MRM: Masked Relation Modeling 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

Abstract

Modern deep learning techniques on automatic multimodal medical diagnosis rely on massive expert annotations, which is time-consuming and prohibitive. Recent masked image modeling (MIM)-based pre-training methods have witnessed impressive advances for learning meaningful representations from unlabeled data and transferring to downstream tasks. However, these methods focus on natural images and ignore the specific properties of medical data, yielding unsatisfying generalization performance on downstream medical diagnosis. In this paper, we aim to leverage genetics to boost image pre-training and present a masked relation modeling (MRM) framework. Instead of explicitly masking input data *in previous MIM methods leading to loss of disease-related semantics, we design relation masking to mask out token-wise* feature relation *in both self- and cross-modality levels, which preserves intact semantics within the input and allows the model to learn rich disease-related information. Moreover, to enhance semantic relation modeling, we propose relation matching to align the sample-wise relation between the intact and masked features. The relation matching exploits inter-sample relation by encouraging global constraints in the feature space to render sufficient semantic relation for feature representation. Extensive experiments demonstrate that the proposed framework is simple yet powerful, achieving state-of-the-art transfer performance on various downstream diagnosis tasks. Codes are available at* https://github.com/CityU-AIM-Group/MRM.

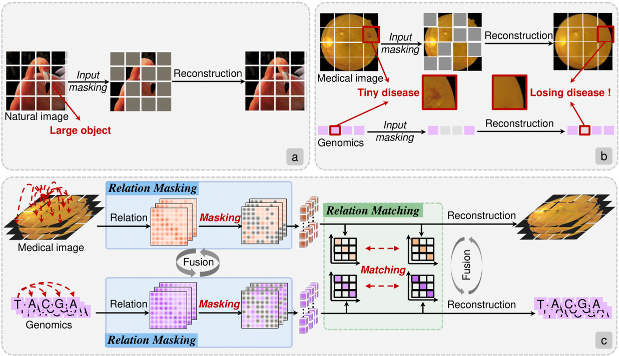

Figure 1. **Comparison of different masking strategies on natural and medical data.** **(a)**: Existing MIM methods mask out input natural images and infer the missing content to learn semantic representations via reconstruction task. **(b)**: Recent pre-training approaches for disease diagnosis explicitly employ MIM on input medical data (e.g., medical images and genome), whilst they are prone to lose tiny disease regions and cause non-tractable reconstruction. **(c)**: Our method masks token-wise feature relation across multimodal data and matches sample-wise relation between the intact and masked features, preserving intact semantic regions and enriching relation information.

1. Introduction

In the medical diagnosis [44, 34, 38, 28, 6, 5], the large-scale multimodal biobank data, *e.g.*, images and genetics, is necessary for a reliable diagnosis, overcoming the limited scale and disease information of a single-modality dataset. However, the prohibitive expert annotations of large-scale datasets make it difficult to train a conventional deep model [44, 45, 41, 46, 47]. Particularly, in this multimodal scenario, the requirements of the experts in various medical fields prevent enough annotation access, severely limiting the grounding of the automatic diagnosis system.

To address this, the most prospering trend [16, 43, 4, 34, 13, 35, 36, 12, 27] is self-supervised pre-training, *e.g.*, masked image modeling (MIM) [16, 43, 13, 42], aiming to train a label-free model with adequate generalization capacity. Existing MIM methods [16, 43, 4, 25, 13, 42, 49] mask out a high portion of patches within input images and infer the missing content, as suggested in Figure 1 (a). They make use of contextual information to glance the semantics and reconstruct the entire images, which performs a mask-and-reconstruct task to pre-train the model without annotation and transfer meaningful representations to various

[*] 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.

downstream tasks for improving label-efficient fine-tuning.

Though achieving great success, most works [16, 43, 4, 40, 10, 25, 19] are designed for natural images, ignoring the essential differences between medical data and natural images. Therefore, we empirically find that existing MIM CANNOT works well in medical data (see Table 1), and even totally fails to reconstruct diseases (see Figure 3). The reasons derive from a critical observation regarding the significant data differences, which can be summarized into two challenges. Firstly, compared with natural images, there are **limited semantic regions** in medical data. As shown in Figure 1 (a), the semantic-rich foreground is always the main body of the natural image, while the rest non-informative background region only represents a minority part. Differently, in the medical image (Figure 1 (b)), the majority of regions are backgrounds, and the informative disease regions are usually on a tiny scale. Under the strategy of masking entire tokens in existing MIM methods [16, 43, 4, 40, 10, 25, 19], if the disease tokens are masked out, the disease-related semantic is totally missing with a *catastrophic information loss*, leading to a non-tractable reconstruction. This issue also exists between genomics and natural images. The semantic regions in genomics, *i.e.*, the disease-related patterns, mainly lie in a minority of genome segments [28, 5, 7]. Hence, instead of masking the whole input token, these observations motivate us to delve into the masking of token-level relation, which preserves abundant semantic discriminability and adequate self-supervision, as illustrated in Figure 1 (c) Left.

The second challenge is the **limited semantic relation**. In a natural image, the background and foreground relationship, *e.g.*, the *bird* in the *sky* and the *person* in a *room*, tends to be prosperous and abundant, serving a critical role in semantic-level learning [39, 24]. In contrast, in each medical data sample, the disease-aware relation is limited and insufficient to provide enough discriminative evidence. The reason lies in that the medical datasets are usually collected from the same human organ, *e.g.*, the *fundus*, containing redundant and similar anatomical patterns, *e.g.*, the *capillary*, which severely prevent the relation modeling between the disease and the complex medical scene. This challenge hampers reliable relation learning in existing MIM methods, and may incur an inevitable overfitting to non-informative relation within the background [40, 25, 23, 3]. Hence, considering the limited semantic relation in each data sample, we are committed to going beyond the self-supervised learning for an independent and individual data sample, and propose to encourage global constraints for exploiting inter-sample relation (see Figure 1 (c) Right).

