ContIG: Self-supervised Multimodal Contrastive Learning for Medical Imaging with Genetics

Aiham Taleb[1*] Matthias Kirchler[1,2*] Remo Monti[1] Christoph Lippert[1,3]

[1] Hasso Plattner Institute for Digital Engineering, University of Potsdam, Germany

[2] TU Kaiserslautern, Germany

[3] Hasso Plattner Institute for Digital Health at the Icahn School of Medicine at Mount Sinai, NYC, USA

{firstname.lastname}@hpi.de

Abstract

High annotation costs are a substantial bottleneck in applying modern deep learning architectures to clinically relevant medical use cases, substantiating the need for novel algorithms to learn from unlabeled data. In this work, we propose ContIG, a self-supervised method that can learn from large datasets of unlabeled medical images and genetic data. Our approach aligns images and several genetic modalities in the feature space using a contrastive loss. We design our method to integrate multiple modalities of each individual person in the same model end-to-end, even when the available modalities vary across individuals. Our procedure outperforms state-of-the-art self-supervised methods on all evaluated downstream benchmark tasks. We also adapt gradient-based explainability algorithms to better understand the learned cross-modal associations between the images and genetic modalities. Finally, we perform genome-wide association studies on the features learned by our models, uncovering interesting relationships between images and genetic data.[1]

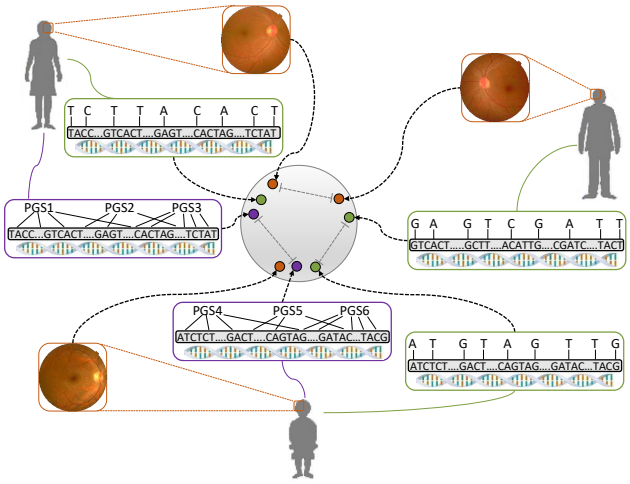

Figure 1. Overview of our contrastive learning method from imaging and genomic data. It learns representations by bringing the modalities of each individual closer in the embedding space, and apart from different individuals'. In this example, the modalities are retinal fundus images (in brown), SNP data (in green), and polygenic risk scores (PGS) (in purple). Our method handles missing modalities (*e.g.* missing PGS for the person in the upper right).

1. Introduction

Medical imaging plays a vital role in patient healthcare. It aids in disease prevention, early detection, diagnosis, and treatment. However, efforts to employ machine learning algorithms to support in clinical settings are often hampered by the high costs of required expert annotations [41]. At the same time, large-scale biobank studies have recently started to aggregate unprecedented scales of multimodal data on human health. For example, the UK Biobank (UKB) [96] contains data on $500,000$ individuals, including a wide range of imaging modalities such as retinal fundus images and cardiac, abdominal, and brain MRI. Similar studies are currently underway in other countries, such as the Nationale Kohorte (NaKo) [15], BioMe [13], FinnGen [34], Estonia Biobank [14], and others. While some of these studies also include phenotypic descriptions, *e.g.* a person's medical history, such data tend to be both highly incomplete and biased due to clinical practices and assessment methods [76], making learning from them challenging and error-prone. On the other hand, genetic data is increasingly abundant. While chip-based genotyping technology has enabled the study of common genetic variation at scale [108], the exponentially decreasing costs of genomic sequencing is driving progress for rare genetic variation [80]. Due to these advances, the UKB and other biobanks often contain a rich array of ge-

*Equal contribution

[1]Source code at: https://github.com/HealthML/ContIG

netic and genomic measurements. Genetic data is generally less susceptible to bias factors, and most diseases have at least a partially genetic cause, with some genetic disorders being exclusively attributed to genetic mutations [110]. Similarly, most other traits – not directly related to diseases –, *e.g*. height and human personality, are also strongly influenced by genetics [64, 130]. Complementary imaging-genetics datasets are increasingly also available in other application settings, *e.g*. plant breeding [116].

