Supplementary Material of MRM for Medical Image Pre-Training with Genetics

Qiushi Yang[1] Wuyang Li[1] Baopu Li[3] Yixuan Yuan[1,2*]

[1]City University of Hong Kong [2]The Chinese University of Hong Kong [3]Independent Researcher

{qsyang2-c, wuyangli2-c}@my.cityu.edu.hk. bpli.cuhk@gmail.com yxyuan@ee.cuhk.edu.hk

In the supplementary material, we firstly introduce the data pre-processing in Section A. Then, we provide the implementation details of MRM in Section B. The comparison on the visualization of reconstructed images is exhibited in Section C. Finally, we discuss the broader impact and limitations of our work in Section D.

A. Data Pre-processing

In this section, we introduce the data pre-processing including the dataset pre-processing, image pre-processing and genetics pre-processing.

A.1. Dataset Pre-processing

To achieve a fair comparison, we follow previous works [17, 20] and employ dataset pre-processing techniques as following.

A.1.1 UK Biobank Genetic Modalities

During the pretraining phase using UK Biobank data, we choose the following feature dimensions. For the raw SNPs, we uniformly sample every 100^th SNP from 22 Chromosomes (excluding the X and Y chromosomes), resulting in 7,854 SNPs per sample. For PGS, we used 481 scores for a wide variety of different traits downloaded from the PGS Catalog [16]. We created burden scores for 18,574 protein-coding genes [17].

A.1.2 Diabetic Retinopathy detection (APTOS)

For diabetic retinopathy detection, we adopt the APTOS 2019 Blindness Detection dataset [1], containing 3,662 retinal fundus images with five categories including five levels of disease severity. Note that the categories are mutually inclusive, for instance, class three also includes the levels of two and one. Therefore, we formulate the task as a multi-class classification task by employing multi-hot form for all labels. For example, class three is encoded as [1,1,1,0,0] and two as [1,1,0,0,0]. We divide the whole dataset into 80% as the training set and 20% as the test set. The Quadratic Weighted Kappa (QwKappa) [6] is adopted as the metric to measure the agreement between the prediction and ground truth.

A.1.3 Retinal Fundus Disease Classification (RFMiD)

For retinal fundus disease classification, we utilize the Retinal Fundus Multi-disease Image Dataset (RFMiD) [13] including 3,200 images across 45 disease classes. Following prior work [17], we exclude two classes (”HR” and ”ODPM”) since they have no positive cases, and use the remaining 43 classes for training and evaluation. As mentioned before, we convert these classes to multi-hot labels and solve the task as multilabel classification. We randomly split 80% as the training set and 20% as the test set. The area under the ROC curve (ROC-AUC) is used as the metric to evaluate the classification results.

A.1.4 Pathological Myopia Segmentation (PALM)

We use the Pathologic Myopia challenge dataset [8] for pathological myopia segmentation, consisting of 400 images with segmentation masks, which contains three categories including peripapillary atrophy (311 cases), optic disc (all cases), and detachment (12 cases). Considering that the detachment is much rarely available, we ignore it and merely consider two classes, i.e., the peripapillary atrophy and optic disc to fine-tune the

[*] Corresponding author.

This work was supported by Hong Kong Research Grants Council (RGC) General Research Fund 14220622, 14204321, and Innovation and Technology Commission-Innovation and Technology Fund ITS/100/20.

model. We use 800 images as the training split and 400 images as test split. The dice score is adopted as the segmentation evaluation metric.

A.1.5 Cardiovascular Risk Prediction (UKB)

To predict the cardiovascular risk factors of (sex, age, BMI, SBP, DBP, smoking status) from retinal fundus images, we utilize 102,219 images from the UKB [16] database, which is randomly split 80% as training set and 20% as test set. Two models are leveraged to train the task. The first model aims to perform the classification task (sex and smoking status to binary classification). The second one targets to predict the continuous variables including age, BMI, SBP, and DBP, formulated as a regression task. Due to different scales of loss values, we adopt two models for two tasks respectively. We normalize the values of continuous factors by standardization (normalizing to zero mean and scaling to unit variance). We finally impute the missing values of the factors by adopting the mean value for continuous factors and median value for discrete factors.