To combat above challenges, as shown in Figure 1 (c), we present MRM, a masked relation modeling from a unified view of *relation*, containing **relation masking** and **relation matching**, to rationally pre-train multimodal medical images with genetics. To preserve intact semantic information within raw input, we devise relation masking strategy to allow the model to learn disease-related semantics. Instead of masking out input data, the relation masking investigates *token-wise relation* within feature representations in both self- and cross-modality levels and masks out the relation among all multimodal tokens. The relation masking endows the model to explicitly learn global dependency from raw data without missing disease-related semantic information. Furthermore, to improve the semantic relation modeling, the relation matching is designed to provide global constraint by aligning the feature relation across multiple samples. Specifically, relation matching exploits *sample-wise relation* in both self- and cross-modality levels to encourage the relation consistency between the intact and masked features. This enjoys the complementary advantages of per-sample pixel-wise reconstruction loss and boosts the transfer ability of the model. With the pre-trained model, we can obtain the feature representation that can be transferred to supervised downstream diagnosis tasks for boosting label-efficient fine-tuning, alleviating the severe demand for specialized annotations.

In summary, our contributions fall into four parts:

- We identify the challenges of current MIM methods on medical data, and present MRM, a masked relation modeling using multimodal medical data to facilitate image representation learning.
- Towards the issue of limited semantic regions in medical data, we design relation masking to mask token-wise feature relation across self- and cross-modality. Different from MIM explicitly masking inputs, relation masking preserves disease semantics within inputs, endowing a powerful mask-and-construct task.
- Moreover, to enrich the semantic relation among diseases, the relation matching is proposed to capture abundant disease-related relation by aligning sample-wise feature relation between intact and masked features in both self- and cross-modality levels.
- Extensive transfer evaluation on various downstream tasks using two public medical pre-training datasets demonstrate that our framework performs superior transfer ability over state-of-the-art methods.

2. Related Work

2.1. Visual Pre-training

In recent years, self-supervised visual pre-training [8, 9, 15, 17, 11, 16, 43, 48, 4] has achieved exceptional transfer performance in various downstream visual tasks, which

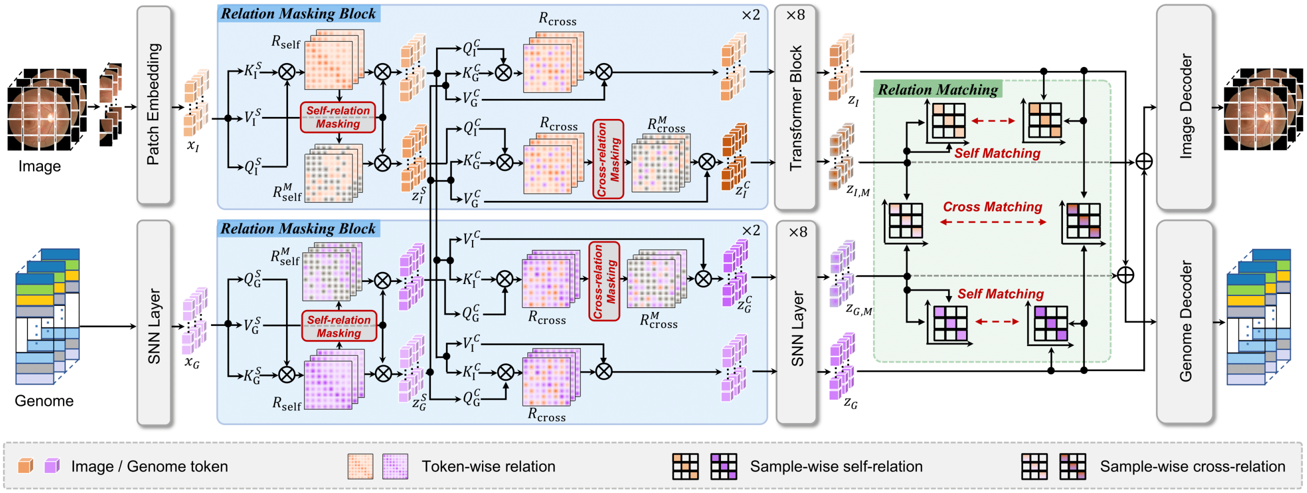

Figure 2. **Overview of masked relation modeling (MRM).** MRM contains relation masking to mask out token-wise feature relation with preserving disease-related semantics, and the relation matching to enforce sample-wise relation consistency for global semantic modeling.

can be divided into two categories. The first is *contrastive-based* methods [8, 15, 17, 35, 35] that aim to project similar samples nearby while pull dissimilar ones far away in the feature space. For instance, SimCLR [8] imposes two embeddings from a sample under different views consistency via InfoNCE loss, and BYOL [15] merely pushes the positive pairs together with a momentum encoder for boosting transfer performance. Moreover, inspired by the progress of masked language modeling on natural language processing [20], *masked image modeling (MIM)-based* approaches [16, 48, 43, 4, 3, 10, 3] are proposed to pre-train visual models and exhibit remarkable transfer performance. MAE [16] formulates a pixel-wise reconstruction task to mask out a high portion of input patches and infer the masked ones with visible tokens via an autoencoder. Afterwards, many followups [43, 3, 19, 10] devise MIM models to improve the representation learning. However, they mainly focus on natural images and ignore the essential differences between the medical data and natural images, thereby delivering unsatisfying transfer results [27, 49].

