Joint Learning of Multi-level Tasks for Diabetic Retinopathy Grading

Xiaofei Wang, Mai Xu, Jicong Zhang, Lai Jiang, Liu Li, Ningli Wang, Hanruo Liu, and Zulin Wang

***Abstract*—Diabetic retinopathy (DR) is a leading cause of permanent blindness among the working-age people. Automatic DR grading can help ophthalmologists make timely treatment for patients. However, the existing grading methods are usually trained with high resolution (HR) fundus images, such that the grading performance decreases a lot given low resolution (LR) images, which are common in clinic. In this paper, we mainly focus on DR grading with LR fundus images. According to our analysis on the DR task, we find that: 1) image super-resolution (ISR) can boost the performance of both DR grading and lesion segmentation; 2) the lesion segmentation regions of fundus images are highly consistent with pathological regions for DR grading. Based on our findings, we propose a convolutional neural network (CNN)-based method for joint learning of multi-level tasks for DR grading, called DeepMT-DR, which can simultaneously handle the low-level task of ISR, the mid-level task of lesion segmentation and the high-level task of disease severity classification on LR fundus images. Moreover, a novel task-aware loss is developed to encourage ISR to focus on the pathological regions for its subsequent tasks: lesion segmentation and DR grading. Extensive experimental results show that our DeepMT-DR method significantly outperforms other state-of-the-art methods for DR grading over three datasets. In addition, our method achieves comparable performance in two auxiliary tasks of ISR and lesion segmentation.**

***Index Terms*—Multi-task Learning, Retinal Fundus Images, Diabetic Retinopathy, Deep Neural Networks.**

I. Introduction

Diabetic retinopathy (DR) is one of the most feared complications of microvascular retinal changes triggered by diabetes, which is also the leading cause of vision loss in working-age adults (20 to 65 years) worldwide [46]. The incidence of serious DR is around 127 million all over the world in 2012 and is predicted to grow to 191 million by 2030 [46]. In fact, most vision loss caused by DR can be avoided through early detection and treatment [3]. Thus, it is important to detect DR in an early stage. However, due to the leaking resources of ophthalmologists and professional screening equipments, it is hard to conduct manual DR screening over all suspected patients. It is therefore essential to develop an automatic DR detection system.

In the past few years, deep neural networks (DNNs) have been widely used as a powerful tool to learn features for automatic DR severity grading [1], [44], [47]. For example, Adly *et al.* [1] proposed a binary-tree-based multiclass classifier of DNNs for DR grading, while Zhou *et al.* [47] designed a multi-cell structure for detecting the early stage of DR, in which the depth of DNN is increased along with the scale growth of input images. In addition to using only image-level supervision, some works also utilized the lesion location information to boost the DR grading performance [21], [48]. For instance, Lin *et al.* [21] presented an attention fusion network (AFN) that fuses the extracted lesion maps and original fundus images for identifying different DR grades. However, the existing DR grading methods are trained by high resolution (HR) fundus images, and their grading performance significantly decreases given low resolution (LR) images [47], as shown by the missed diagnosed examples of LR fundus images in Figure 1.

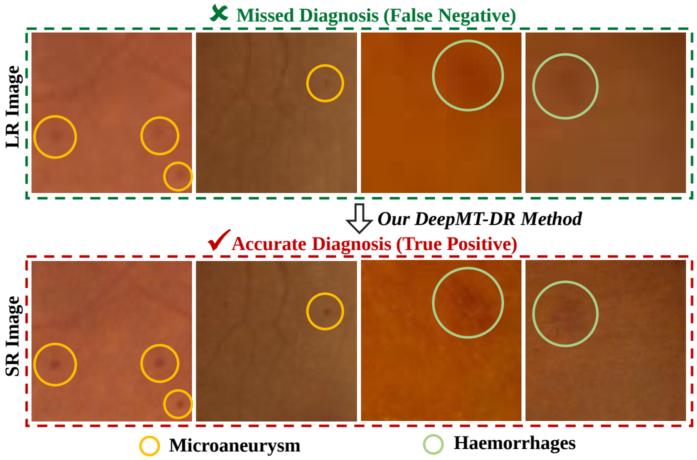

✘ Missed Diagnosis (False Negative)
LR Image
Our DeepMT-DR Method
✓ Accurate Diagnosis (True Positive)
SR Image
○ Microaneurysm
○ Haemorrhages

Fig. 1. Examples of false negative DR grading for the real-world LR fundus images (the first row) and true positive DR grading for the corresponding super-resolved (SR) images generated by our DeepMT-DR method (the second row). In these examples, we show the lesion portions of both LR and SR images. The lesions of LR images are inconspicuous and easily missed-detected by most DR grading methods, inevitably leading to the missed diagnosis, especially for the cases in the early stages. In contrast, the lesions in the SR images are more remarkable, thus leading to accurate diagnosis.

In clinic, there are lots of LR fundus images [25], mainly due to the following reasons. 1) Many LR retinal imaging devices are currently used in clinic [45], [26], e.g., Topcon TRC-NW5S/TRV-50/TRC-50IA and Canon CR6–45NM, the resolutions of which are around 550×550. 2) The acquired fundus images are downsampled to LR before transmitting to the medical systems (e.g., PACS [2]), due to the limited storage capacity and high cost of data migration [2]. Besides, DR diagnosis in clinic highly depends on the detected retinal pathologies, such as microaneurysms, which are usually rather small and inconspicuous on LR fundus images. This hinders the early detection of the disease. Therefore, it is intuitive to apply image super-resolution (ISR) and lesion segmentation to boost the performance of DR grading on LR images.

In this paper, we propose a novel DR grading method, which simultaneously handles the auxiliary tasks of ISR and lesion segmentation on fundus images. Firstly, we thoroughly analyze

X. Wang, M. Xu, J. Zhang, L. Jiang and Z. Wang are with the Beihang University, Beijing 100191 China; L. Li is with the Imperial College London, London SW72BX UK; N. Wang and H. Liu are with the Beijing Institute of Ophthalmology, Beijing 100730 China; M. Xu is the corresponding author of this paper (E-mail: Maixu@buaa.edu.cn).

the correlation among the tasks of ISR, lesion segmentation and DR grading, demonstrating the potential gain of multi-task learning (MTL). Then, based on our findings, we propose a hierarchical deep multi-task learning based DR grading method, called DeepMT-DR, the framework of which is briefly shown in Figure 2. Our DeepMT-DR consists of three tasks, i.e., the ISR, lesion segmentation and DR grading, regarded as the low-, mid- and high-level vision tasks [43], [28], respectively. These three tasks are processed in a hierarchy by our DeepMT-DR framework, consistent with our task correlation analysis. In addition to the forward propagation across the tasks, it is also important to guide the ISR with the downstream tasks, i.e., lesion segmentation and DR grading. For this purpose, we propose a novel task-aware loss, which makes the ISR process focus on the pathological regions for both lesion segmentation and DR grading.

To the best of our knowledge, our work is the first attempt of jointly learning the multiple medical tasks at low-, mid- and high-levels. The main contributions of this paper are as follows. (1) We thoroughly analyze the correlation among the tasks of ISR, lesion segmentation and DR grading, indicating that these three tasks can benefit from each other. (2) We propose a deep multi-task learning method, i.e., DeepMT-DR, for the main task of DR grading and the auxiliary tasks of both ISR and lesion segmentation. (3) We verify through extensive experiments that our method advances the state-of-the-art (SOTA) in DR grading, and also achieves comparable performance in the ISR and lesion segmentation.

This paper is an extended version of our conference paper [42] with substantial improvements as follows. (1) In addition to the review on MTL in the natural image domain, this journal paper adds a thorough literature review on MTL in the medical image domain. (2) In this journal paper, new findings are investigated to verify the necessity of joint learning paradigm. Besides, we also thoroughly analyze our findings with additional network structures and over more datasets. (3) The architecture of DeepMT-DR in this journal paper is significantly advanced as follows. We first propose a novel multi-scale feature integration structure and add it to the lesion segmentation subnet. We further design a novel lesion-oriented feature selection process to optimize the proposed GMSV algorithm for information feedback in DeepMT-DR. Consequently, the performance of our method has been improved in all 3 tasks. (4) We comprehensively validate the model effectiveness through more SOTA methods and experimental settings. More importantly, we further verify the significance and real-world impact of our method, by conducting additional experiments over a real-world LR dataset.

II. Related work

A. Medical Image Analysis for Fundus Images

In the field of medical image analysis, automatic disease grading, lesion segmentation and ISR are three cross-level tasks to advance the computer-aided diagnosis, especially for DR in retinal fundus images [1], [18], [24].