A.1.6 TCGA

For the pathology images-based pre-training and transfer evaluation, we use TCGA-GBM with TCGA-LGG dataset [18] to conduct the pre-training, which consists of 736 paired samples of pathology slides and genetic profiles. We resize the curated pathology slides with the shape of 224 × 224 as inputs. Considering each patient has multiple curated slides, we select one of them associated with one genetic profile as an input pair. During the phase of transfer evaluation on downstream dataset, we leverage glioma grading (GG) to evaluate the performance. The GG can improve the treatment planning for accurate determination. The TCGA dataset contains WHO grading labels including grade II, III and IV. We fine-tune the pre-trained model on training set with 80% data split and evaluate the performance on 20% data split. The accuracy and ROC-AUC are employed as the metrics to measure the classification results.

A.2. Image Pre-processing

A.2.1 Image Quality Control

The UK Biobank contains many retinal fundus images with bad quality (e.g. completely black or extremely overexposed). We conduct two steps of quality control to filter out the outliers. In the first step, we only consider images in which a simple circle-detection algorithm [**?**] can find a circle. Then, we discard the top and bottom 0.5% brightest and darkest remaining images.

A.2.2 Image transformations

The images are cropped to the circles detected by the stage of image quality control, and reshape to 224×224. In the pre-training phase, we randomly transform images by a rotation of up to 20° and flip the image horizontally with a 50% probability. We also follow the common practice of normalizing all image intensities via the mean and standard deviation from ImageNet [15].

A.3. Genetics Pre-processing

A.3.1 Raw SNPs

The raw SNPs are a cross section of all SNPs collected on microarray chips, collecting above 800k genetic variants in total across all chromosomes. More information on data collection can be found at https://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=263. We encode each SNP in another way: 0 represents no difference from the reference genome, 1 indicates one copy of the chromosome differs and the other does not, and 2 means both copies of the chromosome differ. We use SNPs as continuous variables and fill in missing values with mode imputation. To reduce the number of feature dimensions from 800k, which are hard to handle, we only take every 100-th SNP from the full microarray since SNPs are strongly correlated with the genome. We also exclude SNPs on the sex chromosomes as they need special statistical treatment and they are not common between genetic males and females. Hence, we leverage 7,854 SNPs in our models.

A.3.2 Polygenic Risk Scores

For computing polygenic risk scores, we downloaded all PGS weight files included in the PGS Catalog, a collection of published PGS. The PGS files provide weights for a linear model to compute risk scores from the raw genetic data. To have a large intersection of available SNPs for our UKB population and the weights provided by the PGS catalog, instead of using the raw microarray data from Appendix A.3.1, we used imputed data. The imputed data uses prior knowledge about correlations between SNPs collected and not collected on the respective microarray ("linkage disequilibrium", LD) to infer the missing features with high accuracy. Imputed data was precomputed by the UKB, and more information can be found at https://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=100319. Using the imputed data, we computed 481 polygenic scores for our cohort using the PLINK software [21], ignoring scores that gave errors or that only recorded genome positions in a different reference genome build.

We obtain all PGS weight files from the PGS Catalog, which are drawn from the published PGS. The PGS files provide weights for a linear model to calculate risk scores from the raw genetic data. To leverage more SNPs that are available for our UKB population and the weights from the PGS catalog, instead of the raw microarray data, we utilize imputed data, which adopts previous knowledge about relations among SNPs collected and not collected on the respective microarray to predict the missing features with high accuracy. The UKB pre-processes the imputed data, and more information can be found at https://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=100319. With the imputed data, we calculate 481 polygenic scores for our cohort using the PLINK software [17], and neglect the scores that have errors or only using genome positions in a different reference genome build.

A.3.3 Burden Scores

We utilize the Functional Annotation and Association Testing Pipeline to add functional information to all the genetic variants in the UK Biobank 200k exome sequencing release [16]. We make burden scores for all protein coding genes from protein loss of function and missense variants that are likely to be harmful. We only focus on rare variants with minor allele frequencies below 1%. For these variants 41% are ”singletons”, i.e. they merely appear once in our sample. We gave each participant a binary vector of length 18,574 that match the number of protein coding genes. For each gene, the vector entry is 1 if the participant has at least one possibly harmful variant in that gene, or 0 if no possibly harmful variants are seen in that gene for that participant. This coding helps rare-variant association studies to combine the effects of many rare variants within genes, where it can increase statistical power and release the computational cost of multiple testing.

