus image with different lesions is annotated. The arrows indicate the main DR-related lesions, among which the yellow, red and blue arrows represent haemorrhages, microaneurysms and exudates, respectively. (b) The lesion regions marked by the green bounding box in (a) are zoomed in, and we can observe that pixel appearances of the same lesion region tend to be similar. (c) Our model can achieve DR grading and lesion discovery jointly by using only the severity level labels.

Recently, with the development of deep learning, rapid and automatic DR diagnostic models have been proposed based on pixel-level supervision [46], or patch-level supervision [41, 23, 40]. However, their flexibility and scalability are limited in the actual deployment because the annotation of fundus images requires manual labeling by experienced domain experts [22, 27, 32]. In addition, the identification of lesion regions in fundus images is also very important, since it provides visual instructions for ophthalmologists to assist their diagnosis [12, 26, 30]. However, most existing methods treat DR grading and lesion discovery as two independent tasks, and they both require lesion annotations as a guide for learning. To overcome these issues, in [34], a weakly supervised learning model based on DR severity grades has been proposed for simultaneously grading DR and highlighting lesion regions. Unfortunately, it tends to be biased on the most important lesion regions while ignoring trivial lesion information contained in the fundus images, which may impair the performance of lesion location. Furthermore, the less discriminating regions found at a certain severity level may be important for other severi-

*Equal contribution
†Corresponding author

ty grades. Therefore, it is desirable to design an effective model to obtain more complete lesion regions and their importance for DR grading.

Based on the above discussions, to achieve accurate DR grading and complete lesion discovery simultaneously, we need to consider the following three aspects. (1) As shown in Figure 1 (a), the distribution of lesion areas contained in fundus images is usually relatively sparse. Besides, the appearance of pixels in the same lesion region is similar, but is different from the background pixels, as shown in Figure 1 (b). Therefore, it is necessary to model the correlation between pixels for robust feature learning. (2) The importance of different lesion regions in each image should be considered. The observation is that not all lesion information is beneficial to a particular DR severity level, and even some lesion information is noise signal. Therefore, we should evaluate and adaptively fuse the contribution of each lesion region. (3) As shown in Figure 1 (a), each fundus image may contain multiple different lesions. Moreover, even fundus images of the same severity grade may contain inconsistencies in the type and number of lesions. Thus, it is desirable to make the lesion-aware features diverse, namely, capturing the corresponding lesion features from as many lesion regions as possible. Besides, since each region indicates a specific type of lesion or a combination of different lesions, the compactness of the lesion features should also be considered. In other words, the lesion features obtained from the same lesion filter are encouraged to approach each other to form a more compact distribution. In this way, each lesion region can suggest more explicit lesion semantics and different regions are combined to form a complete lesion discovery, as shown in Figure 1 (c).

Motivated by the above observations, we propose a novel lesion-aware transformer (LAT) for DR prediction and lesion discovery in a unified deep model via an encoder-decoder structure including a pixel relation based encoder and a lesion filter based decoder. In the **pixel relation based encoder**, we propose a self-attention mechanism to adapt to pixel appearance variations. In specific, we model the correlation of pixels to capture full-image context information. In other words, it is to realize the aggregation of lesion pixels with similar appearances and the suppression of cluttered background pixels. In the **lesion filter based decoder**, we design a self-attention module and a cross-attention module to learn lesion-aware filters for lesion discovery in a given dataset. In the self-attention module, we model the interactions between lesion filters to increase their discrepancies. In the cross-attention module, given an input fundus image, we treat the pixels of feature map as keys and values. And we store each lesion filter as a query, then the corresponding region activation map based on the similarity between a specific query and keys can be obtained. Each lesion region activation map denotes the spatial distribution of one specific lesion. With the region activation map, we can get the lesion-aware features by adaptively blending values. Without the specific lesion information as supervision signals, it is difficult to learn the lesion filters well. Therefore, to learn lesion filters well with only the severity level labels, we design two mechanisms including a lesion region importance learning mechanism and a lesion region diversity learning mechanism to constrain the lesion-aware features. For the importance learning mechanism, we introduce an importance prediction module to evaluate and adaptively fuse the contribution of each lesion region. For the diversity learning mechanism, we adopt a triplet loss based on the hard negative mining strategy to achieve the diversity and compactness of the lesion-aware features simultaneously. Then, based on the lesion-aware features, we add a classification module containing a global consistency constraint loss for DR grading. By optimizing the encoder-decoder structure and the classification module jointly, the lesion-aware filters can be learned through the whole dataset during training. As a result, we can achieve DR prediction and lesion discovery in a unified deep model.

To sum up, the contributions of this work can be summarized as follows: (1) We propose a novel lesion-aware transformer (LAT) to achieve DR grading and lesion discovery jointly in a unified deep model via an encoder-decoder structure including a pixel relation based encoder and a lesion filter based decoder. (2) To the best of our knowledge, this is the first work to formulate lesion discovery as a weakly supervised lesion localization problem via a transformer decoder. To learn lesion filters well with only image-level labels, we design two effective mechanisms including lesion region importance and lesion region diversity. (3) Extensive experimental results on three challenging benchmarks including Messidor-1, Messidor-2 and EyePACS demonstrate that the proposed LAT performs favorably against state-of-the-art DR grading methods.

2. Related Work

In this section, we briefly overview methods that are related to diabetic retinopathy assessment, weakly supervised object localization and attention-based transformers.