DR Grading. In recent years, several methods have been developed for DR grading based on optical fundus images, which can be divided into 2 categories: heuristic methods [29] and deep learning methods [4], [40], [44], [1]. The heuristic methods extract the hand-designed features with medical DR priors on DR, e.g., the optic disk, blood vessels and the presence of retinal lesions [29]. However, these methods rely heavily on the experience of human experts, inevitably leading to low efficiency and poor generalization ability [29]. On the other hand, many DNNs [4], [40], [44], [1] have been proposed for DR severity grading. For example, Bravo *et al.* [4] proposed simply combining the VGG-16 and Inception-4 networks, while Vo *et al.* [40] proposed a novel structure that integrates the kernels with multiple losses (CKML) to learn the features for DR grading. Wang *et al.* [44] presented a gated attention model that utilizes the fundus images in various modalities for DR grading. In addition to improving the grading accuracy, the recent works [47], [49] also focused on the practical applications of DR grading. For instance, Zhou *et al.* proposed an image synthesis method for DR grading, in order to solve the problem of data paucity [49]. In clinic, many fundus images are captured or stored in LR; however, none of the existing works has applied ISR for precise DR grading on LR fundus images.

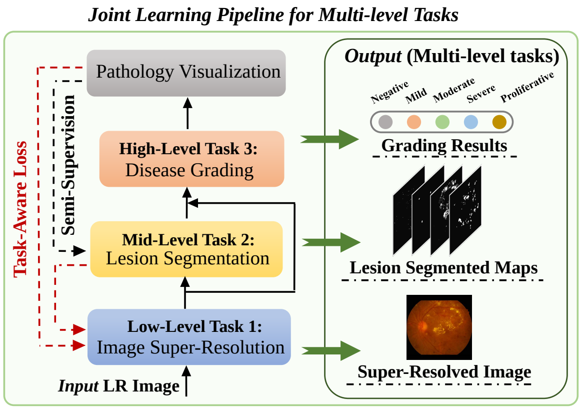
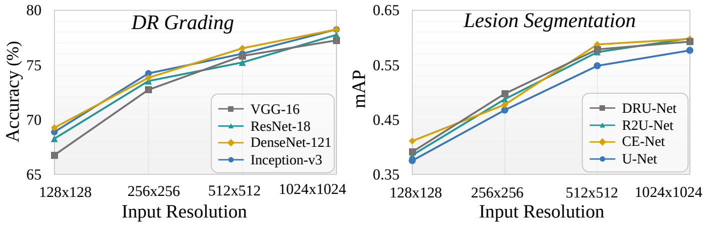

Joint Learning Pipeline for Multi-level Tasks
Task-Aware Loss
Semi-Supervision
Pathology Visualization
High-Level Task 3: Disease Grading
Mid-Level Task 2: Lesion Segmentation
Low-Level Task 1: Image Super-Resolution
Input LR Image
Output (Multi-level tasks)
Negative
Mild
Moderate
Severe
Proliferative
Grading Results
Lesion Segmented Maps
Super-Resolved Image

Fig. 2. Brief framework of the proposed DeepMT-DR method.

Lesion Segmentation. Lesion segmentation is an important task to provide pathological guidance to the diagnosis process, especially for detecting retinal lesions in fundus images. Most recently, taking fundus images as inputs, several fully trainable U-shaped structures [7], [18] have been proposed to generate lesion segmentation with corresponding probability distribution. Specifically, Feng *et al.* [7] proposed a U-Net based structure with short and long skip connections for exudates segmentation. Kou *et al.* designed a recurrent residual U-Net structure for segmenting microaneurysms [18]. However, only a few works [48] have associated lesion segmentation with other medical tasks.

Image Super-Resolution. Recently, there have been several medical ISR methods used to improve the quality of indistinct pathological areas, especially for ocular diseases based on fundus images [25], [31]. The existing ISR methods are mainly on the top of generative adversarial network (GAN)-based [19], [25] or CNN-based [31] structures. Specifically, Mahapatra *et al.* [25] proposed a saliency-based GAN network for ISR in retinal fundus images, while Ren *et al.* [31] designed a multi-scale ISR structure with deep residual connections.

Unfortunately, none of the previous works has studied the relationship of the above three tasks. In this paper, we thoroughly analyze the correlation of these multi-level tasks and propose a



(a) (b)

Fig. 3. Task correlation analysis. (a) Accuracy of DR grading *vs.* input resolution ranging from 128 × 128 to 1024 × 1024. (b) Mean average precision (mAP) of retinal lesion segmentation *vs.* varying input resolutions.

unified framework to jointly learn these tasks.

B. Multi-Task Learning in Deep Neural Networks

Multi-task learning is a learning paradigm in machine learning, to leverage the shared information in related tasks. For natural images, the effectiveness of DNN-based MTL has been verified in many vision tasks, such as ISR [30], instance segmentation [12] and depth estimation [23]. The existing MTL frameworks can be divided into two categories: hard parameter sharing [12], [30] and soft parameter sharing [23], [27]. Specifically, hard parameter sharing refers to sharing the hidden layers of all tasks, while keeping several task-specific output layers independent. For example, He *et al.* [12] proposed the mask R-CNN structure to simultaneously detect objects and output the segmentation mask for each instance. For soft parameter sharing, each task has its own structure and parameters, the distance of which is encouraged to be similar through certain regularization. For example, Liu *et al.* [23] designed the multi-task attention network (MTAN) with soft-attention modules for the joint tasks of semantic segmentation and depth estimation, while Misra *et al.* [27] proposed a novel cross-stitch sharing unit to combine the activations from multiple tasks. However, few methods [17] have considered the information sharing across multi-level tasks, particularly the information flow from the high- to low-levels.

Meanwhile, for medical images, a few MTL methods have also been proposed to simultaneously improve the performance of the main task or the auxiliary tasks [22], [21], [9]. For example, Liu *et al.* [22] proposed a novel MTL network with margin ranking loss (MTMR-Net) to jointly process the tasks of lung nodule classification and attribute score regression. Similarly, Gao *et al.* [9] designed a censored regression loss to jointly predict the lung cancer and cancer-free progression from censored heterogenous clinical images. Nevertheless, most existing methods are designed for the high-level tasks, i.e., image classification or regression. Different from the existing methods, we propose a deep multi-task learning based DR grading (DeepMT-DR) method, which is a first attempt to jointly process the multiple medical tasks at low-, mid- and high-levels. Specifically, we design a novel hierachical structure to combine the three tasks according to our findings, instead of simply sharing the parameters. Besides, a novel task-aware loss is proposed to encourage the ISR to focus on the pathological regions for its subsequent tasks.

III. Task Correlation Analysis

In this section, we thoroughly analyze the correlation among the tasks of disease grading, lesion segmentation and ISR for DR problem, through which we demonstrate the potential gain of multi-task learning. According to the analysis, three findings are investigated as follows.

Finding 1: The performance of DR grading and lesion segmentation can be improved, when the input fundus images are super-resolved.

Analysis: Figure 3 (a) shows the results of DR grading by various algorithms [35], [16], [14], [15] at different input resolutions over a large scale DR image dataset, i.e., the DDR [20] dataset. Specifically, the fundus images are downsampled to 1024×1024 for unifying the size of input images. Then, these 1024×1024 images are downsampled by 2, 4 and 8 scales. Consequently, each fundus image has four resolutions varying from 128×128 to 1024×1024, as the input to DR grading. Note that for fair comparison of DR grading at different resolutions, we use spatial pyramid pooling (SPP) [13] before the dense-connection layers in the algorithms. As can be seen in the Figure 3 (a), the grading accuracy obviously improves along with the increase of input resolution. For example, the accuracy can be improved from 69.3% to 78.3% using DenseNet-121 [15]. This indicates the potential improvement of DR grading after applying ISR to fundus images.

Similarly, we also segment lesions in fundus images at varying resolutions. For fair comparison, the LR images are upsampled to 1024 × 1024 by the bicubic SR algorithm, and then supervised with the HR (i.e., 1024 × 1024) segmentation labels. As can be seen in Figure 3 (b), the segmentation performance improves along with increased input resolution. This implies the lesion segmentation task can benefit from the ISR task. Finally, the analysis of this finding is completed.

Finding 2: Joint training of ISR and DR grading performs better than independent training of these two tasks. Similar results can be found for of ISR and lesion segmentation.