Unlabelled medical images carry valuable information about organ structures, and an organism's genome is the blueprint for biological functions in the individual's body. Clearly, integrating these distinct yet complementary data modalities can help create a more holistic picture of physical and disease traits. Integrating these data types, however, is non-trivial and challenging. The human genome consists of three billion base pairs, yet most genetic differences between individuals have little effect. This leads to challenges both in terms of computational aspects, and in terms of statistical efficiency. Unfortunately, it is not clear a priori which parts of the genome are relevant and which are not. Typically, genome-wide association studies (GWAS) [50, 69] use statistical inference techniques to discover relationships between genetic variations and particular physical or disease traits. To date, thousands of scientific works have found more than $300,000$ genetic-phenotype associations [51]. However, even now a large portion of known or presumed heritability of traits is not yet accounted for by the individual genome-trait associations, a phenomenon known as "missing heritability" [70]. Better methods to find – and explain – the relationships between genetic and imaging modalities may help close this gap.

Therefore, the growing number of biobanks of *unlabeled* multimodal (*i.e*. imaging-genetics) data, calls for solutions that can: (i) learn semantic data representations without requiring expensive expert annotations, (ii) integrate these data modalities end-to-end in an efficient manner, and (iii) explain discovered cross-modal correspondences (associations). Self-supervised (unsupervised) representation learning offers a viable solution when unlabeled data is abundant and labels are scarce. These methods witnessed a surge of interest after proving successful in several application domains [54]. The representations learned by these methods facilitate data-efficient fine-tuning on supervised downstream tasks, reducing significantly the burden of manual annotation. Furthermore, such methods allow for integrating multiple data modalities as distinct views, which can lead to considerable performance gains. Despite the recent advancements in self-supervised methods, *e.g*. contrastive learning, only little work has been done to adopt these methods in the medical domain. In fact, we are not aware of any prior work that leverages self-supervised representation learning on combined imaging and genetic modalities. We believe self-supervised learning has the potential to address the above challenges inherent to the medical domain.

Contributions. **(i)** We propose a self-supervised method, called ContIG, that can learn from multimodal datasets of *unlabeled* medical images and genetic data. ContIG aligns these modalities in the representation space using a contrastive loss, which enables learning semantic representations in the same model end-to-end. Our approach handles the case of multiple genetic modalities, in conjunction with images, even when the available modalities vary across individuals. **(ii)** We adapt gradient-based explainability algorithms to better understand the learned cross-modal correspondences (associations) between the images and genetic modalities. Our method discovers interesting associations, and we confirm their relevance by cross-referencing biomedical literature.

Our work presents a framework on how to exploit inexpensive self-supervised solutions on large corpora (*e.g*. Biobanks) of (medical) images and genetic data.

2. Related Work

Self-supervised learning with pretext tasks. These methods learn an embedding (representation) space by deriving a proxy (pretext) task from the data itself, requiring no human labels. The learned embeddings will also be useful for real-world downstream tasks, afterwards. A large body of works relied on such proxy tasks [17, 30, 39, 71, 75, 122]. A comprehensive review of similar works is provided in [54]. The limitation of such methods is the need to design handcrafted proxy tasks to learn representations.

Contrastive learning approaches [18, 23–25, 32, 40, 44, 46, 47, 72, 106, 112, 120] circumvent the above challenge by maximizing mutual information between related signals in contrast to others, by employing Noise Contrastive Estimation [43]. Contrastive methods advanced the results of unsupervised learning on ImageNet [29]. However, unlike our method, these methods process uni-modal images only.

Multimodal learning. Learning from multimodal data poses several inherent challenges, such as: multimodal fusion, alignment, and representation [12, 73]. Prior works, some of which are self-supervised, learn from a variety of modalities, such as: image and text (vision and language) [8, 55, 62, 67, 97, 98, 104, 113], image and audio [3, 5, 6, 9, 77, 78], audio and text [1, 118], and multi-view (multimodal) images [84, 90, 92, 105]. More recent works employed contrastive learning for multimodal inputs (image and text captions) [2, 81, 85, 119, 123, 124]. We follow this line of work, and we extend contrastive pretraining to novel modalities, *i.e*. images and genetics, for the first time.

Self-supervision on medical images. Early works of self-supervision in the medical context [11, 52, 61, 66, 88, 94, 114, 117, 121] made assumptions about input data, limiting their generalization to other target tasks. Then, many

works proposed employing proxy tasks for self-supervision from medical scans [16, 22, 53, 53, 95, 100, 102, 126, 127]. A review of similar works is in [101]. Recently, contrastive learning [19, 48, 65, 103] has been applied to medical scans, where it also showed promising results. Our work, as opposed to these works, utilizes multiple modalities (images and one or more genetic modalities) to improve the learned representations by capturing imaging-genetic relationships.

Deep learning from both genetics and images. In addition to its successful applications to medical imaging [68], deep learning also found success in applications on genomics [33, 58, 111, 128, 129]. There is a growing number of recent works that utilize deep neural networks to jointly learn from both modalities, such as [7, 10, 21, 28, 37, 42, 57, 74, 107, 125]. However, these methods are all either highly application specific or fully supervised. Notably, we are not aware of any prior work leveraging the self-supervised framework (with contrastive loss functions) to improve representation learning from both imaging and genetic data.