Diabetic Retinopathy Assessment. Early methods on automatic diabetic retinopathy assessment involve two tasks including DR grading and lesion discovery. For lesion discovery, in order to assist ophthalmologists to make accurate diagnosis, a series of approaches based on pixel-level [12, 38, 8, 26] or patch-level annotation [29, 30] have been proposed. Recently, deep features have become popular for DR grading [18, 14, 36, 22]. Unlike handcrafted features, deep features are more discriminative. Generally, existing DR grading methods can be divided into two main categories. The first category is to use lesion information to assist DR classification [3, 23]. In specific, Antal and Balin-

t [3] detect microaneurysms and predict the DR severity level based on the presence or absence of microaneurysms. In [23], Lin *et al*. extract lesion information with the original image for DR grading. The second category only uses image-level supervision for DR grading [22, 14]. In [22], Li *et al*. present a novel attention network for DR prediction by exploring the internal relationship with diabetic macular edema. However, the above methods treat DR grading and lesion discovery as two independent tasks. In order to alleviate this limitation, several methods [40, 46] have been proposed to achieve two tasks simultaneously. In specific, Yang *et al*. [40] propose a two-stage framework for both lesion detection and DR grading by using the lesion annotations. In [46], a collaborative learning mechanism is proposed for both lesion segmentation and DR grading. Although the above methods have achieved remarkable progress, most of them require pixel-level or patch-level lesion annotations, which is time-consuming and laborious. Recently, Wang *et al*. [34] adopt attention maps to highlight the suspicious regions and predict DR grading based on both suspicious patches and the fundus image. However, this method may tend to the most important lesion regions and impair the performance of lesion discovery. Different from previous methods, here, we formulate lesion discovery as a weakly supervised lesion localization problem via a transformer decoder. To learn lesion filters well with only image-level labels, we design two effective mechanisms including lesion region importance and lesion region diversity.

Weakly Supervised Object Localization. Weakly Supervised Object Localization (WSOL) aims to infer object positions and categories simultaneously with only image-level labels. In order to achieve this goal, Zhou *et al*. [45] utilize the Class Activation Mapping (CAM) to implement both object classification and localization. Later, Grad-CAM [28] and CCAM [39] have been proposed to obtain more robust localization performance. The inherent intuition inside the CAM-based methods is that the classification networks have the ability for mining the discriminative object regions. Unfortunately, a common issue for these methods is that they tend to focus on the most discriminative object regions which results in poor localization performance. To mitigate this issue, several methods [7, 42, 37, 24] explore objects context information to expand the most discriminative region to the entire object. Motivated by the above methods for WSOL, we formulate lesion discovery as a weakly supervised lesion localization problem. To achieve this goal, we design a novel lesion-aware transformer by considering lesion importance and lesion diversity for lesion filter learning with only image-level labels.

Attention-based Transformers. Since Vaswani *et al*. [31] have proposed the attention-based transformer, it has been widely applied in machine translation [15, 16], speech recognition [5], word representation learning [11] and object detection [4, 43, 47]. Transformer models introduce multi-head attention layers, similar to Non-Local Neural Networks [33], which can scan through each element of a sequence and updated it by aggregating global information from the whole sequence. Our model applies the idea of transformer to learn the lesion-aware filters through the encoder-decoder structure, so that we can discover as many lesions as possible and predict the final severity level at the same time.

3. Lesion-Aware Transformer Network

In this section, we describe the details of LAT for DR prediction and lesion discovery in a unified deep model via an encoder-decoder structure including a pixel relation based encoder and a lesion filter based decoder.

3.1. Overview

In DR grading, given a fundus image $\mathbf{I}$, let $\mathbf{F} \in \mathbb{R}^{H\times W\times D}$ denote the feature map extracted from a backbone network (*e.g*. ResNet50 [19]), where H, W and D denote the height, width and channel number of the feature map, respectively. And each image is associated with a ground truth label $\mathbf{z} \in \mathbb{R}^C$, where C represents the number of severity grades. During testing, given an image, the outputs are a predicted DR severity level $\tilde{\mathbf{y}}$ and the corresponding lesion activation map. As illustrated in Figure 2, the proposed LAT includes a pixel relation based encoder and a lesion filter based decoder. The pixel relation encoder is designed to adapt to pixel appearance variations by modeling the correlation of pixels, and the lesion filter based decoder is proposed to learn lesion-aware filters for lesion discovery.

3.2. Pixel Relation based Encoder

To capture the full-image context information, we model the correlation of pixels and aim to generate the enhanced feature map to adapt to pixel appearance variations by using a self-attention mechanism. In specific, we first utilize a convolution layer to reduce the channel dimension of the feature map to a smaller dimension L, and then flatten the spatial dimensions into one dimension to produce the new feature map $\mathbf{F} \in \mathbb{R}^{HW\times L}$. Then, for the n-th head, we obtain the queries, keys and values based on the feature map $\mathbf{F}$, denoted as $\mathbf{Q}_n, \mathbf{K}_n, \mathbf{V}_n \in \mathbb{R}^{HW\times L/8}$.

$$\mathbf{Q}_n = \mathbf{F}\mathbf{W}_n^Q, \mathbf{K}_n = \mathbf{F}\mathbf{W}_n^K, \mathbf{V}_n = \mathbf{F}\mathbf{W}_n^V, \qquad (1)$$

where $n = 1, 2, \ldots, N$, and N is the number of heads in the multi-head attention mechanism. In this work, N is set to 8. $\mathbf{W}_n^Q, \mathbf{W}_n^K, \mathbf{W}_n^V \in \mathbb{R}^{L\times L/8}$ are linear projections. Then, we calculate the attention weight $\mathbf{S}_n \in \mathbb{R}^{HW\times HW}$, which