Analysis: We conduct the experiment to analyze the necessity of jointly training ISR and its subsequent tasks, i.e., DR grading and lesion segmentation. Specifically, the ISR and DR grading tasks are trained in joint and independent manners, respectively. Note that we use the Mahapatra *et al.* [25] and the Resnet-18 [14] as the DNNs for ISR and DR grading. Besides, we use the bicubic interpolation as a simple non-learning algorithm for ISR, as the baseline. Besides, the upscale factors are 2, 4 and 8 for the images with resolution of 512 × 512, 256 × 256 and 128 × 128, respectively. The results of both joint and independent training are shown in Figure 4 (a). This figure shows that the DR grading accuracy in the setting of joint training performs better than that of independent training. Similarly, as shown in Figure 4 (b), joint training of ISR and lesion segmentation (with the Mahapatra *et al.* [25] and U-Net [32] as the DNNs for ISR and lesion segmentation) also outperforms independent training of these two tasks.

Finding 3: The lesion segmentation regions of fundus images are highly consistent with pathological regions for DR grading, indicating that the segmented lesions are more important than the background for DR grading.

Analysis: Here, we further study the correlation between lesion segmentation and DR grading. To be specific, we apply the commonly used network visualization algorithm, i.e., the Grad-CAM [34], to generate the evidence map of the final decision from the DR grading network, i.e., Resnet-18. Figure 5

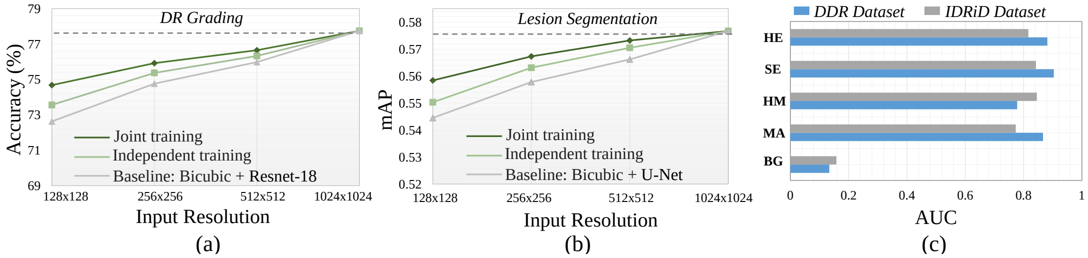


Fig. 4. Task correlation analysis. (a) Accuracy of DR grading on the settings of joint training and independent training of ISR and DR grading network. Note that the input images are all upscaled to 1024 × 1024. (b) mAP of retinal lesion segmentation on the settings of joint training and independent training of ISR and lesion segmentation network. (c) AUC of DR grading network visualization maps in the background (BG) and lesion segmented regions of MA, HM, SE and HE.

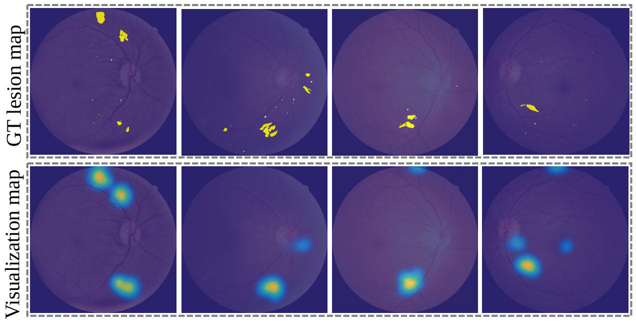


Fig. 5. Ground-truth lesion segmentation maps and their corresponding visualization results of the DR grading network, i.e., Resnet-18.

shows some examples of the ground-truth lesion segmentation maps and their corresponding visualization results of the DR grading network. We then calculate the AUC values between the evidence map and different lesion segmented regions (including microaneurysms (MA), haemorrhages (HM), soft exudates (SE) and hard exudates (HE)) as well as background (BG) regions. Figure 4 (c) shows that the AUC results of lesion segmented regions are significantly higher than those of background over both IDRiD and DDR datasets, e.g., the AUC result of the lesion regions is 0.87 with the comparison of 0.13 for the background over the DDR dataset. This means that the lesion segmented regions are more important than the background for DR grading.

IV. Methodology

A. Structure of DeepMT-DR

According to the above analysis, the tasks of ISR, lesion segmentation and DR grading are closely related with each other. Therefore, it is intuitive to jointly learn these three multi-level tasks in a unified framework. In this paper, we propose a novel DeepMT-DR framework for the high-level task of DR grading, simultaneously handling the low-level task of ISR and the middle-level task of lesion segmentation. The framework of DeepMT-DR is shown in Figure 6. As can be seen, given an LR fundus image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ (with height H, width W and 3 channels of RGB) as the input, our DeepMT-DR model can output the DR severity grade $\tilde{l} \in \mathbb{R}^5$, the super-resolved image $\tilde{\mathbf{Y}}$[1] $\in \mathbb{R}^{4H\times 4W\times 3}$ and the segmented maps $\tilde{\mathbf{S}} \in \mathbb{R}^{4H\times 4W\times 4}$ of 4 lesion types: MA, HE, SE and EX. Note that in our 5-class DR grading task, grade 0 to 4 refer to negative, mild, moderate, severe and proliferative DR, respectively.

The structure of DeepMT-DR consists of 3 subnets, corresponding to the multi-level tasks of the ISR, lesion segmentation and DR grading. Detailed structures are described as follows.

ISR Subnet: As shown in Figure 6, the ISR subnet consists of several cascaded components, i.e., 5 feature extraction layers and 2 up-scaling layers. Specifically, the feature extraction layers are developed to extract pathological information from the LR fundus images. Then, for generating the super-resolved fundus images, the output of the feature extraction layers is processed by 2 up-scaling layers with the upscale factor of 2 in each layer. The detailed structure of the ISR subnet is illustrated in Figure 1 of the supplementary material.

Lesion Segmentation Subnet: Given the output super-resolved image $\tilde{\mathbf{Y}}$ from the ISR subnet, the lesion segmentation subnet is developed to segment the retinal lesions corresponding to DR. The schematic diagram of the lesion segmentation subnet is shown in the orange part of Figure 6. As seen in this figure, a U-shaped structure, consisting of 5 down-transition (DT) layers and 4 up-transition (UT) layers, is designed to extract the features for precisely localizing the lesions. Specifically, as the input, the super-resolved image $\tilde{\mathbf{Y}}$ is progressively contracted and down-sampled through 5 DT layers. Similarly, the contracted features are progressively expanded and up-sampled through 4 UT layers. Subsequently, the outputs of the last DT layer and each UT layer are further processed by a convolutional layer to generate the intermediate segmentation maps at different scales. Assuming that $\hat{\mathbf{S}}_i$ is the segmentation map at the i-th scale, the final segmentation lesion $\hat{\mathbf{S}}$ is calculated as follows:

$$\hat{\mathbf{S}} = \mathrm{softmax}\Big(\sum_{i=1}^{5} \gamma_i \cdot \mathrm{UP}(\hat{S}_i, 2^{5-i})\Big). \quad (1)$$

In the above equation, $\{\gamma_i\}_{i=1}^{5}$ are the hyper-parameters to balance the segmentation maps, and $\mathrm{UP}(\cdot, t)$ is the t-time upscale operation. The detailed structure of the lesion segmentation subnet can be found in Figure 2 of the supplementary material.

DR Grading Subnet: Taking the super-resolved image $\tilde{\mathbf{Y}}$ and the lesion segmented map $\tilde{\mathbf{S}}$ as the inputs, the DR grading subnet is developed on the top of ResNet-18 [14], to grade the DR severity. The structure of the DR grading subnet is illustrated in the blue part of Figure 6. Specifically, a spatial pyramid pooling (SPP) [13] layer is added before the fully connected (FC) layer, in order to consider varied lesion sizes and adapt different input resolutions. In addition to the grading result, the visualization map for the grading network is also generated by our GMSV algorithm. This visualization map is further used as the information flow of feedback from the high-level tasks to the low-level task. The details about the visualization map and the information feedback are described in Section IV-B.

B. GMSV Algorithm for Information Feedback

In addition to the hierarchical structure, we also propose an information flow of feedback from the high-level tasks to the

[1] Note that the upscale factor for ISR is set to 4 in this paper, and it can be easily extended to other values in practice.

