Supplementary Material

I. 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 resolution of input fundus images is increased.

Analysis: Figure 1 (a) shows the results of DR grading by various algorithms [12], [14], [3], [4] at different input resolutions over a large scale DR image dataset, i.e., the DDR [6] 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) [2] before the dense-connection layers in the algorithms. As can be seen in the Figure 1 (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 [4]. 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 1 (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.

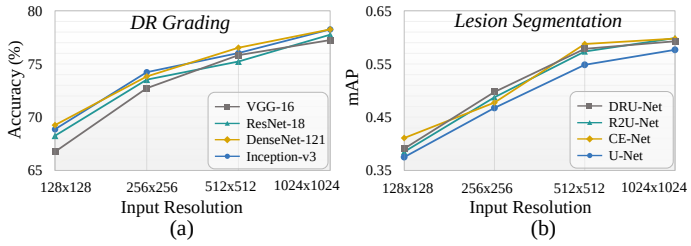

Fig. 1. Task correlation analysis. (a) Accuracy of DR grading *vs.* input resolution ranging from 128 × 128 to 1024 × 1024. (b) Mean average precision (mAP) of retinal lesion segmentation *vs.* varying input resolutions.

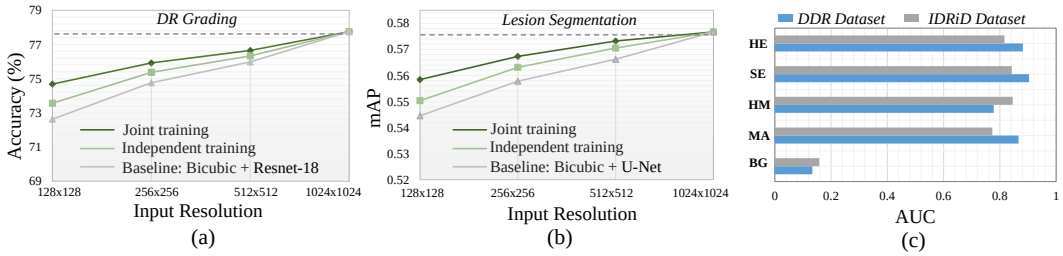

Fig. 2. Task correlation analysis. (a) Accuracy of DR grading on the settings of joint training and independent training of ISR and DR grading network. Note that the input images are all upscaled to 1024 × 1024. (b) mAP of retinal lesion segmentation on the settings of joint training and independent 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.

Finding 2: Joint training of ISR and DR grading performs better than independent training of these two tasks. Similar results can be found for 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.* [9] and the Resnet-18 [3] 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 2 (a). This figure shows that the DR grading accuracy in the setting of joint

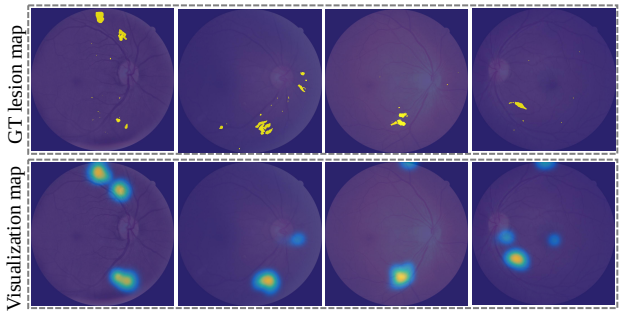

Fig. 3. Ground-truth lesion segmentation maps and their corresponding visualization results of the DR grading network, i.e., Resnet-18.

training performs better than that of independent training. Similarly, as shown in Figure 2 (b), joint training of ISR and lesion segmentation (with the Mahapatra *et al.* [9] and U-Net [10] 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 verify this correlation, we provide both qualitative and quantitative supporting results. Specifically, we apply the commonly used network visualization algorithm, i.e., the Grad-CAM [11], to generate the evidence map of the final decision from the DR grading network, i.e., Resnet-18 [3], which is further utilized as the DR grading subnet of our method. Figure 3 shows some examples of the ground-truth lesion segmentation maps and their corresponding visualization results by the DR grading network. As can be seen from this figure, the pathological areas in the visualization maps significantly overlap with the segmented regions. We then calculate the area under the receiver operating characteristic curve (AUC) values between the evidence maps and different lesion segmented regions (including microaneurysms (MA), haemorrhages (HM), soft exudates (SE) and hard exudates (HE)), and we also calculate the AUC values between the evidence maps and the (BG) regions. Figure 2 (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 versus AUC is 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. From both the qualitative and quantitative results, we can conclude that the lesion segmentation map and visualization map are highly consistent with each other.