3. Method

In Sec. 3.1 we first review some biomedical foundations and motivate the genetic modalities chosen in this work. Then, we describe our contrastive method in Sec. 3.2, and different modality aggregation types. Finally, we detail the explanation methods for genetic features in Sec. 3.3.

3.1. Modalities of Genetic Data

The basic building blocks of DNA, which encodes the biological functions needed for the development of an organism, are called nucleotides. A long sequence of the four nucleotides Adenine (A), Thymine (T), cytosine (C), and Guanine (G) make up the genome - the "recipe" needed to build an organism [49]. A relatively small fraction of the genome codes for proteins, while the remaining parts have regulatory or structural functions. However, over generations, genetic mutations occur, for example substituting one nucleotide for another, *e.g.* A to C. Some of these genetic changes can alter physical traits (*e.g.* eye color), or cause disease (*e.g.* Alzheimer's). "Genotyping" is the process of measuring these genetic changes [86]. The most frequently measured type of changes are single-nucleotide-polymorphisms (SNPs), where a single pair of nucleotides is altered at a specific position in the genome.

There are three billion base pairs in the human genome, but typically only a small fraction of them is measured, due to cost and technological restraints. Even if large parts of the sequence are available, as is the case for whole-genome sequencing studies, working with the raw data is not feasible, both in terms of *statistical efficiency* – most of those base pairs carry no causal signal and only add noise to the estimation process – and in terms of *computational efficiency*. For these reasons, most studies record only a small subset of all nucleotides, usually on the order of several hundred thousand to several million SNPs. Furthermore, human traits of interest are constructed by a spectrum of different genetic architectures. At the same time, due to evolutionary dynamics, some SNPs exhibit their possible variations frequently in a population ("common" variants), while other SNPs are identical for the overwhelming majority of the population with only few individuals having mutations ("rare" variants) – a form of class imbalance. Therefore, in this work we consider *three* genetic modalities to that emphasize different aspects of human physiology.

Complex traits are traits that are influenced by a large number of causal factors, including relatively common genetic variations. One example is height, which is determined to a large degree by many SNPs all across the human genome [115]. Many common diseases and impairments are complex traits, which makes them especially relevant to human health applications [35]. To best encode genetic architectures associated with complex traits, we utilize **polygenic risk scores (PGS)** [31]. PGS aggregate many, mostly common, SNPs into a single score that reflects a person's inherited susceptibility to a specific disease [56]. The individual SNPs are weighted based on their strength of association with the disease. By using many different PGS for different traits and diseases we can get a multi-faceted view of an individual's complex trait predisposition.

Recent advances in DNA sequencing have also enabled assessing the contribution of *rare genetic variants* to heritable traits [109]. Rare variants occur at low frequencies (*e.g.* MAF[2] $< 1\%$ or MAF $\ll 1\%$) in a population. Large genetic effects often negatively affect an individual's health and are strongly selected against by evolution. Hence, in contrast to common variants, many rare variants have a large effect size and predispose for genetic diseases. Rare variants are usually not included in PGS, and due to their low frequencies they pose a challenge for robust statistical models. In this work, we use **burden scores** [60], which aggregate several rare variants within a localized genetic region.

Finally, we also employ a uniformly sampled cross section of the whole genome, by including every k-th SNP that has been genotyped in the respective study. These **raw SNPs** are mostly common variants (due to the biological sampling procedure) and give a broad representation of an individual's genetic composition. This representation likely carries population structure such as ancestry [63], but also tags highly diverse functional information.

These three genetic modalities – polygenic risk scores, burden scores, and raw SNPs – capture complementary aspects[3] and together paint a broad description of an individual's genetic predisposition. We employ them both individ-

[2] minor allele frequency

[3] More than 99% of all pairwise correlations between features from different modalities have absolute value below 0.02.

ually and jointly as contrastive views to medical images.

3.2. Contrastive Learning from Images & Genetics

We assume a dataset of N multimodal samples, one for each individual person. Each sample consists of a medical image paired with multiple genetic modalities. Here, we denote each image by x_v^i, and the corresponding genetic modalities by x_{gm}^i, where $i \in \{1, .., N\}$ is the individual and $m \in \{1, .., M\}$ is the genetic modality. We group images and genetic modalities in batches of size $b > 1$ by the individual modalities: $v = \{x_v^{i_1}, \ldots, x_v^{i_b}\}$ and $g_m := \{x_{gm}^{i_1}, \ldots, x_{gm}^{i_b}\}$. The number of available genetic modalities may vary across individuals.