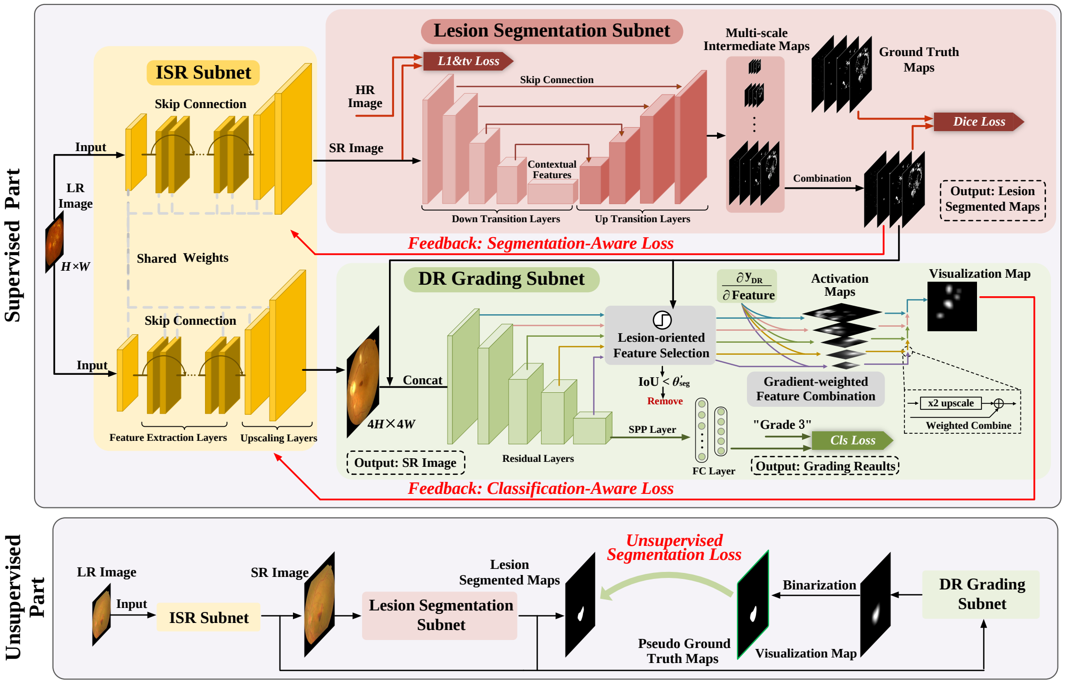


Fig. 6. Framework of the proposed DeepMT-DR for DR grading with multiple tasks. Specifically, the DeepMT-DR consists of three subnets, i.e., the ISR, lesion segmentation and DR grading subnets, for low-, mid- and high-level vision tasks, respectively.

low-level task. Specifically, for obtaining the pathological regions of DR grading, we design a GMSV algorithm to generate the fine-grained evidence map of the final grading decision. Specifically, GMSV is composed of 2 sequential processes, i.e., the lesion-oriented feature selection and gradient-weighted feature combination, for generating the evidence maps of DR grading.

(1) Lesion-oriented feature selection. Recently, it has been observed [41] that the CNN features extracted from fundus images are able to locate the class-relevant objects, such as lesions. Thus, the extracted features with important DR pathologies are firstly selected by our GMSV algorithm, according to the intersection over union (IOU) value between the feature maps and lesion segmentation map $\tilde{\mathbf{S}}$. Let $\mathbf{F}^{i,j} \in \mathbb{R}^{M^j \times N^j}$ denote the i-th feature at the j-th layer, with height M^j and width N^j. The selected set $\mathbf{U}^j$ can be represented as

$$\mathbf{U}^j = \{\mathbf{F}^{i,j} \mid \mathrm{IoU}^{i,j} > \theta_{\text{seg}} \cdot \max_r(\mathrm{IoU}^{r,j})\},$$
$$\text{where } \mathrm{IoU}^{i,j} = \frac{\mathbb{E}\big[(\mathrm{UP}(\mathbf{F}^{i,j}; \dim(\tilde{\mathbf{S}})) > \xi^{i,j}) \cap \tilde{\mathbf{S}}\big]}{\mathbb{E}\big[(\mathrm{UP}(\mathbf{F}^{i,j}; \dim(\tilde{\mathbf{S}})) > \xi^{i,j}) \cup \tilde{\mathbf{S}}\big]}. \quad (2)$$

In (2), $\cap$ and $\cup$ denote the intersection and union operations; $\mathrm{UP}(\mathbf{F}^{i,j}; \dim(\tilde{\mathbf{S}}))$ denotes upsampling $\mathbf{F}^{i,j}$ to the resolution of $\tilde{\mathbf{S}}$; θ_{seg} denotes a threshold of the IoU value; $\xi^{i,j}$ is the learnable parameter to produce the binary mask from the upsampled map.

(2) Gradient-weighted feature combination. Given the selected set $\mathbf{U}^j$, the global-average-pooling (GAP) is applied to obtain the importance weight $w^{i,j}$ for $\mathbf{F}^{i,j}$ in $\mathbf{U}^j$:

$$w^{i,j} = \frac{1}{N^j \times M^j} \sum_{m=1}^{M^j} \sum_{n=1}^{N^j} A^{i,j}_{m,n},$$
$$\text{where} [A^{i,j}_{m,n}]^{M_j,N_j}_{m=1,n=1} = \mathbf{A}^{i,j}, \mathbf{A}^{i,j} = \frac{\partial y_{\text{DR}}}{\partial \mathbf{F}^{i,j}}. \quad (3)$$

In (3), $\mathbf{A}^{i,j}$ indicates the gradient of the final positive DR score[2] y_{DR} with respect to the feature map $\mathbf{F}^{i,j}$ and weight $w^{k,l}$ captures the "importance" of the map of the i-th feature in the j-th layer for positive DR. Finally, given the feature maps $\{\mathbf{F}^{i,j}\}^{K^j,J}_{i=1,j=1}$ and their corresponding importance weights $\{w^{i,j}\}^{K^j,J}_{i=1,j=1}$, we perform a gradient-weighted feature combination to obtain the final visualization map $\mathbf{V}$ as the input to the grading network, the process of which is summarized in Algorithm 1. Specifically, different from the traditional visualization algorithms that only focus on the last layer, we utilize the features of all layers in the network, which can extract fine-grained pathological information in multiple scales.

Algorithm 1: Gradient-weighted feature combination in GMSV.

```
Input: The input fundus image $\mathbf{X}$; the feature maps $\{\mathbf{F}^{i,j}\}^{K^j,J}_{i=1,j=1}$
    of the DR grading network, and their corresponding weights
    $\{w^{i,j}\}^{K^j,J}_{i=1,j=1}$, where $J$ is the total number of layers and $K^j$
    is the number of channels at the $j$-th layer.
Output: The visualization map $\mathbf{V}$ of the grading network.
for $j = J \to 1$ do
    $\mathbf{V}^j \leftarrow \sum_{i=1}^{K^j} w^{i,j}\mathbf{F}^{i,j}$, where $\mathbf{V}^j$ indicates the visualization
      map of the $j$-th layer.
    if $j < J$ then
        $w^j \leftarrow \sum_{i=1}^{K^j} w^{i,j}$, where weight $w^j$ represents the
          importance of the $j$-th layer for DR grading.
        Upsample the $\mathbf{V}^{j+1}$ by a factor of 2, which is denoted as
          $\tilde{\mathbf{V}}^{j+1}$.
        $\mathbf{V}^j \leftarrow \tilde{\mathbf{V}}^{j+1} + w^j\mathbf{V}^j$.
    end
    $j \leftarrow j - 1$.
end
return $\mathbf{V} \leftarrow \mathbf{V}^1$
```

[2] We define the positive DR score by the sum of grades 1 to 4.

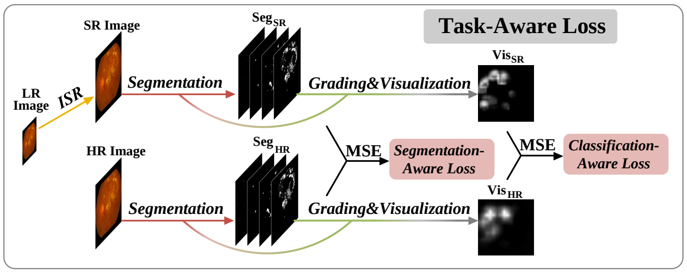


Fig. 7. Pipeline of task-aware loss, i.e., the segmentation-aware Loss and classification-aware Loss.

C. Loss Function

For training our DeepMT-DR network, the loss function mainly includes (1) the task-aware loss guiding the low-level task with the information from the high-level tasks, and (2) the intra-task loss for each single task. Details about the proposed loss function are introduced as follows.

Task-Aware Loss: The task-aware loss is developed to guide ISR to focus more on the most relevant regions of its follow-up tasks. Specifically, the segmentation-aware loss and classification-aware loss are developed for utilizing the feedback information from lesion segmentation and DR grading, respectively. To be specific, let $\tilde{\mathbf{S}}_{\mathbf{sr}}$ and $\tilde{\mathbf{S}}_{\mathbf{hr}}$ denote the segmentation results of the super-resolved image and its corresponding HR image. The segmentation-aware loss is obtained by penalizing the mean square error (MSE) between $\tilde{\mathbf{S}}_{\mathbf{sr}}$ and $\tilde{\mathbf{S}}_{\mathbf{hr}}$:

$$\mathcal{L}_{\text{seg}-\text{aware}} = \|\tilde{\mathbf{S}}_{\mathbf{sr}} - \tilde{\mathbf{S}}_{\mathbf{hr}}\|_2^2. \tag{4}$$

With the guidance of segmentation-aware loss, the ISR subnet is able to focus more on the lesion segmented regions, leading to better segmentation results of the super-resolved image.

