Appendix for "Image Quality-aware Diagnosis via Meta-knowledge Co-embedding"

Haoxuan Che Siyu Chen Hao Chen*
The Hong Kong University of Science and Technology

1. Datasets Details

We used five datasets in our experiments, namely **DRAC** [1], **DeepDR** [2], **EYEQ** [3], **CT-IQAD** [4, 5], and **CXR-IQAD** [6, 7], for five different diagnosis tasks. As previously mentioned, Table 1 in the main section shows the label distribution for each dataset. Here, we provide more details on each dataset and our experimental settings.

DRAC. DRAC is a public competition in the Grand Challenge, which consists of 997 ultra-wide Optical Coherence Tomography Angiography (OCTA) images for diabetic retinopathy (DR) diagnosis. Ultra-wide OCTA images can provide more information than normal OCTA images and can demonstrate retina features in great detail. However, these images usually suffer from image quality problems, such as low-signal and segment artifacts, due to patients' movement and environmental conditions. Medical experts can evaluate DR grade by visually inspecting lesions in the foveal avascular zone and vascular structure. In this dataset, DR is graded into three levels, representing no diabetic retinopathy (NoDR), non-proliferative diabetic retinopathy (NPDR), and proliferative diabetic retinopathy (PDR), respectively. The quality of each image is evaluated separately, and the quality annotation is a binary label with high-quality (HQ) or low-quality (LQ), where LQ images have obvious artifacts. The original setting of this dataset divides it into training and testing sets with 611 and 386 images, respectively, and we follow this setting in our experiments.

DeepDR. DeepDR provides fundus images from more than 1000 patients for DR grading with different quality levels. Fundus imaging is a standard screening tool for DR diagnosis. Unlike other datasets, DeepDR provides dual-view fundus images from the same eyes with different areas in the center. The images in the dataset are evaluated and graded into five levels based on DR lesions. To align with the DRAC dataset, we reorganized the annotations of DeepDR to include three levels: no DR, NPDR, and PDR. Similar to OCTA, fundus images also contain artifacts, and HQ or LQ labels are provided to indicate image quality levels. This dataset comprises a total of 2000 images, and we followed their original data splits, i.e., 1200 training images, 400 validation images, and 400 testing images.

EyeQ. EYEQ is a dataset that has been re-annotated from the EyePACS dataset. The EyePACS dataset originally assessed retinal image quality with a binary annotation basis, but EYEQ selected 28,792 retinal fundus images and re-annotated them into three levels for good, usable, or poor quality. Images in this dataset are graded into five levels of diabetic retinopathy (DR) according to severity. Similar to the DeepDR dataset, we combined the second, third, and fourth levels into the non-proliferative diabetic retinopathy (NPDR) class, while the first and fifth level images are regarded as no DR and proliferative diabetic retinopathy (PDR), respectively. Among the images, we used 17,274 images for training, 2,881 images for validation, and 8,637 images for testing.

CT-IQAD. Our CT-IQAD dataset is composed of 746 computed tomography (CT) images from COVID-X [5] and 2600 from SARS [4]. These datasets were originally collected to evaluate whether patients are infected with COVID-19, and the images are annotated as normal for non-infection and COVID-19 for infected cases. However, due to the nature of COVID-X images being collected in the wild, such as those downloaded from papers, their quality cannot be guaranteed, while SARS collects data from hospital patients with relatively higher quality. Thus, we labeled the images in COVID-X as LQ and those in SARS as HQ. The CT-IQAD dataset is divided into 1641 training images, 235 validation images, and 470 test images.

CXR-IQAD. It consists of chest X-ray (CXR) images from two different sources: [6] for child images and [7] for adult images. All images are labeled as either Normal or Pneumonia. To simulate low-dose CXR images [8], we downsampled 2928 images from [6] and all images from [7] using a bicubic kernel, which are considered LQ. The dataset is divided into three subsets: child images (HQ), low-dose child images (LQ-C), and low-dose adult images (LQ-A). We split the dataset into 5780 training images, 825 validation images, and 1651 test images.

[1] Corresponding author: Hao Chen, email: jhc@cse.ust.hk.