II. Structure of ISR and Lesion Segmentation Subnets

The structure of the proposed ISR subnet is shown in the Figure 4 . As shown, the LR image $\mathbf{X}$ is firstly processed with a convolutional layer followed by a leaky rectified linear unit (Leaky ReLU) [8] activation. Then, the output feature is processed with 5 cascaded feature extraction layers to extract pathological information. Finally, for generating the super-resolved fundus image $\tilde{\mathbf{Y}}$, the output of the feature extraction layers is processed by 2 up-scaling layers with the upscale factor of 2 in each layer.

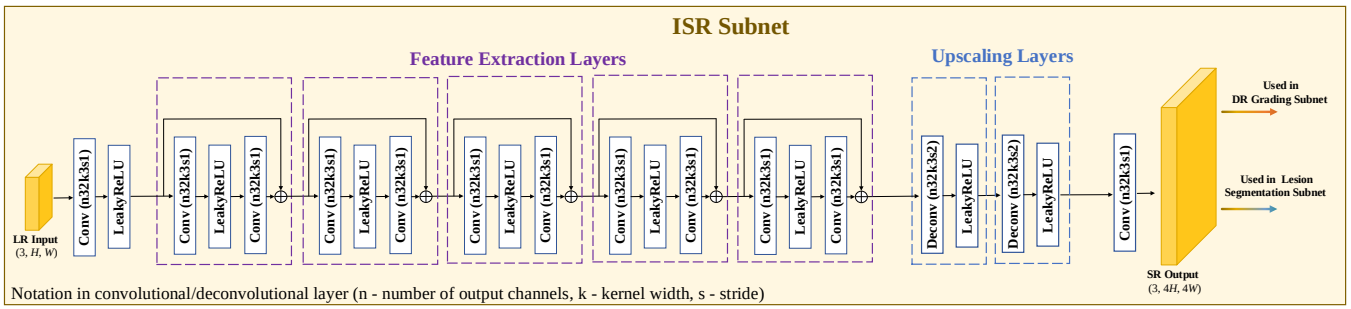

Fig. 4. Structure of the ISR subnet in the proposed DeepMT-DR method. The ISR subnet consists of several cascaded components, i.e., 5 feature extraction layers and 2 upscaling layers. Note that the upscale factor is set to 4 in this paper, and it can be easily extended to other values by adding or removing the upscaling layers.

In addition, the structures of the proposed lesion segmentation subnet is shown in Figure 5. 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 lesion areas. Specifically, as the input, the super-resolved image $\tilde{\mathbf{Y}}$ is progressively contracted and

down-sampled through 5 DT layers. In this way, the contextual features of the retinal images can be captured as the outputs of the last DT layer. Similarly, the contracted features are progressively expanded and up-sampled through 4 UT layers.