B. Implementation Details

Following prior works [7, 20], we adopt ViT-base as the image encoder and SNN network as the genome encoder to learn representations. The ViT and SNN models are trained via AdamW [12] and Adam [10], respectively, both with an initial learning rate of 1×10^{-3}. We use PyTorch [14] to implement our models, and train all models for 50 epochs with the batch size of 256 for UKB and 8 for TCGA. In relation matching, for efficiency, we randomly select 8 pair of multimodal features to perform matching constraint for two datasets. All comparisons [3, 4, 17, 7, 19, 2, 5, 11, 9] share the same settings to achieve a fair comparison. The balanced coefficient λ is 1.0, and the masking ratios τ_I, τ_G for images and genetics are set as 75% and 50%, respectively.

C. Visualization of the Reconstructed Images

To qualitatively verify the effectiveness of relation masking, we visualize the reconstruction images of four samples from different methods. Figure 1 illustrates that MIM-based approaches [7, 19, 2, 5, 11, 9] lose the tiny disease regions, while MRM can reconstruct almost complete disease regions. These advantages indicate that our relation masking in self- and cross-modality levels can preserve the disease semantics thereby improving the quality of the feature representation.

D. Broader Impact and Limitations

The proposed MRM framework, consisting of relation masking and relation matching, can enable the model to capture relation information by token-wise feature masking and sample-wise global relation constraint, thereby learning better feature representation. Extensive experiments across various downstream diagnosis tasks demonstrate that the MRM has superior transfer ability over state-of-the-art methods, and it can also applies single image modality pre-training to achieve compelling performance. Moreover, in this work, we assume the data distribution between downstream datasets and the pre-training dataset is identical, and do not consider the issue of data domain shift. Hence, in future work, we will improve our framework towards data domain shift issue between pre-training dataset and various downstream datasets.

References

[1] Aptos 2019 blindness detection. https://www.kaggle.com/c/aptos2019-blindnessdetection/ Accessed: 2021-11-04. 1

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Proc. ECCV*, pages 348–367. Springer, 2022. 3, 4

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[4] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision trans-