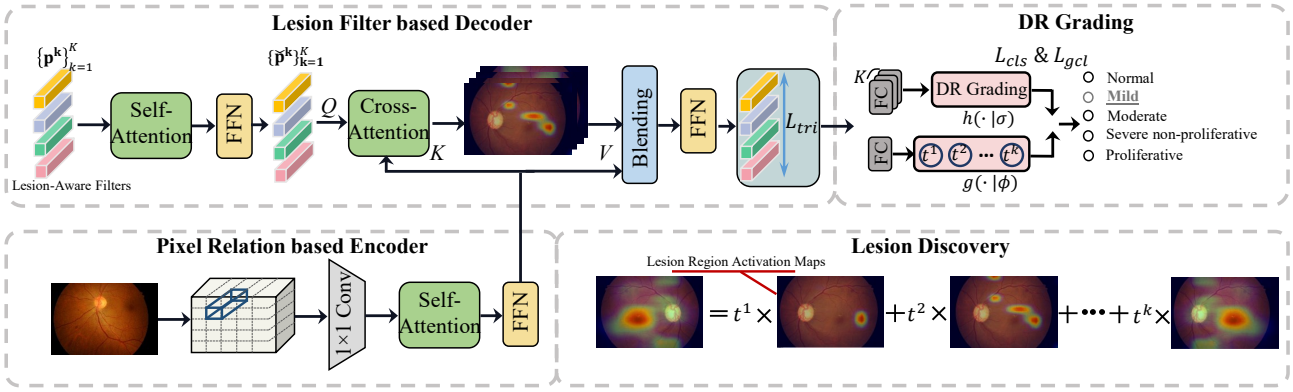


Figure 2: The architecture of our LAT including an encoder-decoder structure and the classification module. By optimizing the encoder-decoder structure and the classification module jointly, the lesion-aware filters can be learned to identify diverse lesion regions for DR grading and lesion discovery. Please refer to the Supplementary Material for details on self-attention, cross-attention and feed-forward network (FFN).

models the interdependencies between different pixels,

$$\mathbf{S}_n = \text{softmax}\left(\frac{\mathbf{Q}_n\mathbf{K}_n^{\mathsf{T}}}{\sqrt{L/8}}\right), \tag{2}$$

where $\sqrt{L/8}$ is a scaling factor. With the attention weight $\mathbf{S}_n$, we can get the output of the head $\mathbf{H}_n \in \mathbb{R}^{HW\times L/8}$ by adaptively blending values,

$$\mathbf{H}_n = \mathbf{S}_n\mathbf{V}_n. \tag{3}$$

We concatenate all single head outputs $\{\mathbf{H}_n\}_{n=1}^N$ along the channel dimension and obtain the final output $\mathbf{H} \in \mathbb{R}^{HW\times L}$ through a projection matrix $\mathbf{W}^O \in \mathbb{R}^{L\times L}$,

$$\mathbf{H} = [\mathbf{H}_1; \mathbf{H}_2; \ldots; \mathbf{H}_N]\,\mathbf{W}^O. \tag{4}$$

The final output $\mathbf{H}$ can be further fed into the feed-forward network (FFN) containing two fully connected layers to produce the enhanced feature map $\tilde{\mathbf{F}} \in \mathbb{R}^{HW\times L}$. Through the self-attention operation, the pixels of the lesion region with similar appearance can be gathered, which also means the better suppression of the messy background pixels caused by under/overexposure and out-of-focus problems.

3.3. Lesion Filter based Decoder

In order to identify different lesion regions, we design the lesion filter based decoder to learn lesion-aware filters. We first learn a set of lesion-aware filters $\mathbf{P} = \{\mathbf{p}^k\}_{k=1}^K$, and each filter $\mathbf{p}^k$ is represented as a L-dimension vector to recognize whether the pixels in the image belong to this lesion region. Then, we use a self-attention mechanism to further incorporate context information from other filters to increase their discrepancies. The implementation is similar to Section 3.2 and only use $\{\mathbf{p}^k\}_{k=1}^K$ to replace $\mathbf{F} \in \mathbb{R}^{HW\times L}$ as the input. In this module, the output is the updated lesion-aware filters $\tilde{\mathbf{P}} = \{\tilde{\mathbf{p}}^k\}_{k=1}^K$. Then we propose a cross-attention mechanism to obtain lesion-aware activation maps $\mathbf{M} \in \mathbb{R}^{H\times W\times K}$. Specifically, we denote lesion-filters $\{\tilde{\mathbf{p}}^k\}_{k=1}^K$ as queries, the enhanced feature map $\tilde{\mathbf{F}} \in \mathbb{R}^{HW\times L}$ as keys and values. Formally,

$$\mathbf{Q}_n = \tilde{\mathbf{P}}\mathbf{W}_n^Q, \quad \mathbf{K}_n = \tilde{\mathbf{F}}\mathbf{W}_n^K, \quad \mathbf{V}_n = \tilde{\mathbf{F}}\mathbf{W}_n^V, \tag{5}$$

where $\mathbf{W}_n^Q, \mathbf{W}_n^K, \mathbf{W}_n^V \in \mathbb{R}^{L\times L/8}$ are linear projections, and $n = 1, 2, \ldots, N$. Then we have

$$\mathbf{S}_n = \text{softmax}\left(\frac{\mathbf{Q}_n\mathbf{K}_n^{\mathsf{T}}}{\sqrt{L/8}}\right), \tag{6}$$

where $\sqrt{L/8}$ is a scaling factor, and $\mathbf{S}_n$ denotes similarities of the n-th head between the enhanced feature map $\tilde{\mathbf{F}}$ and the enhanced lesion-aware filters $\tilde{\mathbf{P}}$. Then, the lesion-aware activation map $\mathbf{M}$ can be calculated as

$$\mathbf{M} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{S}_n, \tag{7}$$

where N is the number of heads, $\mathbf{M} = \{\mathbf{M}^k\}_{k=1}^K$ denotes a set of lesion-aware activation maps, and the $\mathbf{M}^k \in R^{H\times W}$ corresponds to the k-th lesion-aware activation map. Each lesion region activation map denotes the spatial distribution of one specific lesion, that is to say, the activation map has high response values at the pixels belonging to the corresponding lesion. After obtaining the similarities $\mathbf{S}_n$, we can calculate the multi-head attention output $\mathbf{H}$ according to (3) and (4). Then it is fed to the feed-forward network (FFN) to obtain a set of lesion-aware features $\mathbf{X} = \{\mathbf{x}^k\}_{k=1}^K$.