2.2. Medical Self-supervised Learning

Most works [35, 36, 2, 34] on medical self-supervised learning extend the ideas of contrastive learning and develop customized strategies to pre-train models. 3D-SSL [36] adopt a series of proxy tasks to present a 3D versions self-supervised model and achieves compelling transfer results. ContIG [34] crafts a multimodal pre-training framework using image and genome data via contrastive learning and delivers competitive transfer ability. Although efficiency, these contrastive-based approaches strongly rely on well-defined data augmentations to construct positive pairs, which are difficult for medical data, especially for structured data, *e.g.*, genetics. Most recently, some studies [13, 42, 27, 49] adopt MIM to self-supervised pre-train medical models. MedicalMAE [49] introduces a self pre-training paradigm by employing MAE [16] to pre-train the model and achieve superior performance. To perform multimodal pre-training, M^{3}3AE [13] randomly masks out the tokens of image and language inputs and aims to infer the whole data using visible tokens. All of them still adopt input masking strategy, which may incur tiny disease-related region missing and non-tractable reconstruction. Differently, we design a novel relation masking strategy in the feature space and retain intact disease-related information.

2.3. Learning from Images and Genetics

Genetics can provide comprehensive information for many disease diagnosis and treatment planning [38, 28, 6, 5, 7], which attract much attention from researchers. Many works [44, 34, 38] formulate a multimodal learning task by jointly leveraging genetics and medical images to exploit complementary knowledge. Considering that the genetics may be difficult to acquire in many clinical applications, recent studies [34, 7] aim to design pre-training models using images and genetics for improving practical image-based diagnosis. However, they merely use contrastive constraint to perform self-supervised learning, while they fail to adopt the efficient mask-and-reconstruct task to pre-train models. In this work, we aim to leverage genetics with images to boost the visual representation learning via semantic relation modeling, which discovers both inner-sample and inter-sample relation to achieve more effective pre-training.

3. Masked Relation Modeling

In this section, we start by introducing the preliminaries of vision Transformer and masked image modeling. We then present the masked relation modeling (MRM), containing relation masking and relation matching. The overall pre-training and transfer inference of the framework are described at the end of the section.

3.1. Preliminaries

Vision Transformer (ViT). Given an image $x \in \mathbb{R}^{c\times h\times w}$, ViT first splits it into $n = hw/p^2$ non-overlapping patches $x_P = \{x_1; x_2; ...; x_n\}, x_i \in \mathbb{R}^{c\times p\times p}$, where $p \times p$ denotes the size of each patch. Afterwards, the patches x_P are tokenized as a sequence of token embeddings $e = \{e_1; e_2; ...; e_n\}, e_i \in \mathbb{R}^{1\times d}$, where d is the dimension of each token, via a linear projection with positional embeddings. The embeddings e are then processed by a cascade of multi-head self-attention layers to capture global information, where the self-attention is described as follow:

$$\text{Attention}(e) = \text{Softmax}(\frac{1}{\sqrt{d}}QK^T)V, \tag{1}$$

where $K = eW^K$, $Q = eW^Q$, $V = eW^V$, $K, Q, V \in \mathbb{R}^{1\times d}$ is the key, query and value obtained by the linear projection via W^K, W^Q, W^V, respectively. Each layer of the ViT encoder contains a multi-head self-attention block, and the feature representation is produced by multiple layers within the encoder.

Masked Image Modeling (MIM). The input patches x_p are randomly masked out with a high ratio, denoted as masked patches x_M. The remaining visible patches x_V are fed into ViT encoder $f(\cdot)$ to capture features $z = f(x_V)$, which are then sent to a lightweight decoder $h(\cdot)$ to recover the missing patches $\hat{x} = h(z)$ of the input data. The overall framework is optimized by mean-squared error (MSE) reconstruction loss function $\mathcal{L}_{\text{MIM}}$ between the recovered patches $\hat{x}$ and masked patches x_M as:

$$\mathcal{L}_{\text{MIM}} = \mathbb{E}||x_M - h(f(x_V))||_2^2. \tag{2}$$

Overview of Masked Relation Modeling (MRM). As illustrated in Figure 2, the proposed MRM consists of the relation masking strategy to mask out feature relation and preserve intact disease-related semantics, and the relation matching to offer global constraints for relation modeling. With the input image x_I and genome x_G, the ViT encoder f_I and self-normalizing network (SNN) [22] f_G produces masked feature representations $z_{I,M}, z_{G,M}$ for image and genome via relation masking. Parallelly, the intact representations z_I, z_G are obtained by two encoders without relation masking. Afterwards, the masked features $z_{I,M}, z_{G,M}$ are aggregated with the intact features z_G, z_I from the other modality to yield the fused features z_I^F, z_G^F, respectively. These fused feature are then put into the image decoder h_I and the genome decoder h_G to reconstruct original data $\hat{x}_I$ and $\hat{x}_G$. The relation matching is employed on intact and masked feature representations with the data reconstruction loss to jointly optimize the overall framework.

3.2. Relation Masking

Due to the medical data usually appearing tiny disease patterns, current input-level masking strategies may discard disease-related semantics and hardly learn informative feature representation. To address this problem, the relation masking is proposed to mask token-wise feature relation from self- and cross-modality perspectives in a cascading manner, which can maintain intact input disease-related information during pre-training. The operations of relation masking for image and genome are parallel, for brevity, we adopt the image branch to introduce.