2. Implementation Details

We conducted our experiments using the Pytorch framework and ran them on a GeForce RTX™ 3090 GPU. All images were resized to 256×256 and normalized to zero mean and unit variance in intensity values, individually for each dataset, before being divided into batches. The batch size was chosen adaptively to account for differences between image modalities. We used VGG16 as the backbone, along with focal loss and entropy loss for both the Task Net and Meta Learner. We also utilized SGD as the optimizer, with Task Net's learning rate α and weight decay strategy kept constant at 0.01 and 0.0005, respectively, for all datasets. Moreover, the Meta Learner's learning rate β was chosen based on the specific dataset being used. To balance the effectiveness of Task Net and Meta Learner, we applied different weights to the entropy loss for each dataset. We tuned the length of y_ω individually for each dataset to improve performance. A summary of the training parameters used for each dataset is provided below.

DRAC. The number of training epochs is set as 200, and we report the last epoch result because the official data split does not have a validation set. The learning rates for task net and meta learner are both 0.01. The batch size is 8, and the weight of entropy loss is 0.5, and the length of y_ω is 7.

DeepDR. We set the number of training epochs is set as 200 and validate the model per 20 epochs to select the best model and evaluate it on the test set. The learning rates for task net and meta learner are 0.01 and 0.001, respectively. The batch size is 4, and the weight of entropy loss is 0.3, and the length of y_ω is 10.

EyeQ. We train the model for 100 epochs and validate the model per 10 epochs to select the best model. The learning rates for task net and meta learner are 0.01 and 0.001, respectively. The batch size is 4, and the weight of entropy loss is 0.2, and the length of y_ω is 5.

CT-IQAD. Similar to DeepDR, the number of training epochs is 200, and the validation is made per 20 epochs. The learning rates for task net and meta learner are both 0.01. The batch size is 4, and the weight of entropy loss is 0.2, and the length of y_ω is 5.

CXR-IQAD. The number of training epochs is also 200 and validation interval is 20 epochs. The learning rates for task net and meta learner are 0.01 and 0.001, respectively. The batch size is 8, and the weight of entropy loss is 0.3, and the length of y_ω is 7.

We took into consideration the unique characteristics of each dataset when selecting training parameters. For instance, we maintained the Meta Learner learning rate at 0.001 across all datasets except for DRAC and CT-IQAD, where the number and proportion of LQ images are lower than other datasets, and thus set the Meta Learner learning rate to 0.01. We also adjusted the batch size based on the nature of the images in each dataset. OCTA images contain clear lesion information compared to fundus images, which led us to set the batch size to 8 for DRAC and 4 for DeepDR and EYEQ. Additionally, since LQ images in CXR-IQAD are simulated and those degradations appear similar, compared to CT-IQAD, we set their batch sizes to 8 and 4, respectively. To adapt to different datasets, we applied different weights to the entropy loss. For small datasets like DRAC, we set a higher entropy loss weight of 0.5 to provide more encouragement to Meta Learner to generate appropriate y_ω. For larger datasets like EYEQ and CXR-IQAD, we set the weight to 0.2. Finally, for other datasets, we set the weight to 0.3.

3. Comparison Details

Ophthalmic disease assessment. MMCNN [9] employs a multi-cell architecture that performs regression and classification jointly. BIRA-Net [10] uses a two-stream CNN architecture with an attention module and bi-linear strategy. GREEN [11] utilizes a graph convolutional network with a class dependency prior for disease diagnosis tasks. CAB-Net [12] learns discriminative features for each disease category using a categorical attention block.

Multi-task & auxiliary learning. QGNet [13] uses image quality assessment as an auxiliary branch of the model supervised by center loss and weighted softmax loss. CANet [14] explicitly explores the internal relationship between diseases via attention-based modules. Multitask-Net [15] takes advantage of specific task layers to conduct multi-lesion diagnosis. MTMR-Net [16] proposes a margin ranking loss and explicitly leverages the relationship between regression and classification for disease diagnosis. MAXL [17] employs meta-auxiliary learning for self-supervision in primary tasks. DETACH [18] proposes a dual-stream disentangled learning architecture on the task level to explore potential relationships among diseases.