Similarly, the classification-aware loss encourages ISR to concentrate on the pathological regions of DR grading. Specifically, we use our GMSV algorithm to visualize the pathology-areas of DR grading from both super-resolved image $\tilde{\mathbf{Y}}$ and corresponding HR image $\mathbf{Y}$. Given the GMSV algorithm, the classification-aware loss can be defined as follows:

$$\mathcal{L}_{\text{cls}-\text{aware}} = \left\|\left(\text{Vis}(\tilde{\mathbf{Y}}) > \phi_{\text{vis}}\right) - \left(\text{Vis}(\mathbf{Y}) > \phi_{\text{vis}}\right)\right\|_2^2, \tag{5}$$

where $\text{Vis}(\cdot)$ refers to the visualization process of our GMSV algorithm; $\oplus$ denotes channel-wise concatenation; and ϕ_{vis} is the threshold to generate the binary mask from $\text{Vis}(\mathbf{Y})$. Benefiting from the classification-aware loss, the super-resolved image tends to highlight pathological regions for DR grading, leading to a better grading result.

Intra-Task Loss: In addition to the above task-aware loss, we further propose the intra-task loss to encourage our network to perform better in each single task, i.e, ISR, lesion segmentation and DR grading.

- For the ISR task, we measure the ℓ_1-norm difference between the super-resolved image $\tilde{\mathbf{Y}}$ and the ground-truth HR image $\mathbf{Y}$ as follows:

  $$\mathcal{L}_{\text{img}}^{\text{isr}} = \|\tilde{\mathbf{Y}} - \mathbf{Y}\|_1. \tag{6}$$

  Besides, to consider the spatial smoothness in the generated image, we calculate the total variation loss to $\tilde{\mathbf{Y}}$:

  $$\mathcal{L}_{\text{tv}}^{\text{isr}} = \sum_{w=1}^{W}\sum_{h=1}^{H}(\|\tilde{\mathbf{Y}}_{w,h+1} - \tilde{\mathbf{Y}}_{w,h}\|_2^2 + \|\tilde{\mathbf{Y}}_{w+1,h} - \tilde{\mathbf{Y}}_{w,h}\|_2^2), \tag{7}$$

  where W and H are the width and height of the image $\tilde{Y}$.
- For lesion segmentation, we develop a *semi-supervised* training method, which utilizes the fine-grained visualization maps as pseudo segmentation ground-truth maps to co-train the lesion segmentation subnet.

  (1) Supervised part. For those data with segmentation annotation, we adopt dice loss [8] to measure the overlap area between segmentation result $\tilde{\mathbf{S}}$ and its ground-truth lesion mask $\mathbf{S}$:

  $$\mathcal{L}_{\text{seg}}^{\text{fully}} = 1 - \frac{2\|\tilde{\mathbf{S}} \circ \mathbf{S}\|_1}{\|\tilde{\mathbf{S}}\|_1 + \|\mathbf{S}\|_1}, \tag{8}$$

  where $\circ$ denotes the hadamard product.

  (2) Unsupervised part. The information flow of the unsupervised part is illustrated in Fig. 6. As shown, for those data without segmentation annotation, the predicted segmentation mask is supervised by the network visualization map. Mathematically, the unsupervised segmentation loss is defined as follows:

  $$\mathcal{L}_{\text{seg}}^{\text{un}} = \|\tilde{\mathbf{S}} - (\text{Vis}(\mathbf{Y}) > \phi_{\text{vis}})\|_2^2, \tag{9}$$

  where $\text{Vis}(\cdot)$ refers to the visualization process of our GMSV algorithm, and ϕ_{vis} is a shared parameter in (5).
- For the DR grading task, we develop a weighted cross-entropy loss to consider not only the classification accuracy but also the importance of different misclassified cases. Specifically, for each training sample, the classification loss is penalized by the distance between the predicted label $\tilde{l}$ and its groundtruth label l. Mathematically, our loss function for the DR grading can be formulated as

  $$\mathcal{L}_{\text{cls}} = -\frac{|\tilde{l} - l|}{\sum_{i=0}^{C-1}(|i-l|+1)} \sum_{j=0}^{C-1} 1\{j=l\}\log \tilde{p_j} \tag{10}$$

  In (10), C is the total number of the DR severity grades; $\tilde{p_j}$ indicates the predicted probability of the j-th grade and $1\{\cdot\}$ denotes the indicator function.

Overall Loss for ISR: Finally, by combining the task-aware and the intra-task losses for ISR, the overall loss function for our ISR subnet can be formulated as

$$\mathcal{L}_{\text{isr}} = \lambda_{\text{img}}\mathcal{L}_{\text{img}}^{\text{isr}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}}^{\text{isr}} + \lambda_{\text{sa}}\mathcal{L}_{\text{seg}-\text{aware}} + \lambda_{\text{ca}}\mathcal{L}_{\text{cls}-\text{aware}}. \tag{11}$$

where $\lambda_{\text{img}}, \lambda_{\text{tv}}, \lambda_{\text{sa}}$ and λ_{ca} are hyper-parameters to balance the intra-task, segmentation-aware and classification-aware losses.

V. Experimental Results

In this section, we mainly focus on evaluating the performance of the proposed DeepMT-DR method on the medical tasks of ISR, lesion segmentation and DR grading tasks. Before performance evaluation, we present the datasets in our experiments. Besides, due to the space limitation, the implementation details are introduced in the supplementary material.

A. Datasets

In our experiments, we evaluate the performance of our DeepMT-DR method on two public DR datasets, i.e., DDR [20] and EyePACS [10]. These two datasets have 13,673 and 88,702 retinal fundus images for DR grading, respectively. In DDR, there are only 757 fundus images annotated with the pixel-wise segmentation for four retinal lesions, including MA, HM, SE and HE. In our experiments, we use the default setting of training, validation and test sets of these two datasets. Note that all images are cropped out the backgrounds and then downsampled to 1024×1024, seen as HR images. Subsequently, to obtain LR-HR pairs, all HR images are downsampled by a factor of 4, to generate the LR images at resolution of 256×256[3].

In addition to the downsampled LR images, we also collected a real-world LR fundus image dataset called Real-LR, including 898 LR fundus images. In Real-LR, 40 images (with resolution of 584×565) were sourced from a public dataset, DRIVE [38], and 858 images (with resolution of 470×380) were acquired by LR imaging devices from Beijing Tongren Hospital. In Real-LR, 611 and 287 images are annotated as the DR negative and positive samples, respectively, according to the screen results of the doctors in Beijing Tongren Hospital. Since there is no lesion segmentation annotation in Real-LR, the test set is only used for evaluating the performance of DR grading. The Real-LR dataset is public online: https://www.dropbox.com/s/b232l3ncktxehx1/Real-LR.zip?dl=0.

B. Performance Evaluation

Evaluation on the Main Task of DR Grading. We evaluate the DR grading performance of our DeepMT-DR method over DDR and EyePACS datasets, compared with 10 other SOTA methods of ResNet-18 [14], Inception-v3 [39], DenseNet-121 [15], M-Net [44], CKML [40], VNXK [40], MMCNN [47], Adly *et al.* [1], Lin *et al.* [21] and Wang *et al.* [42]. Note that Wang *et al.* [42] is the conference version of our method. For fair comparison, all compared methods are conducted on the fundus images super-resolved (using the SOTA ISR method [31]) or downsampled to the required resolutions of these methods. In our experiments, we apply 4 metrics to evaluate the performance of DR grading: accuracy, precision, F_1-score and the Cohen's kappa coefficient [5]. Note that the larger values of these 4 metrics indicate more accurate DR grading. Table I tabulates the results of accuracy, precision, F_1-score and kappa for our and other methods, which are averaged over the test sets of the DDR and EyePACS datasets. As shown in this table, our DeepMT-DR method performs considerably better than all other methods over both datasets, in terms of all 4 metrics. Specifically, on the DDR dataset, our method achieves at least 1.1%, 0.9%, 1.2% and 2.4% improvements in accuracy, precision, F_1-score and kappa, respectively. Similar results can be found in the EyePACS dataset. These significant improvements are mainly due to the following aspects. (1) We design a hierarchical structure for the joint learning of the tasks of lesion segmentation, ISR and DR grading, such that the performance of DR grading can be boosted by lesion segmentation and ISR. (2) We propose the task-aware loss to guide the low-level tasks with the information feedback from the high-level tasks.(3) We develop the GMSV algorithm to extract fine-grained pathological information from the DR grading task, which is then used to guide the training process. The corresponding analysis can be found in the ablation study of Section V-C.