Self-modality relation masking. In the i-th relation masking block, given z_I^0 as the input image token, we first calculate K_I^S, Q_I^S, V_I^S as its key, query and value, respectively. Then, the normalized feature dependency in tokens is obtained by computing the relation between the key K_I^S and query Q_I^S of the image feature as follows:

$$R_{\text{self}} = \text{Softmax}(\frac{1}{\sqrt{d}}Q_I^S \cdot (K_I^S)^T). \tag{3}$$

This self-modality relation R_{self} reflects the token-wise semantic correspondence within the image feature. The tokens with stronger relation contain more interactive and informative semantics. To enable the model to capture intra-modality feature representation via the reconstruction task, on the top of self-modality relation, we aim to break the strong dependency across informative regions by relation masking. Specifically, we mask out high-intensity elements with a ratio τ_I in self-modality relation:

$$R_{\text{self}}^M[p,q] = \begin{cases} R_{\text{self}}[p,q] & \text{if } R_{\text{self}}[p,q] < r_{\text{self}}^p \\ 0 & \text{else,} \end{cases} \tag{4}$$

where r_{self}^p refers to the intensity of the top-τ_I element among the p-th row of the self-modality relation R_{self} matrix. Note that the masked elements are selected in each row of the self-modality relation matrix to remove the semantic correspondence for all tokens with the same intensity. Afterwards, the masked relation matrix R_{self}^M is forwarded to produce the token $z_I^S = R_{\text{self}}^M V_I^S$ for subsequent data flow. With the self-modality relation masking, the relative important dependency within the modality is discarded and we enforce the model to reconstruct the original image with the remaining weak feature relation. Thus, the model can capture disease-related feature representation of images.

Cross-modality relation masking. In an attempt to incorporate multimodal knowledge, after the self-modality self-attention layer, we perform cross-modality attention between the image and genome features. Precisely, obtained the image and genome tokens z_I^S, z_G^S masked via self-modality relation masking, we compute token-wise cross-modality relation to capture semantic correspondence between image and genome features:

$$R_{\text{cross}} = \text{Softmax}(\frac{1}{\sqrt{d}} Q_I^C \cdot (K_G^C)^T), \tag{5}$$

where K_G^C means the key of genome feature z_G^S, and Q_I^C represents the query of image feature z_I^S in the cross-attention layer. This cross-modality attention can amalgamate two modality knowledge and improve the feature representation towards disease-related information. Then, to allow the model to learn complementary multimodal knowledge and the relation between two modalities, we perform cross-modality relation masking by removing strong semantic correspondence between image and genome features:

$$R_{\text{cross}}^M[p,q] = \begin{cases} R_{\text{cross}}[p,q] & \text{if } R_{\text{cross}}[p,q] < r_{\text{cross}}^p \\ 0 & \text{else,} \end{cases} \tag{6}$$

where r_{cross}^p denotes the intensity of the top-τ_I element among the p-th row of the cross-modality relation matrix R_{cross}. With the masked relation R_{cross}^M, we can obtain the fused image token $z_I^C = R_{\text{cross}}^M V_G^C$, where V_G^C means the value of the genome feature.

Reconstruction with relation masking. We feed the input data into two networks shared parameters, where the first network consists of a ViT encoder to yield intact image feature z_I and a SNN encoder with self-attention blocks to produce genome feature z_G. Parallelly, the second network employs the proposed self- and cross-modality relation masking in first two attention blocks to generate masked features $z_{I,M}$, $z_{G,M}$ for image and genome, respectively. Afterwards, the masked features are incorporated with the intact feature from the other modality and produce amalgamated features $z_I^F = \text{concat}[z_{I,M}; z_G]$, $z_G^F = \text{concat}[z_{G,M}; z_I]$ for images and genome. The fused features are then put into the decoders $h_I(\cdot), h_G(\cdot)$ to reconstruct image and genome $\hat{x}_I = h_I(z_I^F)$, $\hat{x}_G = h_G(z_G^F)$. The overall loss function for the reconstruction of image and genome is:

$$\mathcal{L}_{\text{recon}} = ||h_I(z_I^F) - x_I||_2^2 + ||h_G(z_G^F) - x_G||_2^2. \tag{7}$$

The relation masking strategy exploits the token-wise feature relation for mask-and-reconstruct task. It is worth noting that although the strong relation is removed, the intrinsic information within the data is retained [32]. Therefore, with the original complete image and genome as inputs, our relation masking can preserve the intact disease-related semantics. With the relation masking-based reconstruction task, the self-modality relation is encouraged to be recovered by the model to capture disease-related information within each modality, and the cross-modality relation is enforced to reconstruct to learn abundant multimodal knowledge to improve the disease-related representation for effective downstream diagnosis tasks transfer.

3.3. Relation Matching

Considering that the disease-aware relation among diseases is limited in medical data, to provide sufficient semantic relation, we propose relation matching, a global constraint to align the sample-wise relation across self- and cross-modality samples to perform global constraints in the feature space.

Self-modality matching. Assume one minibatch contains B multimodal pairs $\{x_I^i, x_G^i\}_{i=1}^B$, we can obtain the intact feature representations $\{z_I^i, z_G^i\}_{i=1}^B$ and masked ones $\{z_{I,M}^i, z_{G,M}^i\}_{i=1}^B$ via the proposed relation masking strategy. Firstly, we compute the sample-wise relation $\{R_{\text{II}}^{i,j}\}_{i,j=1}^B, \{R_{\text{GG}}^{i,j}\}_{i,j=1}^B$ across all intact features of image and genome, respectively:

$$\begin{aligned} R_{\text{II}}^{i,j} &= \text{sim}(z_I^i; z_I^j) \\ R_{\text{GG}}^{i,j} &= \text{sim}(z_G^i; z_G^j), \end{aligned} \tag{8}$$

where $\text{sim}(\cdot;\cdot)$ denotes the similarity between two features, here we adopt the cosine similarity. Similarly, we acquire the sample relation among masked features:

$$\begin{aligned} R_{\text{II,M}}^{i,j} &= \text{sim}(z_{I,M}^i; z_{I,M}^j), \\ R_{\text{GG,M}}^{i,j} &= \text{sim}(z_{G,M}^i; z_{G,M}^j). \end{aligned} \tag{9}$$