Other adaptable methods. Mixup [19] is a classic augmentation method that mixes images to improve robustness. Mixstyle [20] conducts the mixing of feature statistics of training samples across domains to improve model robustness. Augmix [21] is a simple data processing technique that generates augmented images automatically to improve model performance on unseen domains. DDAIG [22] uses adversarial training to generate perturbed images to improve generalization ability.

4. Class Activation Map Visualization

In this paper, we use class activation mappings (CAMs) to perform a qualitative analysis of how MKCNet works. As shown in Figure 1, we present additional samples of OCTA images from DRAC and fundus images from DeepDR and EyeQ. In the DRAC dataset, both Vanilla and MKCNet exhibit desired attention on the high-quality image. However, Vanilla is susceptible to misleading signs caused by image degradations, whereas MKCNet is more robust in handling such degradations. In contrast with OCTA images, Vanilla may disregard vascular structures or lesions that are relevant for diagnosing optic disc conditions in fundus images. Additionally, as shown in the last row, Vanilla may be easily distracted by large black areas around fundus images, whereas MKCNet focuses on anatomical structures or lesions instead of artifacts. Overall, MKCNet performs well in evaluating both OCTA and fundus images.

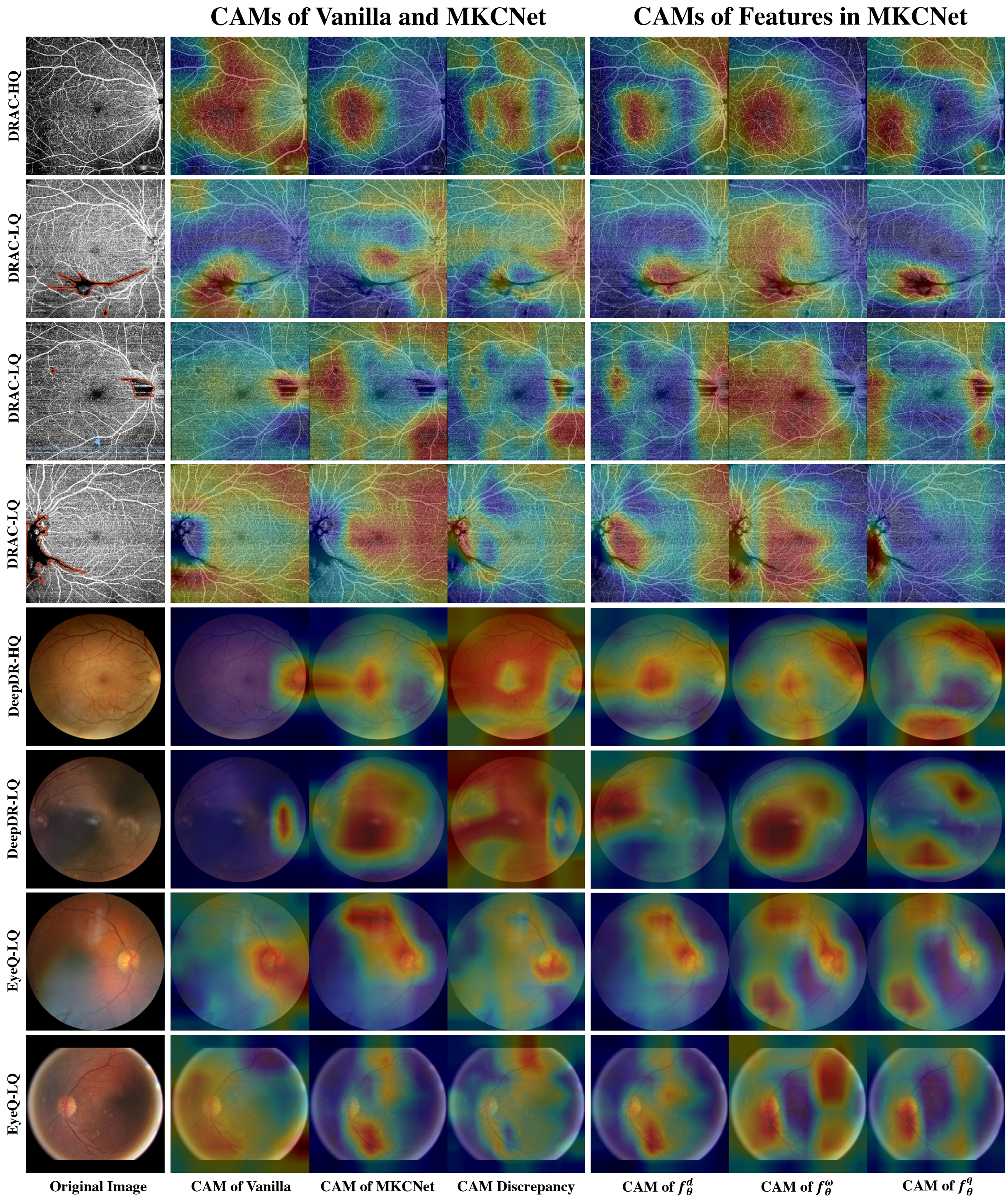

Figure 1. Qualitative analysis via CAM visualization on DRAC, EyeQ, and DeepDR.

References

[1] DRAC. The diabetic retinopathy analysis challenge website. https://drac22.grand-challenge.org. Accessed October 20, 2022. 1

[2] Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, et al. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, page 100512, 2022. 1

[3] Huazhu Fu, Boyang Wang, Jianbing Shen, Shanshan Cui, Yanwu Xu, Jiang Liu, and Ling Shao. Evaluation of retinal image quality assessment networks in different color-spaces. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 48–56. Springer, 2019. 1

[4] Eduardo Soares, Plamen Angelov, Sarah Biaso, Michele Higa Froes, and Daniel Kanda Abe. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. *MedRxiv*, 2020. 1

[5] Jinyu Zhao, Yichen Zhang, Xuehai He, and Pengtao Xie. Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865*, 490, 2020. 1

[6] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018. 1

[7] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2097–2106, 2017. 1