Without the specific lesion information as supervision signals, it is difficult to learn the lesion filters well. Therefore, to learn lesion filters well with only the severity level labels, we design two mechanisms to constrain the lesion-aware features, the details are as follows.

Lesion Region Importance Learning Mechanism. Considering that not all lesion information is beneficial to a particular DR level, we should evaluate and incorporate the contribution of each lesion region. In specific, we design an importance prediction module $g(\cdot|\phi)$, parameterized by ϕ, to evaluate the importance for lesion-aware features $\{\mathbf{x}^k\}_{k=1}^K$ and generate importance weights $\{\mathbf{t}^k\}_{k=1}^K$. The prediction module $g(\cdot|\phi)$ is a linear layer followed by a sigmoid operation to output probabilities between 0 and 1.

$$t^k = g(\mathbf{p}^k|\phi), \tag{8}$$

Lesion Region Diversity Learning Mechanism. Meanwhile, we adopt a triplet loss [44] based on the hard negative mining strategy [20] to simultaneously achieve the diversity and the compactness of the lesion-aware features. The introduced triplet loss is based on a mini-batch of T images. Therefore, we rewrite the lesion-aware features $\{\mathbf{x}^k\}_{k=1}^K$ as $\{\mathbf{x}_m^k\}_{k=1}^K$, where $m = 1, 2, \ldots, T$. The triplet loss is trained on a series of triplets, and each triplet consists of an anchor with the label k, a positive lesion feature with the same label and a negative lesion feature with different label. By treating the $\mathbf{x}_m^k$ as the anchor, the hardest positive pair and negative pair distances are defined as follows

$$\begin{aligned} d_+(m,k) &= \min_q \frac{\langle \mathbf{x}_m^k, \mathbf{x}_q^k \rangle}{\| \mathbf{x}_m^k \|_2 \| \mathbf{x}_q^k \|_2}, \\ d_-(m,k) &= \max_{z,n} \frac{\langle \mathbf{x}_m^k, \mathbf{x}_n^z \rangle}{\| \mathbf{x}_m^k \|_2 \| \mathbf{x}_n^z \|_2}, \end{aligned} \tag{9}$$

where $q \neq m, z \neq k$. $q, n = 1, 2, \ldots, T$ and $z = 1, 2, \ldots, K$.

The triplet loss function is adopted to reduce the distances of the hardest positive pairs and increase the distances of the hardest negative pairs.

$$L_{tri} = \sum_{m=1}^{T} \sum_{k=1}^{K} \left[d_+(m,k) - d_-(m,k) + \alpha \right]_+ . \tag{10}$$

Where α is the margin between positive and negative pairs, and $[b]_+ = \max(b, 0)$. In this way, each lesion region can suggest more explicit lesion semantics and different regions are combined to form a complete lesion discovery.

3.4. DR Grading

For DR grading, we add a classification module, which contains a global consistency loss based on the lesion-aware features. In specific, lesion part-aware features $\{\mathbf{x}_m^k\}_{k=1}^K$ are fed to generate the DR severity level prediction $\{\mathbf{y}_m^k\}_{k=1}^K$. $\mathbf{y}_m^k \in \mathbb{R}^C$ is corresponding to the k-th grade prediction in the m-th image by the classification module, which consists of K fully connected layers and the k-th layer is denoted as $h(\cdot|\sigma^k)$, parameterized by σ^k.

$$\mathbf{y}_m^k = h(\mathbf{x}_m^k|\sigma^k), \tag{11}$$

where $k = 1, 2, \ldots, K$ and $m = 1, 2, \ldots, T$. The final DR prediction $\tilde{\mathbf{y}}_m \in \mathbb{R}^C$ can be calculated by a weighted sum operation as follows

$$\tilde{\mathbf{y}}_m = \sum_{k=1}^{K} t_m^k \cdot \mathbf{y}_m^k. \tag{12}$$

The classification loss is given by the cross entropy loss between ground truth labels $\{\mathbf{z}_m\}_{m=1}^T$ and the predicted labels $\tilde{\mathbf{y}}_m$:

$$L_{cls}(\mathbf{z}_m, \tilde{\mathbf{y}}_m) = -\frac{1}{T} \sum_{m=1}^{T} \sum_{c=1}^{C} z_m^c \cdot \log \tilde{y}_m^c. \tag{13}$$

Besides, for the m-th image, assume that the ground truth label $\mathbf{z}_m$ corresponding to the c-th category, the overall lesion feature $\mathbf{o}_m^c \in \mathbb{R}^L$ is calculated by a weighted sum strategy with importance weights. Here, the importance weights $\{\mathbf{t}^k\}_{k=1}^K$ are also rewritten as $\{\mathbf{t}_m^k\}_{k=1}^K$ for the m-th image.

$$\mathbf{o}_m^c = \sum_{k=1}^{K} t_m^k \cdot \mathbf{x}_m^k. \tag{14}$$

Then, the local center of a particular category c in a mini-batch can be calculated as

$$\mathbf{o}^c = \frac{1}{T_c} \sum_{m=1}^{T_c} \mathbf{o}_m^c, \tag{15}$$

where T_c represents the number of fundus images for c-th in a minibatch. Meanwhile, we maintain a randomly initialized memory bank $\{\mathbf{b}^c\}_{c=1}^C$, and $\mathbf{b}^c \in \mathbb{R}^L$ can be represented as the c-th class center. The memory bank will update with moving average. Specifically, $\boldsymbol{b}^c = (1 - \eta_{t^c})\boldsymbol{b}^c + \eta_{t^c}\boldsymbol{o}^c$, where $\eta_{t^c} = e^{-t^c}$ is the updating rate of the class c, and t^c counts the number of category c in previous mini-batches. Then, we introduce the global consistency loss to align the local center $\mathbf{o}^c$ with the corresponding global class center by

$$L_{gcl} = \frac{1}{C} \sum_{c=1}^{C} \left\| \mathbf{b}^c - \mathbf{o}^c \right\|_2 . \tag{16}$$

We can alleviate problems such as Grades 0 and 1 that are difficult to distinguish by pushing class-specific local centers in each minibatch approaching their corresponding global centers, thereby improving DR grading performance.