We aim to ensure the relation consistency between intact and masked features. The matching objective for self-modality feature relation can be formulated as:

$$\mathcal{L}_{\text{self}} = \frac{1}{2B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} ||R_{\text{II}}^{i,j} - R_{\text{II,M}}^{i,j}||_2^2 + ||R_{\text{GG}}^{i,j} - R_{\text{GG,M}}^{i,j}||_2^2. \tag{10}$$

Cross-modality matching. Furthermore, to incorporate multimodal information and bridge the gap of different modalities, we compute the cross-modality relation:

$$\begin{aligned} R_{\text{IG}}^{i,j} &= \text{sim}(z_I^i; z_G^j), \\ R_{\text{IG,M}}^{i,j} &= \text{sim}(z_{I,M}^i; z_{G,M}^j), \end{aligned} \tag{11}$$

where $\{R_{\text{IG}}^{i,j}\}_{i,j=1}^B$ represents the cross-modality relation across intact multimodal features, and $\{R_{\text{IG,M}}^{i,j}\}_{i,j=1}^B$ means the relation across masked ones. We intend to ensure the cross-modality relation to be invariant between masked and intact features. The cross-modality feature relation can be calculated as:

$$\mathcal{L}_{\text{cross}} = \frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} ||R_{\text{IG}}^{i,j} - R_{\text{IG,M}}^{i,j}||_2^2. \tag{12}$$

Table 1. Comparison with state-of-the-art pre-training algorithms via fine-tuning evaluation on four downstream retinal image-based tasks.

Method	Masking Level		APTOS	RFMiD	PALM	CRP	
	Input	Relation	QwKappa ↑	ROC-AUC ↑	Dice-Score ↑	MSE ↓	ROC-AUC ↑
Training from scratch (baseline)	–	–	83.25	92.91	79.44	3.405	58.05
Contrastive-based Pre-training Methods							
SimCLR [8]	–	–	84.50	93.75	73.18	3.426	63.92
MoCo v3 [17]	–	–	85.03	94.02	75.73	3.421	65.33
ContIG [34]	–	–	86.27	94.59	80.25	3.184	73.57
MIM-based Pre-Training Methods							
MAE [16]	✓	✗	84.06	92.71	74.28	3.488	64.29
SimMIM [43]	✓	✗	84.67	93.80	72.73	3.450	66.08
MultiMAE [3]	✓	✗	85.59	94.26	75.92	3.411	67.51
M^{3}AE [13]	✓	✗	86.85	94.72	77.35	3.394	68.39
mc-BEiT [25]	✓	✗	87.15	94.37	77.40	3.372	66.38
AttMask [19]	✓	✗	87.73	94.91	78.59	3.262	69.40
MRM	✗	✓	89.83	96.31	81.45	3.107	75.90

With the self- and cross-modality feature relation, we can formulate the overall relation matching objective as:

$$\mathcal{L}_{\text{match}} = \mathcal{L}_{\text{self}} + \mathcal{L}_{\text{cross}}. \tag{13}$$

The relation matching encourages global constraints in the feature space to provide sufficient semantic relation, which enjoys the complementary advantages of pixel-wise data reconstruction loss in Eq. (7).

3.4. Pre-training and Transfer Inference

In the pre-training phase, the image and genome as multimodal inputs are fed into the model. We adopt the proposed relation masking to generate masked features, and employ the relation matching as the global constraint with the data reconstruction loss to jointly optimize the overall framework:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{recon}} + \lambda \mathcal{L}_{\text{match}}, \tag{14}$$

where λ denotes the balanced coefficient to control contributions of the reconstruction task and the relation matching. During the phase of downstream image-based fine-tuning, we discard the genome branch and utilize the image encoder to extract the feature representation without relation masking. The pre-trained encoder following a randomly initial task-relevant head is fine-tuned on downstream tasks for evaluation.

4. Experiments

We pre-train MRM on two multimodal datasets involving retinal images with genetics, and pathology images with genetics, respectively. To evaluate the quality of the representations learned on retinal images, we transfer the pre-trained model to four downstream retinal image-based tasks. As for the model pre-trained on pathology images, we assess it on a downstream pathology image-based task. The gene-image association analysis is performed to study the relation between images and genetics towards disease diagnosis. Finally, we extend our method to image pre-training to verify the effectiveness on single modality.

4.1. Experimental Setup

4.1.1 Retinal Images with Genetics

Pre-training. We use UKB [33] to conduct the self-supervised pre-training, which is one of the largest multimodal images and genetics datasets. Following the previous work ContIG [34], we leverage the retinal fundus dataset as the pre-training set, which contains 155,238 images, and the genetics including 155,238 Raw-SNP samples, 145,206 PGS samples and 93,216 burden scores.

Downstream Transfer. We introduce four tasks to evaluate and compare the effectiveness of the model pre-training.

1) Diabetic Retinopathy Detection (APTOS). The APTOS [1] is a fundus disease dataset including 35,126 2D images across five categories. All images are resized as 224 × 224 for efficient fine-tuning. We split 80% as the training set to fine-tune the pre-trained MRM model, and the remaining 20% as the test set is used to evaluate the performance. The Quadratic Weighted Kappa (QwKappa) [14] is adopted as the metric to measure the agreement between the prediction and ground truth.