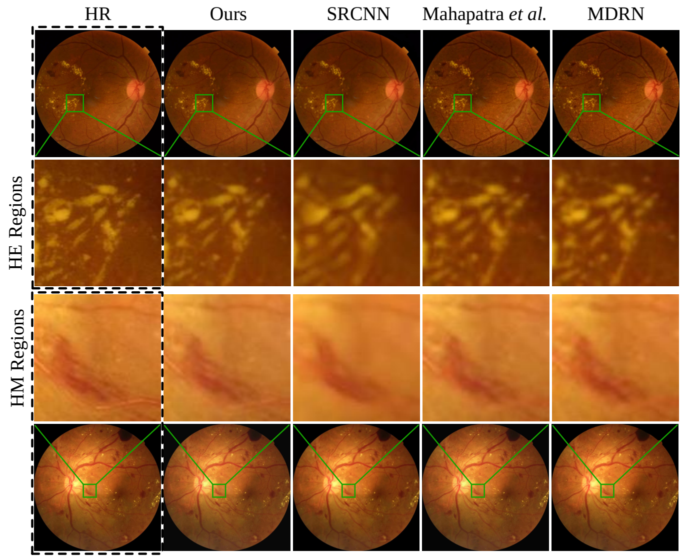


Fig. 8. Qualitative ISR results of our DeepMT-DR and other methods. We crop and zoom-in the pathological region to compare the performance of ISR.

Evaluation on the Auxiliary Task of ISR. Here, we evaluate the performance of our DeepMT-DR method in the auxiliary task of ISR, over DDR dateset. Specifically, we compare the ISR performance of our and 4 SOTA ISR methods, i.e., SRCNN [6], Mahapatra *et al.* [25], MDRN [31] and Wang *et al.* [42], over the DDR dataset, with the upscale factor of 4. Experimental results are shown in Table IV, in terms of Peak Signal-to-Noise Ratio (PSNR), structural similarity (SSIM). As shown, our DeepMT-DR method outperforms all compared methods. Moreover, Figure 8 shows the super-resolved images generated by our DeeMT-DR and 3 compared methods. As shown, our method can clearly super-resolve the LR fundus images, in particular the pathological areas. The above results verify the effectiveness of our DeepMT-DR method in the task of ISR.

Evaluation on the Auxiliary Task of Lesion Segmentation. We further validate the performance of DeepMT-DR method in the auxiliary task of lesion segmentation, via comparing with 5 SOTA methods: U-Net [32], DRU-Net [18], HGN [33], CE-Net [11] and Wang *et al.* [42]. Table III reports the results of average precision (AP) and area under the receiver operating characteristic curve (AUC). As shown, our DeepMT-DR method achieves the best performance for the lesions of MA, HM and HE. Additionally, Figure 9 also compares the subjective segmentation results in all kinds of lesions (MA, HM, SE and HE) generated by our DeeMT-DR and 4 compared methods. As seen, the lesion segmented masks by our method are more close to the ground-truth than other methods. As such, we can conclude that our DeepMT-DR method is effective in the task of lesion segmentation.

Evaluation on the Real-world LR Fundus Images. The above experiments mainly focus on the fundus images that are downsampled to LR. Here, we further evaluate the DR grading performance of our DeepMT-DR and other 9 SOTA methods over a dataset of real-world LR fundus images, i.e., Real-LR. Table II tabulates the results of diagnosis accuracy of our and compared methods. As shown in this table, our DeepMT-DR

[3] Note that we choose 256×256 as the main experimental setting for considering the extreme practical scenario. Besides, another input resolution of 512×512 is used in Section V-D.

TABLE I
MEAN (STANDARD DEVIATION) VALUES IN TERMS OF PERCENTAGE FOR DR GRADING METRICS BY OUR AND OTHER METHODS OVER THE TEST SET OF DDR AND EYEPACS DATASETS.

Approaches	Attributes			Evaluation on DDR				Evaluation on EyePACS			
	For DR grading	MTL*	SSL**	Accuracy	Precision	F_1-score	Kappa	Accuracy	Precision	F_1-score	Kappa
ResNet-18 [14]				73.2(0.9)	77.3(0.6)	72.5(0.8)	70.8(0.7)	78.6(1.3)	79.2(0.7)	78.3(0.6)	77.6(0.4)
Inception-v3 [39]				74.4(0.5)	78.9(0.2)	75.5(0.6)	71.9(0.7)	79.2(0.6)	79.5(0.3)	78.7(0.5)	77.7(0.9)
DenseNet-121 [15]				74.0(0.7)	79.2(0.3)	73.0(0.4)	68.9(0.6)	80.6(0.8)	80.8(0.6)	79.8(1.3)	79.5(0.4)
M-Net [44]	✓			78.9(0.5)	80.1(0.8)	78.4(0.3)	74.2(0.5)	81.8(0.6)	82.7(0.4)	82.5(0.6)	81.3(0.4)
CKML [40]	✓			77.5(0.6)	80.0(0.7)	77.0(0.6)	73.3(0.4)	81.1(0.6)	81.3(0.4)	80.7(0.5)	80.1(0.5)
VNXK [40]	✓			78.3(0.8)	81.6(1.2)	78.2(0.7)	74.7(0.6)	81.8(0.7)	82.1(0.5)	80.8(0.6)	80.6(0.3)
MMCNN [47]	✓	✓		79.5(0.3)	81.6(0.8)	79.4(0.4)	75.7(0.8)	83.1(0.6)	83.7(0.2)	82.9(0.7)	82.6(0.8)
Adly *et al.* [1]	✓			78.6(0.5)	81.2(0.7)	78.3(1.1)	73.9(0.4)	81.3(0.6)	82.3(0.4)	81.6(0.4)	80.9(0.5)
Lin *et al.* [21]	✓	✓		79.2(0.4)	80.6(0.6)	78.7(1.0)	74.9(0.3)	82.8(0.5)	83.3(0.2)	82.2(0.5)	81.8(0.7)
Wang *et al.* [42] †	✓	✓	✓	82.5(0.8)	82.2(0.5)	81.8(0.6)	77.8(0.8)	85.7(0.7)	86.0(0.8)	83.9(0.5)	83.7(0.6)
DeepMT-DR (Ours)	✓	✓	✓	**83.6(0.5)**	**83.1(0.3)**	**83.0(0.4)**	**80.2(0.7)**	**86.9(0.5)**	**87.1(0.3)**	**85.7(0.6)**	**85.2(0.2)**

* MTL refers to multi-task learning. ** SSL refers to semi-supervised learning. † Wang *et al.* [42] is the conference version of our method.

method achieves averagely 87.6% accuracy in DR grading, with at least 6.8% improvement over other SOTA methods. This verifies the significance of our method in the real-world scenarios of LR image based DR diagnosis.

TABLE II
MEAN VALUES IN TERMS OF PERCENTAGE FOR DR GRADING ACCURACY BY OUR AND OTHER METHODS OVER THE REAL-LR DATASET.

	ResNet-18	GoogLeNet	DenseNet-121	M-Net	CKML
Accuracy	77.6	78.2	77.4	78.8	78.6
	VNXK	MMCNN	Adly *et al.*	Lin *et al.*	Ours
Accuracy	79.3	80.8	78.5	79.6	**87.6**

C. Ablation Study

We ablate different components of our DeepMT-DR method to thoroughly analyze their effects on DR grading.

Ablation on ISR. First, we conduct the ablation experiments on ISR in our DeepMT-DR method with the following 2 experimental settings: (1) training without ISR, i.e., remove the ISR subnet and the related task-aware loss; (2) training ISR independently, i.e., the ISR subnet is fixed in the second training stage; (3) training the 3 tasks together, i.e., our DeepMT-DR method. Figure 10 (a) shows the DR grading results with above settings. As shown, the DR grading performance of DeepMT-DR greatly degrades, when ISR is removed or trained independently. This indicates the effectiveness of our framework with ISR being an auxiliary task.

Ablation on Lesion Segmentation. We further conduct ablation experiments to evaluate the impact of lesion segmentation on DR grading. Specifically, we compare the performance of DR grading under 3 experimental settings: (1) training without lesion segmentation, i.e., remove the lesion segmentation subnet and related segmentation-aware loss; (2) training lesion segmentation independently, i.e., jointly train the ISR and DR grading subnets with segmented masks generated from the pre-trained lesion segmentation subnet; (3) training the 3 tasks together, i.e., our DeepMT-DR method. DR grading results with above ablation settings are shown in Figure 10 (b). As shown, our DeepMT-DR method performs worse, when lesion segmentation is removed or trained independently. This indicates that lesion segmentation has positive impact as an auxiliary task.