3.5. Joint Training and Inference

By optimizing the encoder-decoder structure and the classification module jointly, the lesion-aware filters can be learned through the whole dataset during training. As a result, our LAT is trained by minimizing the overall objective as follows

$$L_{final} = L_{cls} + \lambda_{tri} L_{tri} + \lambda_{gcl} L_{gcl}. \tag{17}$$

Where λ_{tri} and λ_{gcl} are balance parameters. During the testing stage, for each fundus image, we can get fused activation map $\mathbf{A} \in \mathbb{R}^{H \times W}$ as follows,

$$\mathbf{A} = \sum_{k=1}^{K} t^k \mathbf{M}^k. \tag{18}$$

4. Experiments

4.1. Datasets and Evaluation Metrics

We conduct experiments on three benchmarks including Messidor-1 [10], Messidor-2 [21] and EyePACS [1].

Messidor-1 dataset [10] contains 1,200 fundus images from three French hospitals. However, their grading scale only has four grades, which is slightly different from the five-level international protocol [17, 35]. Therefore, we implement referral classification and normal classification. For referral classification, Grade 0 and Grade 1 are marked as non-referable, while Grade 2 and Grade 3 are considered referable. For normal classification, only Grade 0 is assigned as normal, and other grades are regarded as abnormal. Following the protocol in previous methods [22, 34], we use 10-fold cross validation on the entire dataset.

Messidor-2 dataset [10, 2, 21] is an extension of the original Messidor-1 dataset, which contains 1,748 fundus images, and each image is divided into one of five DR grades. Since there is no published method to compare on this dataset, we mainly adopt it to implement ablation studies to measure the effectiveness of each module of our proposed LAT by calculating Referral AUC and Kappa score.

EyePACS dataset [1, 9] contains 35,126 training images, 10,906 validation images and 42,670 testing images. The grading protocol is the same as the Messidor-2 dataset, with five DR categories. For this dataset, we train the model with the training set and evaluate the performance with the validation set and testing set respectively.

Evaluation Metrics. For the five DR categories, we use the quadratic weighted kappa metric [1], which can effectively reflect the performance of the model on the unbalanced dataset. The value of kappa varies between 0 and 1, with higher values indicating better model performance. And for the referral or normal classification, we evaluate by using the AUC (area under the ROC curve) metric.

4.2. Implementation Details

In this work, we use ResNet50 [19] as our backbone network for feature extraction by removing the global average pooling (GAP) layer and fully connected layer. The fundus images are resized to 512 $\times$ 512 and augmented with random horizontal flips, vertical flips and random cropping. Extra color jitter is adopted to reduce overfitting. There are two kinds of modules in our network including the importance prediction module $g(\cdot|\phi)$ and the classification module $h(\cdot|\sigma)$. The importance prediction module $g(\cdot|\phi)$ consists of a shared fully connected layer with 1 output channel followed by a sigmoid operation. The classification module $h(\cdot|\sigma)$ contains K fully connected layers with C output channels. Empirically, the weight λ_{tri} for the triplet loss and the λ_{gcl} for the global consistency loss are set to 0.04 and 0.01 respectively.

Table 1: Performance comparison with state-of-the-art methods on the Messidor-1 dataset.

Methods	Annotations	Referral AUC	Normal AUC
VNXK [32]	-	0.887	0.870
CKML [32]	-	0.891	0.862
Comp. CAD [27]	-	0.910	0.876
Expert A [27]	-	0.940	0.922
Expert B [27]	-	0.920	0.865
Zoom-in-Net [34]	-	0.957	0.921
AFN [23]	patch	0.968	-
Semi+Adv [46]	pixel	0.976	0.943
CANet [22]	-	0.963	-
LAT (ours)	-	**0.987**	**0.963**

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

DR Grading Performance. We compare the proposed method with various recent DR grading methods including VNXK [32], CKML [32], Comprehensive CAD [27], Expert A [27], Expert B [27], Zoom-in-Net [34], AFN [23], Semi+Adv [46], and CANet [22]. Table 1 shows the comparison of our model with state-of-the-art methods on the Messidor-1 dataset, where '*Annotations*' denotes whether additional lesion information is used as pixel-level or patch-level supervision signals to assist DR grading. According to these results, we can observe that our LAT outperforms all baseline methods by a large margin for referral classification and normal classification, even if some methods [23, 46] use additional lesion information to promote DR classification. Compared to the state-of-the-art method Semi+Adv [46], LAT using only image-level labels surpasses it by 2% on the Normal AUC metric. This is because most methods cannot effectively distinguish between Grade 0 and Grade 1, which degrades Normal AUC score. Our LAT can effectively alleviate this problem by forcing Grade 0 and Grade 1 to be close to the corresponding global center through the global consistency loss, which increases the discriminability. Besides, we can find that LAT obtains 3% Referral AUC and 4.2% Normal AUC gain over Zoom-in-Net [34], which uses the severity level label to implement DR grading and lesion discovery like ours. This can be attributed to the fact that our encoder-decoder architecture can find diverse lesion regions and adaptively fuse corresponding lesion features for comprehensive DR grading. Moreover, clinical experts [27] are also invited to grade on the Messidor-1 dataset. It is worth mentioning that our method outperforms the experts by 4.7% and 4.1% on the AUC of referral and normal settings, respectively.