2) Retinal Fundus Disease Classification (RFMiD). The Retinal Fundus Multi-disease Image Dataset (RFMiD) [29] consists of 3200 annotated retinal fundus images of 46 eye diseases. RFMiD formulates a multi-label classification task since the images may contain multiple conditions. We 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.

3) Pathological Myopia Segmentation (PALM). The PALM dataset [18] contains 1200 images with disc and atrophy segmentation annotations. We use 800 images as the training split and 400 images as test split. The dice score is adopted as the segmentation evaluation metric.

4) Cardiovascular Risk Prediction (CRP). The retinal fundus images of UKB dataset can also be utilized to CRP including age, sex, smoking status, systolic and diastolic blood pressure (SBP, DBP), and body mass index (BMI) [31], formulating a regression with classification

Table 2. Transfer results on downstream pathological image task.

Method	GG	
	Acc ↑	ROC-AUC ↑
Training from scratch (baseline)	73.83	90.16
Contrastive-based Pre-training		
SimCLR [9]	74.24	90.62
MoCo v3 [11]	74.78	90.90
ContIG [34]	75.14	91.25
MIM-based Pre-Training		
MAE [16]	73.75	90.44
SimMIM [43]	73.96	90.80
MultiMAE [3]	74.45	91.03
mc-BEiT [25]	74.82	91.35
AttMask [19]	75.36	91.75
MRM	76.17	92.07

task. We use mean squared error (MSE) as the metric for numerical factors, *i.e.*, age, BMI, SBP, DBP, and the ROC-AUC for the categorical factors, *i.e.*, sex and smoking status.

4.1.2 Pathology Images with Genetics

Pre-training. We use TCGA-GBM with TCGA-LGG dataset [37] 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.

Downstream Transfer. We leverage glioma grading (GG) to evaluate 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.

4.2. Implementation Details

Following prior works [16, 44], 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 [26] and Adam [21], respectively, both with an initial learning rate of 1×10^{-3}. We use PyTorch [30] 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 [8, 11, 34, 16, 43, 3, 13, 25, 19] share the same settings to achieve a fair comparison. The balanced coefficient λ in Eq. (14) is set as 1.0, and the masking ratios τ_I, τ_G for images and genetics are equal to 75% and 50%, respectively.

4.3. Comparison with State-of-the-arts

In this section, to assess the quality of feature representations to various downstream tasks, we fine-tune the encoder with an initial task-relevant head together on the training set, and evaluate the downstream tasks on the test set.

Table 3. Ablation study of each proposed component in relation masking and relation matching on retinal image-based tasks.

Masking		Matching		APTOS	RFMiD	PALM	CRP	
R^M_{self}	R^M_{cross}	$\mathcal{L}_{\text{self}}$	$\mathcal{L}_{\text{cross}}$	QwKappa ↑	ROC-AUC ↑	Dice-Score ↑	MSE ↓	ROC-AUC ↑
				83.65	90.49	72.51	3.496	61.57
✓				86.37	93.48	75.29	3.352	67.21
	✓			85.10	92.71	74.20	3.418	64.82
✓	✓			87.50	94.28	77.23	3.283	70.52
✓	✓	✓		88.27	94.94	78.49	3.216	73.73
✓	✓		✓	88.65	95.51	79.77	3.162	74.27
✓	✓	✓	✓	89.83	96.31	81.45	3.107	75.90

Retinal Images with Genetics. We compare MRM pre-trained on UKB with state-of-the-art methods [8, 11, 34, 16, 43, 3, 13, 25, 19] on four downstream retinal image-based tasks in Table 1. MRM achieves 89.83%, 96.31%, 81.45%, 3.107 and 75.90% scores on APTOS, RFMiD, PALM and CRP tasks, respectively, outperforming other comparisons by a clear margin in all tasks including disease classification, segmentation and regression. Particularly, compared with MIM-based pre-training methods, MRM is superior over the second best one AttMask [19] with 2.10%, 1.40%, 2.86%, 0.155 and 6.50% scores on four tasks. These results reflect that masking out relation and retaining intact information within the input can obtain better feature representation than MIM-based input masking and per-sample reconstruction strategies for medical data.

Pathology Images with Genetics. To verify the versatility of the proposed framework, we pre-train MRM on pathology image-genome dataset TCGA and fine-tune the model on downstream grade grading task to evaluate the transfer results. From Table 2, compared with other pre-training methods, MRM yields state-of-the-art performance with 76.17% accuracy and 92.07% ROC-AUC scores, indicating its effectiveness and flexibility on various types of medical images. Hence, the proposed MRM is able to achieves significantly better self-supervised leaning.

4.4. Ablation Study

We ablate the effectiveness of each component in MRM on retinal image dataset.

Effectiveness of relation masking. We study the effectiveness of the proposed self- and cross-modality relation masking. We start from a baseline that trains the autoencoder to construct the multimodal data, we first mask out a portion of self-modality relation in both image and genome branches. From Table 3, the self-modality relation masking brings performance gains by 2.72%, 2.99%, 2.78%, 0.144 and 5.64% scores on APTOS, RFMiD, PALM and CRP tasks, respectively. Then, we further mask out the cross-modality relation, the results show consistent increases by 1.13%, 0.80%, 1.94%, 0.069 and 3.31% on four tasks. To

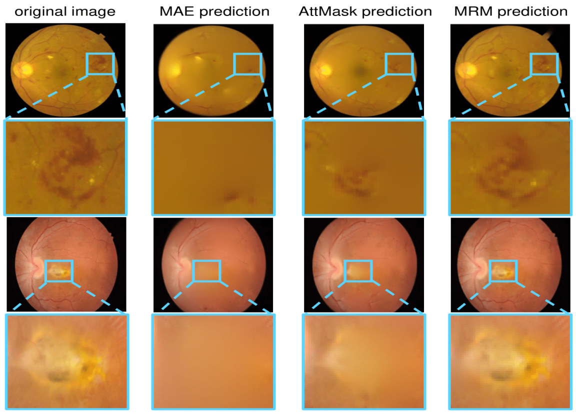

Figure 3. **Comparison of reconstruction results of different methods.** From left to right suggests the original input and the reconstructed images by MAE [16], AttMask [19] and our MRM. We can observe that MRM can preserve the disease regions framed in blue while MIM-based methods lose them.