Ablation on Task-Aware Loss. Moreover, we conduct ablation experiments to evaluate the effect of our task-aware loss, by removing the segmentation-aware loss (SAL) and classification-aware loss (CAL). Figure 10 (c) presents the DR grading results of our DeepMT-DR method under 4 experimental settings. As shown, both the SAL and CAL contribute to the DR grading performance. To conclude, the above results verify the

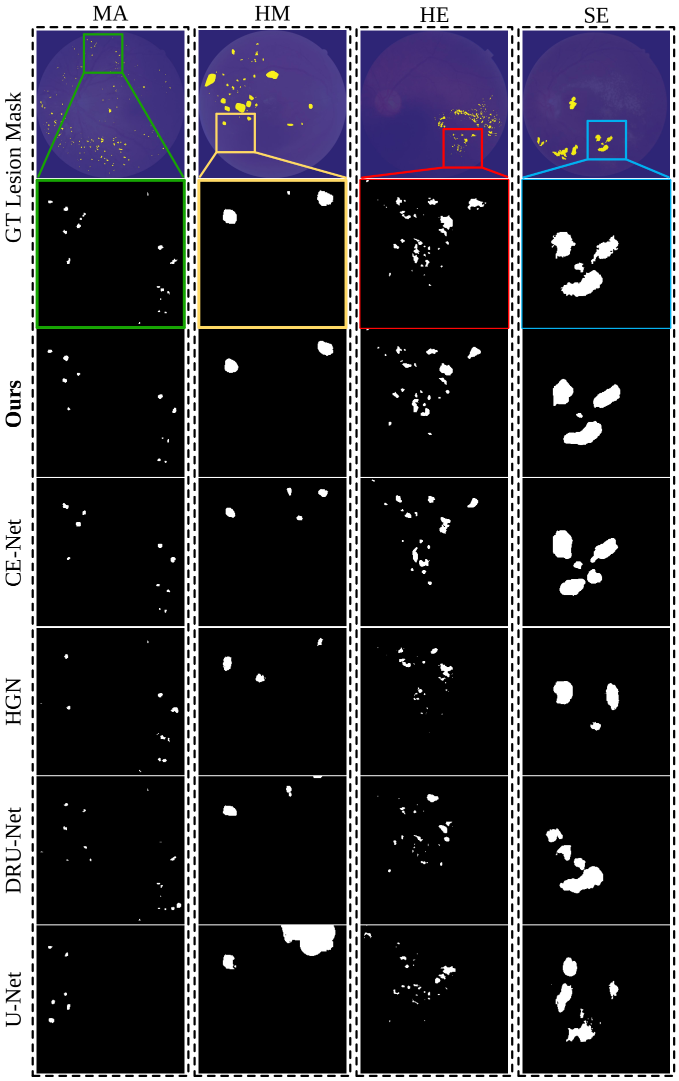
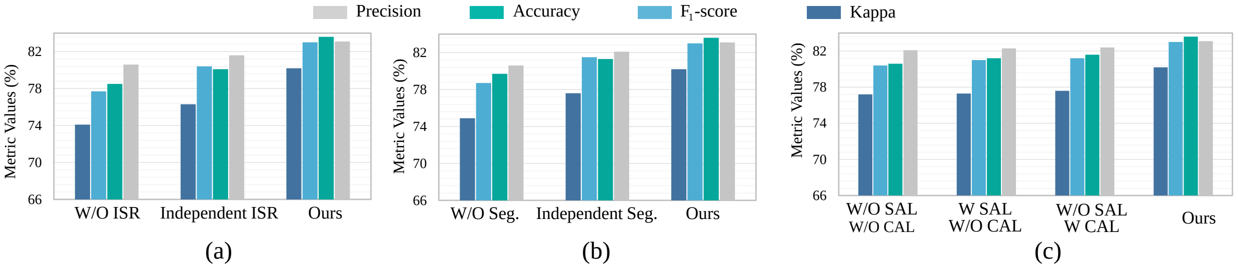


Fig. 9. Qualitative multi-lesion segmentation results. We coarsely mark some regions to compare the lesion segmentation results of our DeepMT-DR method with those of other methods. The first row to the fourth row show the segmentation results of MA, HM, SE and HE, respectively.

TABLE III
MEAN (STANDARD DEVIATION) VALUES IN TERMS OF PERCENTAGE FOR AUC AND AP OF OUR AND OTHER LESION SEGMENTATION METHODS OVER DDR DATASET.

Leiosns	Microaneurysms (MA)		Haemorrhages (HM)		Hard Exudates (HE)		Soft Exudates (SE)		Averaged Result	
Methods	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
U-Net [32]	95.7(0.6)	45.5(0.4)	95.3(0.5)	49.4(0.2)	96.5(0.3)	65.5(0.3)	97.2(0.9)	58.4(0.2)	96.2	54.7
DRU-Net [18]	98.1(0.7)	49.3(0.1)	95.8(0.4)	50.4(0.2)	98.1(0.5)	67.0(0.1)	98.5(0.6)	62.8(0.3)	97.6	57.4
HGN [33]	97.4(0.4)	48.1(0.1)	96.1(0.6)	51.8(0.6)	98.4(0.8)	67.7(0.5)	98.6(0.4)	63.1(0.2)	97.2	57.7
CE-Net [11]	97.6(0.7)	48.5(0.1)	96.4(0.6)	52.3(0.8)	98.2(0.2)	67.6(0.7)	**99.4**(0.7)	**65.2**(0.5)	97.9	58.4
Wang *et al.* [42]	98.7(0.6)	50.1(0.1)	97.1(0.7)	55.4(0.3)	99.0(0.7)	71.8(0.2)	99.2(0.8)	64.7(0.1)	98.5	60.5
Ours	**98.8**(0.5)	**50.1**(0.1)	**97.3**(0.5)	**55.5**(0.5)	**99.0**(0.8)	**72.0**(0.2)	99.2(0.8)	65.0(0.2)	**98.6**	**60.6**



(a) (b) (c)

Fig. 10. The DR grading results of the ablation experiments on ISR (a), lesion segmentation (b) and task-aware loss (c), over DDR dataset.

TABLE IV
MEAN (STANDARD DEVIATION) VALUES IN TERMS OF PSNR AND PERCENTAGE FOR SSIM BY OUR AND OTHER ISR METHODS OVER DDR DATASET.

	Ours	Wang *et al.*	SRCNN	Mahapatra *et al.*	MDRN
PSNR (dB)	**39.9**(0.3)	39.7(0.5)	36.8(0.4)	38.1(0.6)	38.9(0.6)
SSIM	**0.892**(0.08)	0.883(0.06)	0.772(0.03)	0.836(0.08)	0.856(0.10)

effectiveness of the task-aware loss.

TABLE V
MEAN VALUES IN TERMS OF PERCENTAGE FOR DR GRADING ACCURACY BY OUR AND OTHER NETWORK VISUALIZATION ALGORITHMS OVER DDR DATASET.

	W/O vis	GBP [37]	Grad-CAM [34]	SmoothGrad [36]	Ours
Acc.	81.7	82.0	82.2	82.3	**83.6**

Ablation on GMSV algorithm. Finally, we evaluate the impact of network visualization (by our GMSV algorithm) on DR grading. To this end, we conduct the ablation experiments by removing the GMSV algorithm and directly using DR scores as the classification-aware loss. This model is denoted as W/O vis in Table V. Furthermore, we also compare the DR grading performance in Table V, using our GMSV and 3 other commonly used network visualization algorithms, including GBP [37], Grad-CAM [34] and SmoothBP [36]. As shown, the performance of DR grading degrades when removing the visualization algorithm. Meanwhile, our DeepMT-DR method with GMSV performs the best among all visualization algorithms. This indicates the advantage of our GMSV algorithm in extracting the pathological information for DR grading.

D. Analysis on Experimental Settings

Influence of Input Resolution. We evaluate the performance of our DeepMT-DR method on the tasks of ISR, lesion segmentation and DR grading, when the resolution of input fundus images varies between 512 × 512 and 256 × 256. Table VI tabulates the results of DR grading (in accuracy and kappa), lesion segmentation (in AUC and AP) and ISR (in PSNR and SSIM) of our method at input resolutions of 512 × 512 and 256 × 256. As shown in this table, the performance of all 3 tasks is slightly improved, after the resolution of input images is increased from 256 × 256 to 512 × 512. For instance, the grading results of 512 × 512 are 84.3% and 82.1% in accuracy and kappa, whereas those of 256 × 256 are 83.6% and 80.2%, respectively. To summarize, the above results imply the performance improvement of our method given higher input resolution.