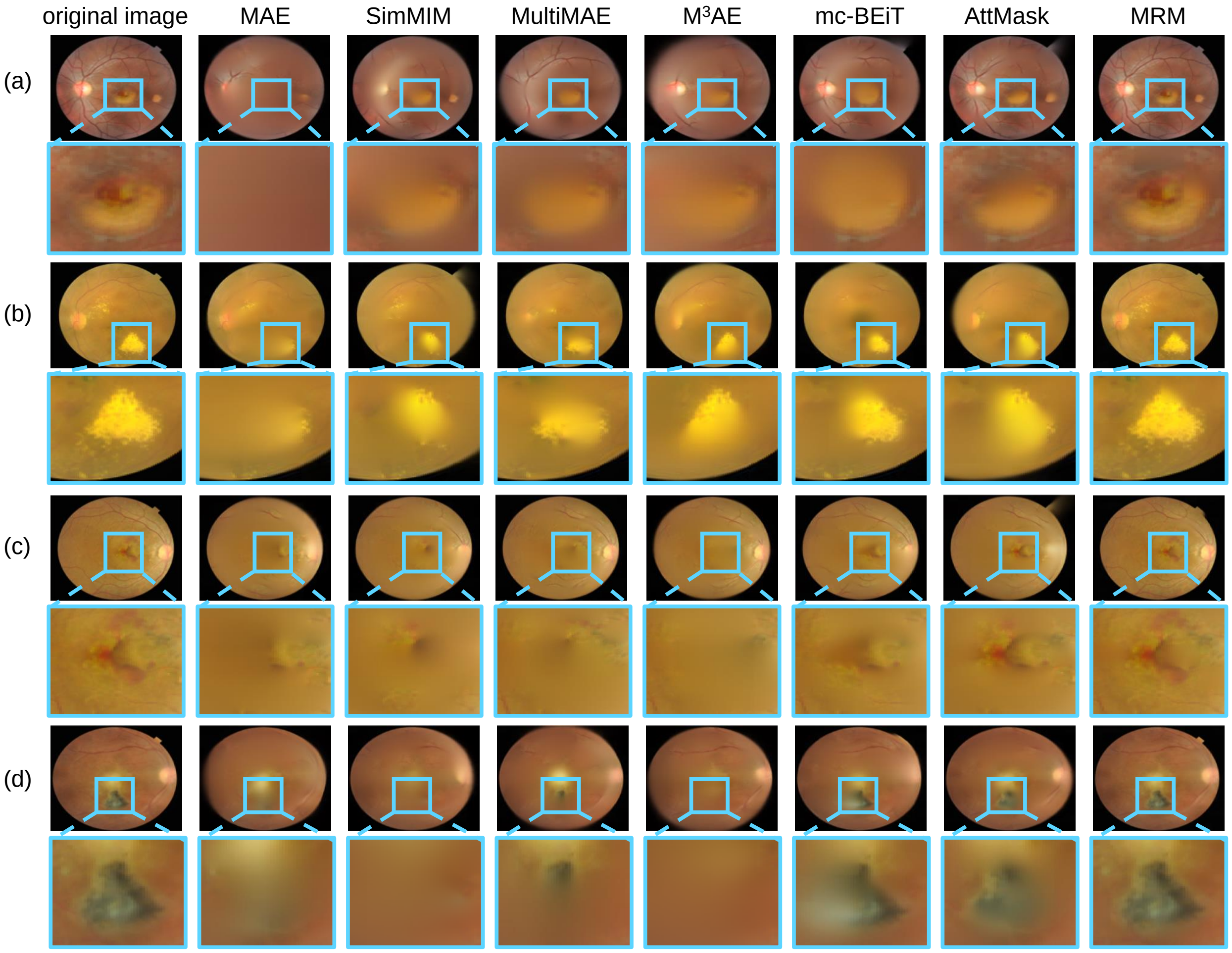

Figure 1. **Comparison of reconstruction results of different methods.** From left to right suggests the original input and the reconstructed images by MAE [7], SimMIM [19], MultiMAE [2], M^{3}AE [5], mc-BEiT [11], AttMask [9] and our MRM. We can observe that MRM can preserve the disease regions framed in blue while MIM-based methods lose them.

formers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 3

[5] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *Proc. MICCAI*, pages 679–689. Springer, 2022. 3, 4

[6] Jacob Cohen. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull*, 70(4):213, 1968. 1

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. CVPR*, pages 16000–16009, 2022. 3, 4

[8] Jose Ignacio Orlando Hrvoje Bogunovic Xu Sun Jingan Liao Yanwu Xu Shaochong Zhang Huazhu Fu, Fei Li and Xiulan Zhang. Palm: Pathologic myopia challenge. 2019. 1

[9] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *Proc. ECCV*, pages 300–318. Springer, 2022. 3, 4

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proc. ICLR*, 2015. 3

[11] Xiaotong Li, Yixiao Ge, Kun Yi, Zixuan Hu, Ying Shan, and Ling-Yu Duan. mc-beit: Multi-choice discretization for image bert pre-training. In *Proc. ECCV*, pages 231–246. Springer, 2022. 3, 4

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *Proc. ICLR*, 2019. 3

[13] Samiksha Pachade, Prasanna Porwal, Dhanshree

Thulkar, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Luca Giancardo, Gwenolé Quellec, and Fabrice Mériaudeau. Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research. *Data*, 6(2):14, 2021. 1

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 3

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 2

[16] Hannah Spitzer, Kai Kiwitz, Katrin Amunts, Stefan Harmeling, and Timo Dickscheid. Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In *Proc. MICCAI*, pages 663–671. Springer, 2018. 1, 2, 3

[17] Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In *Proc. CVPR*, pages 20908–20921, 2022. 1, 3

[18] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemp. Oncol.*, 2015(1):68–77, 2015. 2

[19] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proc. CVPR*, pages 9653–9663, 2022. 3, 4

[20] Xiaohan Xing, Zhen Chen, Meilu Zhu, Yuenan Hou, Zhifan Gao, and Yixuan Yuan. Discrepancy and gradient-guided multi-modal knowledge distillation for pathological glioma grading. In *Proc. MICCAI*, pages 636–646. Springer, 2022. 1, 3