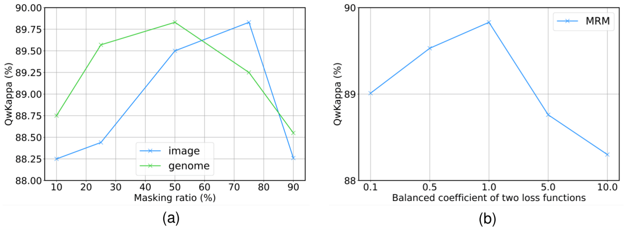

Figure 4. **Ablation study.** (a) Masking ratios τ_I and τ_G for image and genome. (b) Balanced coefficient λ of two loss functions.

Table 4. Results of gene-image association analysis.

Method	Found Regions ↑
Training from scratch	4
SimCLR [9]	5
MoCo v3 [11]	6
ContIG [34]	10
MAE [16]	8
SimMIM [43]	10
MultiMAE [3]	13
mc-BEiT [25]	14
AttMask [19]	16
MRM	18

qualitatively verify the impact of the relation masking, we visualize the reconstruction images of different methods. Figure 3 illustrates that MIM-based approaches [16, 19] 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.

Impact of relation matching. To observe the impact of the self- and cross-modality relation matching, on the top of the model with relation masking, we employ the self-modality matching constraint. As implied in Table 3, self-modality matching leads to the transfer result improvements by 0.77%, 0.66%, 1.26%, 0.067 and 3.21% scores on APTOS, RFMiD, PALM and CRP tasks, respectively. When adding the cross-modality relation matching, the performance continuously grows by 1.56%, 1.37%, 2.96%, 0.109 and 2.17% scores on four downstream tasks. The results illustrate that both self- and cross-modality relation matching can boost the representation learning, and they can render complementary constraints to achieve sample relation invariant when masking the feature relation.

Masking ratios. To understand the influence of masking ratios τ_I and τ_G for images and genetics, we fine-tune the pre-trained model on diabetic retinopathy detection task under different masking ratios. Firstly, we mask the self- and cross-modality relation for image features with different ratio τ_I including 10%, 25%, 50%, 75% and 90%. From Figure 4 (a), the best choice for τ_I is 75%, and the performance decreases when diminishing or enlarging the ratio. As the ratio reduces from 75% to 10%, the transfer result drops remarkably by 1.35%. Moreover, we analyze the effect of masking ratio τ_G for genetics and find that 50% is good for representation learning and transfer results.

Loss balanced coefficient λ. We study the sensitivity of MRM towards the balanced coefficients λ for controlling two loss functions $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{match}}$ in Eq. (14), as shown in Figure 4 (b). It can be observed that the performance rises with the balanced coefficients λ increasing from 0.1 to 1.0, while it plunges when λ is larger than 1.0. These suggest that the proposed relation matching leads to beneficial representation learning with the appropriate range of λ, either too strong or too weak of the relation matching constraint damage the quality of the feature representation.

4.5. Gene-image Association Analysis

To obtain an interpretable understanding on how genetics can improve images representation learning, we conduct gene-image association analysis for all models. This association analysis is a statistical tool for finding individual genetic regions correlated with image features according to the disease traits. The better the image representation, the more associated regions are expected to be discovered. Table 4 shows the number of independent regions each model finds, where we observe that MRM finds the most disease-related regions from genetics, explaining why genetics can improve image representation learning in our model.

4.6. Extension to Single Modality Pre-training

To study the application of the proposed MRM on single modality, we pre-train MRM using merely image modality on retinal dataset from UKB, and evaluate the feature representation on downstream tasks. Specifically, during the pre-training phase, we remove the cross-modality attention layers and adopt self-attention with self-modality relation

Table 5. Transfer results with single image modality pre-training on retinal image-based tasks.

Method	APTOS	RFMiD	PALM	CRP	
	QwKappa ↑	ROC-AUC ↑	Dice-Score ↑	MSE ↓	ROC-AUC ↑
Training from scratch	79.26	89.27	77.58	3.446	53.56
SimCLR [8]	81.43	90.51	72.52	3.453	59.25
MoCo v3 [17]	82.20	91.82	73.97	3.437	61.73
ContIG [34]	82.89	91.55	79.17	3.306	67.40
MAE [16]	78.31	88.62	73.05	3.503	60.32
SimMIM [43]	79.91	89.47	71.39	3.468	61.09
MultiMAE [3]	80.55	90.91	74.52	3.427	62.14
mc-BEiT [25]	82.59	91.78	75.84	3.388	63.61
AttMask [19]	83.80	92.36	76.50	3.285	64.77
MRM	85.46	93.11	80.26	3.245	70.15

masking to capture the representation. The image feature is directly fed into the image decoder to reconstruct the image without multimodal fusion. As for the relation matching, only self-modality matching is employed for image features. Table 5 indicates that MRM exhibits the best transfer ability across four downstream tasks, verifying the effectiveness of MRM with single image modality pre-training.