TABLE VI
MEAN VALUES IN TERMS OF METRICS FOR THE 3 TASKS OF ISR, LESION SEGMENTATION AND DR GRADING WITH VARIED INPUT RESOLUTIONS OF OUR METHOD OVER DDR DATASET.

Task	DR grading		Segmentation		ISR	
	Accuracy	Kappa	AUC	AP	PSNR	SSIM
512×512	84.3	82.1	98.9	61.1	45.2	0.915
256×256	83.6	80.2	98.6	60.6	39.9	0.892

Amount of Pixel-wise Segmentation Supervision. In our method, we propose a novel semi-supervised training strategy to overcome the paucity of pixel-wise segmentation annotations. Thus, we conduct additional experiments to evaluate the effectiveness of the proposed strategy on the DR grading performance, via training with different amounts of pixel-wise segmentation annotations. The experimental results show that the DR grading accuracy of our method decreases by 0.9%, 1.3% and 2.2%, when training with 75%, 50% and 25% pixel-wise supervision, respectively. It can be seen that our method performs better than most compared methods, even with 25% pixel-wise segmentation supervision, implying the effectiveness of the proposed semi-supervised training strategy.

VI. CONCLUSION AND FUTURE WORK

This paper has proposed a multi-task learning method, called DeepMT-DR, for the main task of DR grading and the auxiliary tasks of ISR and lesion segmentation on fundus images. First, we thoroughly analyzed the correlation among the tasks of DR grading, ISR and lesion segmentation, and then found the potential gain of jointly learning these three tasks. Second, we proposed our DeepMT-DR method with a multi-task learning structure and task-aware loss. Besides, the GMSV algorithm was proposed to supervise the ISR task through the feedback information from the high-level tasks of lesion segmentation and DR grading. Finally, the experimental results verified that our DeepMT-DR method significantly outperforms 10 other

SOTA methods in DR grading, and it also achieves comparable performance in the ISR and lesion segmentation tasks.

There are two promising directions for future works. First, our work at the current stage mainly focuses on the low-level image enhancement task of super resolution. Other image enhancement tasks, e.g., image deblurring and image denosing, can be also embedded in our DR grading method, seen as a potential future work. Second, it is also interesting to implement our method in the automatic diagnosis of other diseases, e.g., the nephropathy and skin diseases.

References

[1] M. M. Adly, A. S. Ghoneim, and A. A. Youssif. On the grading of diabetic retinopathies using a binary-tree-based multiclass classifier of cnns. *International Journal of Computer Science and Information Security*, 17(1), 2019.

[2] M. Alhajeri and S. G. S. Shah. Limitations in and solutions for improving the functionality of picture archiving and communication system: An exploratory study of pacs professionals' perspectives. *Journal of digital imaging*, 32(1):54–67, 2019.

[3] N. Asiri, M. Hussain, F. Al Adel, and N. Alzaidi. Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. *Artificial intelligence in medicine*, 2019.

[4] M. A. Bravo and P. A. Arbeláez. Automatic diabetic retinopathy classification. In *MICCAI Workshop*, volume 10572, page 105721E, 2017.

[5] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[6] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199, 2014.

[7] Z. Feng, J. Yang, L. Yao, Y. Qiao, Q. Yu, and X. Xu. Deep retinal image segmentation: A fcn-based architecture with short and long skip connections for retinal image segmentation. In *ICNIP*, pages 713–722, 2017.

[8] L. Fidon, W. Li, L. C. Garcia-Peraza-Herrera, J. Ekanayake, N. Kitchen, S. Ourselin, and T. Vercauteren. Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. In *MICCAI Workshop*, pages 64–76, 2017.

[9] R. Gao, L. Li, Y. Tang, S. L. Antic, A. B. Paulson, Y. Huo, K. L. Sandler, P. P. Massion, and B. A. Landman. Deep multi-task prediction of lung cancer and cancer-free progression from censored heterogenous clinical imaging. In *Medical Imaging 2020: Image Processing*, volume 11313, page 113130D. International Society for Optics and Photonics, 2020.

[10] B. Graham. Kaggle diabetic retinopathy detection competition report. *University of Warwick*, 2015.

[11] 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. *TMI*, 2019.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *CVPR*, pages 2961–2969, 2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *TPAMI*, 37(9):1904–1916, 2015.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.

[16] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[17] I. Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6129–6138, 2017.

[18] C. Kou, W. Li, W. Liang, Z. Yu, and J. Hao. Microaneurysms segmentation with a u-net based on recurrent residual convolutional neural network. *Journal of Medical Imaging*, 6(2):025008, 2019.

[19] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017.

[20] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 2019.

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

[22] L. Liu, Q. Dou, H. Chen, I. E. Olatunji, J. Qin, and P.-A. Heng. Mtmr-net: Multi-task deep learning with margin ranking loss for lung nodule analysis. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 74–82. Springer, 2018.

[23] S. Liu, E. Johns, and A. J. Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1871–1880, 2019.

[24] D. Mahapatra, B. Bozorgtabar, and R. Garnavi. Image super-resolution using progressive generative adversarial networks for medical image analysis. *Computerized Medical Imaging and Graphics*, 71:30–39, 2019.

[25] D. Mahapatra, B. Bozorgtabar, S. Hewavitharanage, and R. Garnavi. Image super resolution using generative adversarial networks and local saliency maps for retinal image analysis. In *MICCAI*, pages 382–390, 2017.

[26] I. MALAK, I. MAY, A. E.-A. A. SAAD, and A. FOAD. Asymmetric diabetic retinopathy and carotid insufficiency: A correlative study. *The Medical Journal of Cairo University*, 87(June):1331–1335, 2019.

[27] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, pages 3994–4003, 2016.

[28] S. S. Nathan. A review: Image analysis techniques to improve labeling accuracy of medical image classification. In *SCDM*, volume 700, page 298, 2018.

[29] M. F. Nørgaard and J. Grauslund. Automated screening for diabetic retinopathy–a systematic review. *Ophthalmic research*, 60(1):9–17, 2018.

[30] M. S. Rad, B. Bozorgtabar, C. Musat, M. Marti, H. K. Ekenel, and J.-P. Thiran. Benefiting from multitask learning to improve single image super-resolution. *Neurocomputing*, 2019.

[31] S. Ren, D. K. Jain, K. Guo, T. Xu, and T. Chi. Towards efficient medical lesion image super-resolution based on deep residual networks. *SPIC*, 75:1–10, 2019.

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

[33] M. H. Sarhan, S. Albarqouni, M. Yigitsoy, N. Navab, and A. Eslami. Multi-scale microaneurysms segmentation using embedding triplet loss. In *MICCAI*, pages 174–182, 2019.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IJCV*, 128(2):336–359, 2020.

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[36] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv:1706.03825*, 2017.

[37] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv:1412.6806*, 2014.

[38] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *TMI*, 23(4):501–509, 2004.

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

[40] H. H. Vo and A. Verma. New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space. In *ISM*, pages 209–215, 2016.

[41] X. Wang, M. Xu, L. Li, Z. Wang, and Z. Guan. Pathology-aware deep network visualization and its application in glaucoma image synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 423–431. Springer, 2019.

[42] X. Wang, M. Xu, J. Zhang, L. Jiang, and L. Li. Deep multi-task learning for diabetic retinopathy grading in fundus images. In *AAAI*, 2021.

[43] X. Wang, K. Yu, C. Dong, and C. Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, pages 606–615, 2018.

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

[45] T. D. Weber and J. Mertz. Retina and choroid imaging with transcranial back-illumination. In *Clinical and Translational Biophotonics*, pages CF3B–8. Optical Society of America, 2018.

[46] Y. Zheng, M. He, and N. Congdon. The worldwide epidemic of diabetic retinopathy. *Indian journal of ophthalmology*, 60(5):428, 2012.

[47] K. Zhou, Z. Gu, W. Liu, W. Luo, J. Cheng, S. Gao, and J. Liu. Multi-cell multi-task convolutional neural networks for diabetic retinopathy grading. In *EMBC*, pages 2724–2727, 2018.

[48] Y. Zhou, X. He, L. Huang, L. Liu, F. Zhu, S. Cui, and L. Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. In *CVPR*, pages 2079–2088, 2019.

[49] Y. Zhou, B. Wang, X. He, S. Cui, F. Zhu, L. Liu, and L. Shao. Dr-gan: Conditional generative adversarial network for fine-grained lesion synthesis on diabetic retinopathy images. *arXiv:1912.04670*, 2019.