As shown in Table 2, we also experiment on the EyePACS dataset to test the advancements of our proposed approach. The fundus images collected from this dataset are captured by different types of cameras, so the quality of im-

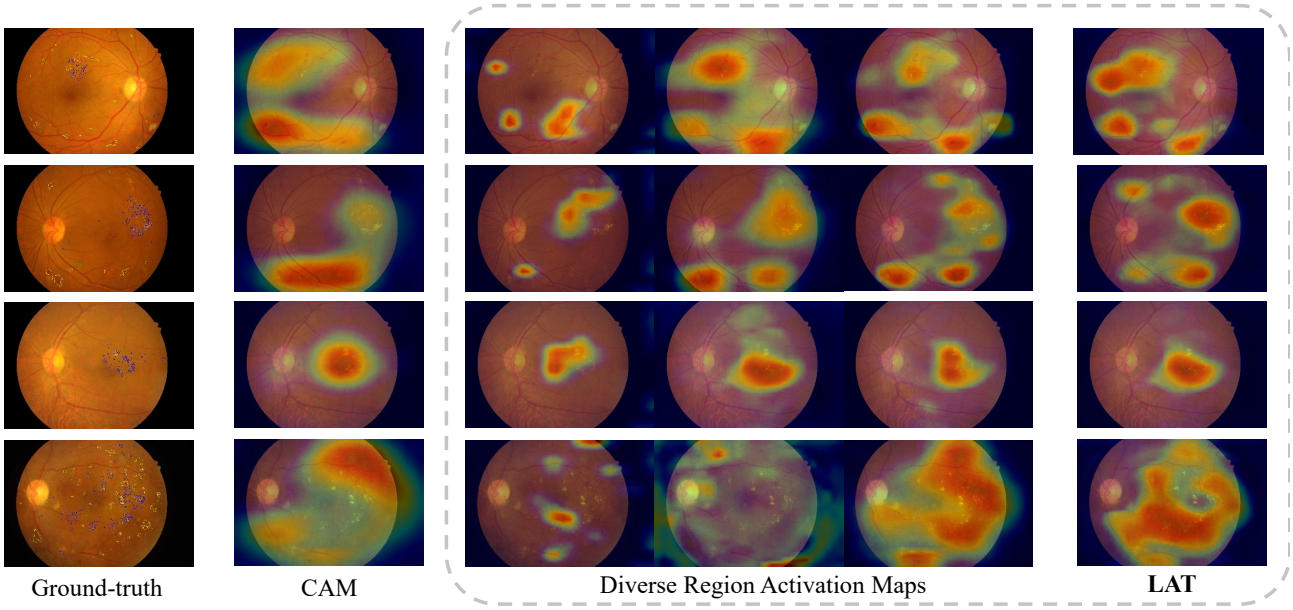


Figure 3: Visualization comparison with CAM [45]. Our method can identify diverse lesion regions through different filters, and adaptively fuse these regions to produce a more complete and accurate activation map. The ground-truth contains microaneurysms, haemorrhages, soft exudates and hard exudates, annotated with green, yellow, green and blue respectively.

Table 2: Performance comparisons of DR grading with state-of-the-art methods on the EyePACS dataset.

Methods	Val set	Test set
	Kappa	Kappa
Min-pooling [1]	0.860	0.849
o_O	0.854	0.844
Reformed Gamblers	0.851	0.839
Zoom-in-Net [34]	0.865	0.854
AFN [23]	0.871	0.859
Semi+Adv [46]	-	0.872
LAT (ours)	**0.893**	**0.884**

ages is relatively low which contains some noises like under/overexposure and out-of-focus problem. Among them, Kappa values of the top three places from the Kaggle challenge [1] are shown, where the top-1 place can achieve 84.9% Kappa score. And our LAT reports 89.3% and 88.4% Kappa on the validation set and the testing set respectively, and sets a new state-of-the-art performance. Besides, LAT obtains 1.2% performance gain over the recent Semi+Adv [46] on the validation set. The results show that our method can adapt to pixel appearance variations by using a self-attention mechanism. In other words, the pixels of the lesion region with similar appearance can be gathered, and the noisy background pixels caused by overexposure or underexposure can be suppressed.

Lesion Discovery Performance. To demonstrate the power of LAT for lesion discovery, we compare it with CAM [45], and its backbone is also set as ResNet50 [19] for a fair comparison. As illustrating in Figure 3, we visualize the different region activation maps generated from different lesion filters for qualitative evaluation. We can observe that different lesion filters can successfully identify multiple regions such as microaneurysms, haemorrhages and exudates. This is because that we explicitly model a set of lesion filters and learn them based on the encoder-decoder structure to focus on different lesions, which greatly improves the performance of lesion discovery. Unlike our method, CAM [45] and Zoom-in-Net [34] do not constrain the obtained attention map for more lesion information, so the network usually only focuses on the most important lesion region for DR grading. Please refer to the Supplementary Material for more visualization results and analysis.

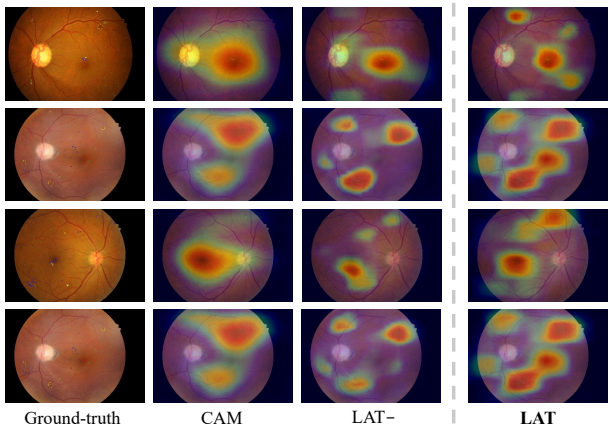


Figure 4: The qualitative comparisons for lesion discovery.