5. 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.

6. Conclusion

In this work, we present MRM framework to jointly leverage medical images and genetics for self-supervised pre-training. Instead of explicitly masking inputs, we design the relation masking to mask out feature relation and enable the model to capture informative patterns, which can retain intact disease-related semantics. Moreover, to enrich semantic relation, we present relation matching by exploiting inter-sample relation to encourage global constraints in the feature space. Extensive experiments verify the effectiveness of MRM on various downstream diagnosis tasks.

References

[1] Aptos 2019 blindness detection. https://www.kaggle.com/c/aptos2019-blindnessdetection/ Accessed: 2021-11-04. 6

[2] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proc. ICCV*, pages 3478–3488, 2021. 3

[3] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Proc. ECCV*, pages 348–367. Springer, 2022. 2, 3, 6, 7, 8, 9

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *Proc. ICLR*, 2022. 1, 2, 3

[5] Rodrigo Bonazzola, Nishant Ravikumar, Rahman Attar, Enzo Ferrante, Tanveer Syeda-Mahmood, and Alejandro F Frangi. Image-derived phenotype extraction for genetic discovery via unsupervised deep learning in cmr images. In *Proc. MICCAI*, pages 699–708. Springer, 2021. 1, 2, 3

[6] Nathaniel Braman, Jacob WH Gordon, Emery T Goossens, Caleb Willis, Martin C Stumpe, and Jagadish Venkataraman. Deep orthogonal fusion: multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. In *Proc. MICCAI*, pages 667–677. Springer, 2021. 1, 3

[7] Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE TMI*, 41(4):757–770, 2020. 2, 3

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2, 3, 6, 7, 9

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020. 2, 7, 8

[10] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022. 2, 3

[11] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 2, 7, 8

[12] Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Mariann Micsinai Balan, and Kevin Brown. Masked image modeling advances 3d medical image analysis. In *Proc. WACV*, 2023. 1

[13] 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. 1, 3, 6, 7

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

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. 2, 3

[16] 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. 1, 2, 3, 6, 7, 8, 9

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2, 3, 6, 9

[18] 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. 6

[19] 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. 2, 3, 6, 7, 8, 9

[20] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*, pages 4171–4186, 2019. 3

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proc. ICLR*, 2015. 7

[22] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Proc. NeurIPS*, volume 30, 2017. 4

[23] Wuyang Li, Jie Liu, Bo Han, and Yixuan Yuan. Adjustment and alignment for unbiased open set domain adaptation. In *Proc. CVPR*, pages 24110–24119, 2023. 2

[24] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proc. CVPR*, pages 5291–5300, 2022. 2

[25] 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. 1, 2, 6, 7, 8, 9

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *Proc. ICLR*, 2019. 7

[27] Yang Luo, Zhineng Chen, and Xieping Gao. Self-distillation augmented masked autoencoders for histopathological image classification. *arXiv preprint arXiv:2203.16983*, 2022. 1, 3

[28] Zhenyuan Ning, Denghui Du, Chao Tu, Qianjin Feng, and Yu Zhang. Relation-aware shared representation learning for cancer prognosis analysis with auxiliary clinical variables and incomplete multi-modality data. *IEEE TMI*, 41(1):186–198, 2021. 1, 2, 3

[29] 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. 6

[30] 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. 7

[31] Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.*, 2(3):158–164, 2018. 6

[32] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. In *Proc. ICLR*, 2020. 5

[33] 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. 6

[34] 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, 6, 7, 8, 9

[35] Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal self-supervised learning for medical image analysis. In *Proc. IPMI*, pages 661–673. Springer, 2021. 1, 3

[36] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *Proc. NeurIPS*, 33:18158–18172, 2020. 1, 3

[37] 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. 7

[38] Rami S Vanguri, Jia Luo, Andrew T Aukerman, Jacklynn V Egger, Christopher J Fong, Natally Horvat, Andrew Pagano, Jose de Arimateia Batista Araujo-Filho, Luke Geneslaw, Hira Rizvi, et al. Multimodal integration of radiology, pathology and genomics for prediction of response to pd-(l) 1 blockade in patients with non-small cell lung cancer. *Nature cancer*, 3(10):1151–1164, 2022. 1, 3

[39] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proc. ICCV*, pages 7303–7313, 2021. 2

[40] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proc. CVPR*, pages 14668–14678, 2022. 2

[41] LI Wuyang, YANG Chen, LIU Jie, LIU Xinyu, GUO Xiaoqing, and YUAN Yixuan. Joint polyp detection and segmentation with heterogeneous endoscopic data. In *Proc. ISBI Workshop*, pages 69–79. CEUR-WS Team, 2021. 1

[42] Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In *Proc. WACV*, pages 3588–3600, 2023. 1, 3

[43] 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. 1, 2, 3, 6, 7, 8, 9

[44] 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, 7

[45] Qiushi Yang, Xiaoqing Guo, Zhen Chen, Peter YM Woo, and Yixuan Yuan. D2-net: Dual disentanglement network for brain tumor segmentation with missing modalities. *IEEE TMI*, 41(10):2953–2964, 2022. 1

[46] Qiushi Yang and Yixuan Yuan. Learning dynamic convolutions for multi-modal 3d mri brain tumor segmentation. In *Brainlesion: MICCAI Workshop*, pages 441–451. Springer, 2021. 1

[47] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proc. ICCV*, pages 3040–3049, 2021. 1

[48] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *Proc. ICLR*, 2022. 2, 3

[49] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Self pre-training with masked autoencoders for medical image analysis. *arXiv preprint arXiv:2203.05573*, 2022. 1, 3