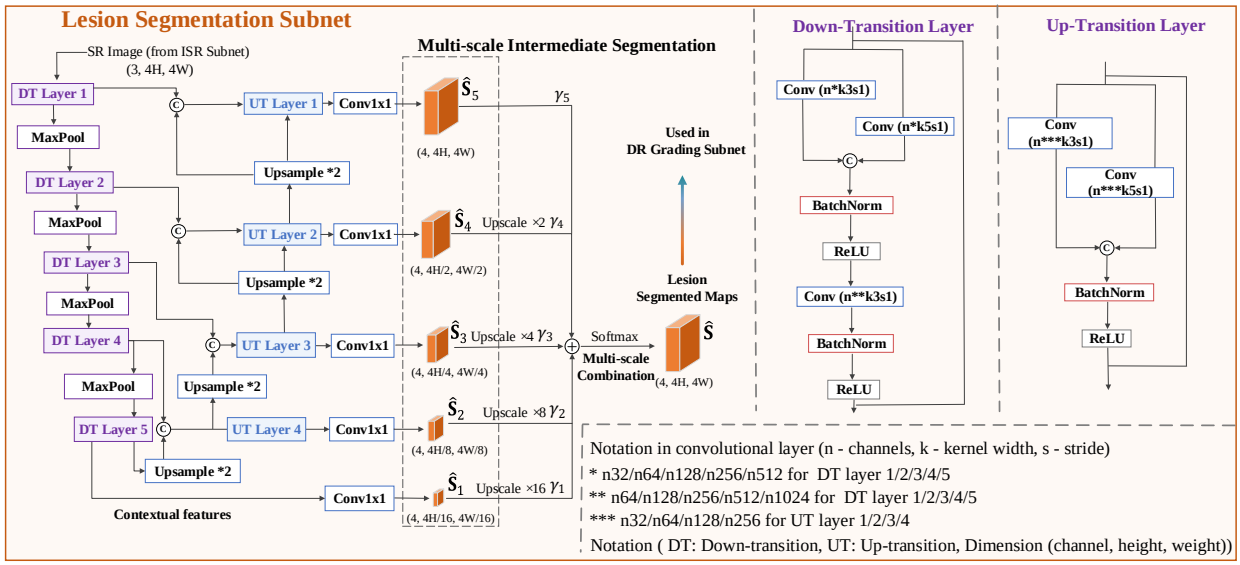

Fig. 5. Structure of the lesion segmentation subnet in the proposed DeepMT-DR method. The lesion segmentation subnet consists of 5 down-transition layers and 4 up-transition layers.

III. Implementation Details and Datasets

Implementation Details. There are two stages for training the DeepMT-DR model. In the first stage, we jointly pre-train the subnets of the two auxiliary tasks, i.e., ISR and lesion segmentation, in order to extract sufficient pathological features for the main task of DR grading. Besides, we also pre-train the DR grading subnet in the first stage, via taking it as a multi-class classification task on DR severity. During the training of the first stage, multiple data augmentation strategies are conducted, including the random rotation, horizontal flips, vertical flips and random changing of the brightness, contrast and saturation of the images. In the second stage, we simultaneously fine-tune the subnets of ISR, lesion segmentation and DR grading, in an end-to-end manner. Note that, in both stages, the parameters are updated using the Adam [5] optimizer, together with the weight decay. The values of key hyper-parameters in both training stages are listed in Table I, in which all of the hyper-parameters are tuned to achieve the best performance over the validation set. All experiments are conducted on a computer with an Intel(R)Xeon E5-2698 CPU@2.70GHz, 256GB RAM and 4 Nvidia Tesla V100 GPUs. Specifically, all of the 4 GPUs are used for training in the first stage, while only 1 GPU is used in the second training stage. It takes around 40 hours for training our model, and the inference time is around 45 ms per fundus image. Additionally, our method is implemented on PyTorch with the Python environment.

TABLE I
VALUES OF SOME KEY HYPER-PARAMETERS IN THE TWO TRAINING STAGES.

Stage I	Batch size	8
	Initial learning rate	1×10^{-4}
	Weight decay	5×10^{-6}
	λ_{img} for $\mathcal{L}_{\text{isr}}$ in equation (11)	1
	λ_{tv} for $\mathcal{L}_{\text{isr}}$ in equation (11)	1×10^{-6}
	λ_{sa} for $\mathcal{L}_{\text{isr}}$ in equation (11)	10
Stage II	Batch size	1
	Initial learning rate	1×10^{-5}
	Weight decay	5×10^{-6}
	Threshold θ_{seg} in equation (2)	0.4
	Threshold $\xi^{i,j}$ in equation (2)	0.5
	Threshold ϕ_{vis} in equation (5) and (9)	0.5
	λ_{img} for $\mathcal{L}_{\text{isr}}$ in equation (11)	1
	λ_{tv} for $\mathcal{L}_{\text{isr}}$ in equation (11)	1×10^{-6}
	λ_{sa} for $\mathcal{L}_{\text{isr}}$ in equation (11)	1
	λ_{ca} for $\mathcal{L}_{\text{isr}}$ in equation (11)	10
	λ_{isr}, λ_{cls} and λ_{seg} for $\mathcal{L}$ in equation (12)	1