Table 3: Evaluation of the effectiveness of different components on the Messidor-2 dataset.

Index	$\mathcal{P}$	$\mathcal{S}$	$\mathcal{C}$	$\mathcal{D}$	$\mathcal{G}$	AUC	Kappa
1	✗	✗	✗	✗	✗	0.941	0.785
2	✓	✗	✗	✗	✗	0.948	0.797
3	✗	✓	✓	✗	✗	0.959	0.821
4	✗	✗	✓	✗	✗	0.952	0.813
5	✓	✓	✓	✗	✗	0.959	0.839
6	✓	✓	✓	✓	✗	0.971	0.842
7	✓	✓	✓	✓	✓	0.979	0.851

4.4. Ablation Studies

To look deeper into our method, we perform a series of ablation studies using ResNet50 as the backbone on the Messidor-2 dataset. Results and analysis are as follows.

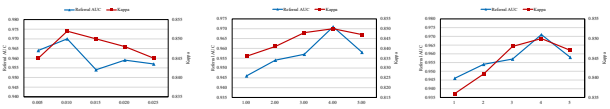
Effectiveness of the lesion region diversity mechanism for lesion discovery. To evaluate the improvement for the region diversity mechanism, we compare two baselines with our final proposed model, including CAM [45] and our method without diversity mechanism (LAT-). Figure 4 illustrates qualitative comparisons for lesion discovery. We can observe that CAM only focuses on the most important lesion region, and LAT- performs better than CAM. This is because the self-attention mechanism in the decoder can incorporate contextual information from other lesion filters to increase their discrepancies. Our method can achieve the best performance for lesion discovery, because the diversity mechanism can further explicitly constrain the lesion features and enable the different filters to find their corresponding regions containing more explicit lesion semantics.

Next, we analyze the effectiveness of each component of our LAT for DR grading, including the pixel relation based encoder ($\mathcal{P}$), the self-attention layer ($\mathcal{S}$) and cross-attention layer ($\mathcal{C}$) of the lesion filter based decoder, the region diversity mechanism ($\mathcal{D}$), and the global consistency loss ($\mathcal{G}$).

Effectiveness of the pixel relation based encoder. In index-2, we only add the self-attention based on the pixel relation based encoder. Compared with the baseline model, the performances is improved by 1.2% in Kappa score. This is because the self-attention mechanism of the encoder can model the correlation of pixels and generate more robust features to adapt to pixel appearance variations.

Effectiveness of the lesion filter based decoder. Based on index-1 and index-3, when the lesion filter based decoder is added, the performance is improved by 1.8% in AUC and 3.6% in Kappa. This demonstrates the importance of learning lesion filters. Based on index-3 and index-4, when the self-attention layer in decoder is added, the performance is further improved. This demonstrates the self-attention can further incorporate context information from other filters to increase their discrepancies.

Effectiveness of the region diversity mechanism. Based on index-5 and index-6, the performance is improved by 1.2% in AUC, which shows the necessity of our diversity mechanism. By adding the triplet loss, the learned lesion filters can be guided to discover diverse lesion regions, and the corresponding features extracted from lesion regions can be adaptively fused together to improve DR grading.



(a) λ_{gcl} (b) λ_{tri} (c) K

Figure 5: Evaluation of the hyperparameters λ_{gcl}, λ_{tri}, and the number of lesion filters K.

Effectiveness of the global consistency loss. Compared with index-6 and index-7, when the global consistency loss is added, the performance is further improved by 0.9% in Kappa. This shows that this loss can alleviate problems such as Grade 0 and 1 that are difficult to distinguish by pushing class-specific local centers approaching their corresponding global centers.

Hyperparameter evaluations. We evaluate how λ_{tri} and λ_{gcl} affect our model learning. Here, λ_{tri} controls the relative importance of the region diversity mechanism, and λ_{gcl} controls the relative importance of the global consistency loss. As shown in Figure 5, our model achieves much better performance when $\lambda_{tri} = 0.04$, $\lambda_{gcl} = 0.01$. We also evaluate the influence of different lesion filters in Figure 5. We conduct experiments to explore the effect of different numbers of lesion filters. The performance continues to grow until $K = 4$, which means that 4 filters are enough to identify lesion regions.

5. Conclusion

In this paper, we propose a novel lesion-aware transformer to achieve DR grading and lesion discovery jointly via an encoder-decoder structure. Specifically, the pixel relation based encoder can effectively adapt to pixel appearance variations. And the lesion filter based decoder is designed to learn lesion-aware filters for lesion discovery. Extensive results on three challenging benchmarks demonstrate that our LAT performs favorably against state-of-the-art other approaches.

6. Acknowledgment

This work was partially supported by the National Key Research and Development Program under Grant No. 2017YFC0820600, Strategic Priority Research Program of Chinese Academy of Sciences (No.XDC02050500), National Nature Science Foundation of China (Grant 62022078, 62021001, 62071122), Open Project Program of the National Laboratory of Pattern Recognition (NLPR) under Grant 202000019, and Youth Innovation Promotion Association CAS 2018166.

References

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

[2] Michael D Abràmoff, James C Folk, Dennis P Han, Jonathan D Walker, David F Williams, Stephen R Russell, Pascale Massin, Beatrice Cochener, Philippe Gain, Li Tang, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA ophthalmology*, 131(3):351–357, 2013.

[3] Balint Antal and Andras Hajdu. An ensemble-based system for microaneurysm detection and diabetic retinopathy grading. *IEEE transactions on biomedical engineering*, 59(6):1720–1726, 2012.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020.