[8] Liming Xu, Xianhua Zeng, Zhiwei Huang, Weisheng Li, and He Zhang. Low-dose chest x-ray image super-resolution using generative adversarial nets with spectral normalization. *Biomedical Signal Processing and Control*, 55:101600, 2020. 1

[9] Kang Zhou, Zaiwang Gu, Wen Liu, Weixin Luo, Jun Cheng, Shenghua Gao, and Jiang Liu. Multi-cell multi-task convolutional neural networks for diabetic retinopathy grading. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2724–2727. IEEE, 2018. 2

[10] Ziyuan Zhao, Kerui Zhang, Xuejie Hao, Jing Tian, Matthew Chin Heng Chua, Li Chen, and Xin Xu. Bira-net: Bilinear attention net for diabetic retinopathy grading. In *2019 IEEE International Conference on Image Processing*, pages 1385–1389. IEEE, 2019. 2

[11] Shaoteng Liu, Lijun Gong, Kai Ma, and Yefeng Zheng. Green: a graph residual re-ranking network for grading diabetic retinopathy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 585–594. Springer, 2020. 2

[12] Along He, Tao Li, Ning Li, Kai Wang, and Huazhu Fu. Cabnet: category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging*, 40(1):143–153, 2020. 2

[13] Kang Zhou, Zaiwang Gu, Annan Li, Jun Cheng, Shenghua Gao, and Jiang Liu. Fundus image quality-guided diabetic retinopathy grading. In *Computational Pathology and Ophthalmic Medical Image Analysis*, pages 245–252. Springer, 2018. 2

[14] 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. 2

[15] Qingyu Chen, Yifan Peng, Tiarnan Keenan, Shazia Dharssi, Elvira Agro, Wai T Wong, Emily Y Chew, Zhiyong Lu, et al. A multi-task deep learning model for the classification of age-related macular degeneration. *AMIA Summits on Translational Science Proceedings*, 2019:505, 2019. 2

[16] Lihao Liu, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Multi-task deep model with margin ranking loss for lung nodule analysis. *IEEE Transactions on Medical Imaging*, 39(3):718–728, 2019. 2

[17] Shikun Liu, Andrew Davison, and Edward Johns. Self-supervised generalisation with meta auxiliary learning. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[18] Haoxuan Che, Haibo Jin, and Hao Chen. Learning robust representation for joint grading of ophthalmic diseases via adaptive curriculum and feature disentanglement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 523–533. Springer, 2022. 2

[19] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2

[20] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2020. 2

[21] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2019. 2

[22] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13025–13032, 2020. 2