TABLE II
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

Datasets. In our experiments, we evaluate the performance of our DeepMT-DR method on two public DR datasets, i.e., DDR [6] and EyePACS [1]. 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. Besides, in DDR, the proportions of negative, mild, moderate, severe and proliferative DR (denoted as grades 0, 1, 2, 3, and 4) are 48.9%, 5.1%, 36.6%, 1.9%, and 7.5%, respectively. Similarly, grades 0, 1, 2, 3, and 4 occupy 73.3%, 6.9%, 15.2%, 2.6%, and 2.0% samples in EyePACS, respectively. In our experiments, we use the default setting of training, validation and test sets of these two datasets of DDR and EyePACS for DR grading. 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, we follow the down-sampling strategy that is commonly used in the ISR field. Specifically, all HR images are downsampled by a factor of 4, to generate the LR images at resolution of 256×256.[1]

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, to ensure the multi-institution and global diversity, 40 images (with resolution of 584×565) were sourced from a public dataset of DRIVE [13], and 858 images (with resolution of 470×380) were acquired by LR imaging devices from Beijing Tongren Hospital. Note that the fundus images in DRIVE were sourced from the DR screening program in the Netherlands [13]. 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 data 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.

IV. 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 II 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.

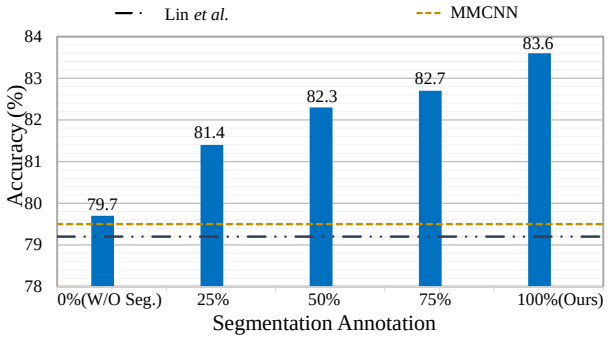

Fig. 6. DR grading accuracy of our method with decreased segmentation annotations used in training stage.

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. As shown in Figure 6, the DR grading accuracy of our method achieves 83.6%, 82.7%, 82.3%, 81.4% and 79.7%, when using 100%, 75%, 50%, 25% and 0% segmentation annotations as training supervision, respectively.

[1]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 **??**.

TABLE III
MEAN VALUES IN TERMS OF PERCENTAGE FOR DR GRADING ACCURACY BY 2 SOTA COMPARED METHODS AN OUR METHOD WITH THE HYPER-PARAMETERS, I.E., ξ AND ϕ_{vis}, OF DIFFERENT VALUES OVER THE TEST SET OF DDR DATASET.

	SOTA compared methods		ξ (ϕ_{vis}=0.5)					ϕ_{vis} (ξ=0.5)				
	Lin *et al.*	MMCNN	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
Accuracy	79.2	79.5	81.9	82.7	**83.6**	83.1	82.3	81.6	82.1	**83.6**	82.0	81.1

In addition, as two best performed compared methods, the results of Lin *et al.* [7] (79.2%) and MMCNN [15] (79.5%) are also shown in the Figure as baselines. According to Figure 6, compared with the baseline models, our method still performs well even when using less segmentation annotations. This implies the effectiveness of the proposed semi-supervised training strategy in the real-world scenarios of disjoint datasets.

Sensitivity analysis of the hyper-parameters. In order to analyze the robustness of our method to the hyper-parameters, we conduct the sensitivity analysis on some important hyper-parameters, i.e., ξ and ϕ_{vis} in Equations (2), (5) and (9). Specifically, ξ is a threshold to produce the binary mask from the upsampled feature map in the DR grading subnet, while ϕ_{vis} is the threshold to generate the binary mask from the visualization map generated by the proposed GMSV algorithm. The experimental results are shown in Table III, in which we train DeepMT-DR with different values of ξ and ϕ_{vis}. Table III shows the DR grading performance at different values of ξ and ϕ_{vis}. From Table III, we can see that the performance of DR grading is slightly changed at different values of ξ and ϕ_{vis}, still better than other competitive SOTA methods, e,g., Lin *et al.* [7] and MMCNN [15] (see the results in Table II). This indicates that our method is only slightly sensitive to the hyper-parameters and is thus can be used for real-world applications.

REFERENCES

[1] B. Graham. Kaggle diabetic retinopathy detection competition report. *University of Warwick*, 2015.
[2] 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.
[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
[4] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.
[5] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.
[6] 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.
[7] 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.
[8] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.
[9] 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.
[10] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.
[11] 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.
[12] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.
[13] 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.
[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.
[15] 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.