[5] William Chan, Chitwan Saharia, Geoffrey Hinton, Mohammad Norouzi, and Navdeep Jaitly. Imputer: Sequence modelling via imputation and dynamic programming. *arXiv preprint arXiv:2002.08926*, 2020.

[6] NH1 Cho, JE Shaw, Suvi Karuranga, Yafang Huang, JD da Rocha Fernandes, AW Ohlrogge, and B Malanda. Idf diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes research and clinical practice*, 138:271–281, 2018.

[7] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *CVPR*, pages 2219–2228, 2019.

[8] Piotr Chudzik, Somshubra Majumdar, Francesco Caliva, Bashir Al-Diri, and Andrew Hunter. Exudate segmentation using fully convolutional neural networks and inception modules. In *Medical Imaging 2018: Image Processing*, volume 10574, page 1057430. International Society for Optics and Photonics, 2018.

[9] Jorge Cuadros and George Bresnick. Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening. *Journal of diabetes science and technology*, 3(3):509–516, 2009.

[10] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Noushin Eftekhari, Hamid-Reza Pourreza, Mojtaba Masoudi, Kamaledin Ghiasi-Shirazi, and Ehsan Saeedi. Microaneurysm detection in fundus images using a two-step convolutional neural network. *Biomedical engineering online*, 18(1):67, 2019.

[13] Rishab Gargeya and Theodore Leng. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7):962–969, 2017.

[14] Rishab Gargeya and Theodore Leng. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7):962–969, 2017.

[15] Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*, 2019.

[16] Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. In *ICLR*, 2018.

[17] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge 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.

[18] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge 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.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[20] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[21] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S Corrado, Lily Peng, and Dale R Webster. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125(8):1264–1272, 2018.

[22] Xiaomeng Li, Xiaowei Hu, Lequan Yu, Lei Zhu, Chi-Wing Fu, and Pheng-Ann Heng. Canet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. *IEEE transactions on medical imaging*, 39(5):1483–1493, 2019.

[23] Zhiwen Lin, Ruoqian Guo, Yanjie Wang, Bian Wu, Tingting Chen, Wenzhe Wang, Danny Z Chen, and Jian Wu. A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion. In *MICCAI*, pages 74–82. Springer, 2018.

[24] Weizeng Lu, Xi Jia, Weicheng Xie, Linlin Shen, Yicong Zhou, and Jinming Duan. Geometry constrained weakly supervised object localization. *arXiv preprint arXiv:2007.09727*, 2020.

[25] Harry Pratt, Frans Coenen, Deborah M Broadbent, Simon P Harding, and Yalin Zheng. Convolutional neural networks for diabetic retinopathy. *Procedia Computer Science*, 90:200–205, 2016.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[27] Clara I Sánchez, Meindert Niemeijer, Alina V Dumitrescu, Maria SA Suttorp-Schulten, Michael D Abramoff, and Bram 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.

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *CVPR*, pages 618–626, 2017.

[29] Ruchir Srivastava, Lixin Duan, Damon WK Wong, Jiang Liu, and Tien Yin Wong. Detecting retinal microaneurysms and hemorrhages with robustness to the presence of blood vessels. *Computer methods and programs in biomedicine*, 138:83–91, 2017.

[30] Mark JJP Van Grinsven, Bram van Ginneken, Carel B Hoyng, Thomas Theelen, and Clara 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.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

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

[33] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.

[34] Zhe Wang, Yanxin Yin, Jianping Shi, Wei Fang, Hongsheng Li, and Xiaogang Wang. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection. In *MICCAI*, pages 267–275. Springer, 2017.

[35] CP Wilkinson, Frederick L Ferris III, Ronald E Klein, Paul P Lee, Carl David Agardh, Matthew Davis, Diana Dills, Anselm Kampik, R Pararajasegaram, Juan T Verdaguer, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–1682, 2003.

[36] Sungmin Woo, Chong Hyun Suh, Sang Youn Kim, Jeong Yeon Cho, and Seung Hyup Kim. Diagnostic performance of prostate imaging reporting and data system version 2 for detection of prostate cancer: a systematic review and diagnostic meta-analysis. *European urology*, 72(2):177–188, 2017.

[37] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *ICCV*, pages 6589–6598, 2019.

[38] Zizheng Yan, Xiaoguang Han, Changmiao Wang, Yuda Qiu, Zixiang Xiong, and Shuguang Cui. Learning mutually local-global u-nets for high-resolution retinal lesion segmentation in fundus images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 597–600. IEEE, 2019.

[39] Seunghan Yang, Yoonhyung Kim, Youngeun Kim, and Changick Kim. Combinational class activation maps for weakly supervised object localization. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2941–2949, 2020.

[40] Yehui Yang, Tao Li, Wensi Li, Haishan Wu, Wei Fan, and Wensheng Zhang. Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. In *MICCAI*, pages 533–540. Springer, 2017.

[41] Yehui Yang, Fangxin Shang, Binghong Wu, Dalu Yang, Lei Wang, Yanwu Xu, Wensheng Zhang, and Tianzhu Zhang. Robust collaborative learning of patch-level and image-level annotations for diabetic retinopathy grading from fundus image. *arXiv preprint arXiv:2008.00610*, 2020.

[42] Chen-Lin Zhang, Yun-Hao Cao, and Jianxin Wu. Rethinking the route towards weakly supervised object localization. In *CVPR*, pages 13460–13469, 2020.

[43] Dong Zhang, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, and Qianru Sun. Feature pyramid transformer. *arXiv preprint arXiv:2007.09451*, 2020.

[44] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*, pages 172–188, 2018.

[45] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.

[46] Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, and Ling Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. In *CVPR*, pages 2079–2088, 2019.

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.