Our method, illustrated in Fig. 2, processes these input modalities with a set of neural network encoders, one per modality. We denote the image encoding by $h_v^i = f_v(x_v^i)$, and the genetics encodings as $h_{gm}^i = f_{gm}(x_{gm}^i)$, with M distinct genetics encoders. The resulting d-dimensional vector representations $h_v^i, h_{gm}^i \in \mathbb{R}^d$ are then processed with projection heads $z_v^i = p_v(h_v^i), z_{gm}^i = p_{gm}(h_{gm}^i)$, respectively, where $z_v, z_{gm} \in \mathbb{R}^d$. Following [23], each projection head is a non-linear MLP with one hidden-layer.

Contrastive Loss with Two Modalities. We first define the contrastive loss assuming N pairs of an image and one genetic modality (x_v^i, x_g^i), with their respective representations (z_v^i, z_g^i). Then, for the image sample in the i^{th} pair, we consider the genetic sample x_g^i as the positive (true) sample among the negative genetic samples of other individuals x_g^k in the same batch. Similarly, the image x_v^i is the positive sample of x_g^i, amongst the negative image samples x_v^k. Therefore, the contrastive loss is the sum of these two parts: i) image-to-genetics $L(v, g)$ (fix the image and contrast genetic samples), and ii) genetics-to-image $L(g, v)$ (fix the genetics and contrast images). Formally, in each step of the training we select a random batch of size $b > 1$ with indices $\{i_1, \ldots, i_b\}$ and use the batch-wise loss function:

$$L(v, g) = -\sum_{j=1}^{b} \log \frac{\exp(\cos(z_v^{i_j}, z_g^{i_j})/\tau)}{\sum_{k=1, k\neq j}^{b} \exp(\cos(z_v^{i_j}, z_g^{i_k})/\tau)}$$
$$\mathcal{L}_{cont}(v, g) = \lambda L(v, g) + (1 - \lambda) L(g, v), \tag{1}$$

where τ is a temperature parameter, cos is the cosine similarity, and $\lambda \in [0, 1]$ is a loss weighting hyperparameter.

Generalizing to Multiple Genetic Modalities. We generalize here the above contrastive loss formulation to the case when there are multiple available genetic modalities, corresponding to the same image sample. Since we aim to improve the learned visual representations mainly, the image modality is used at the center of this training scheme (we deem alternative contrasting schemes a future work). In other words, we contrast the image with each one of the M genetic modalities. Therefore, the generalized multimodal contrastive loss becomes:

$$\mathcal{L}(v, g_1, \ldots, g_M) = \sum_{m=1}^{M} \mathcal{L}_{cont}(v, g_m) \tag{2}$$

This formulation ensures the learned visual representations capture useful information from all available genetic modalities. However, this assumes that every individual has all the genetic modalities, which is not normally the case. Hence, we define two aggregation schemes to handle the missing genetic modalities: i) the "inner" aggregation scheme, which uses only those individuals for which *all* the modalities exist, and ii) the "outer" aggregation scheme, which covers all the individuals, even those with *missing* genetic modalities. In particular, for each $\mathcal{L}_{cont}(v, g_m)$ in Eq. (2), the "outer" aggregation only includes individuals with non-missing data for this specific modality. The "outer" scheme can make better use of all available data. Both schemes allow for training on combinations of existing modalities.

3.3. Genetic Features Explanation

For a given multimodal tuple $x := (x_v, x_{g1}, \ldots, x_{gM})$ of image and genetic representations, we perform feature explanations to understand the contribution of each genetic feature $g_{m,j}$ for the model output. Standard deep learning explainability approaches are not directly applicable in this setting, as they require a simple one-to-one relation from input to output, while the contrastive loss Eq. (2) is computed over batches. Instead, we utilize a fixed reference batch of $b \geq 1$ individuals with images v_r and genetic modalities $g_{m,r}$ $(m = 1, \ldots, M)$ and define the explainer function

$$E(x) := \mathcal{L}(v_r \cup \{x_v\}, g_{1,r} \cup \{x_{g1}\}, \ldots, g_{M,r} \cup \{x_{gM}\})$$

with $\mathcal{L}$ defined as in Eq. (2), but $v_r, g_{1,r}, \ldots, g_{M,r}$ fixed. We can then use standard feature attribution methods such as Integrated Gradients [99] or DeepLift [91] to explain the contribution of all elements in x towards the full batch loss. We can additionally also fix the input image x_v to only consider the attribution of the genetic effects. Note that the explanation will be sensitive to the choice of the reference batch; to minimize this effect, we choose b to be relatively large ($b = 1,000$ in our experiments).

In addition to these *local* instance-specific attributions, we are especially interested in understanding the behavior of our models *globally*. For this, we aggregate many individual explanations, all using the same (independent) reference batch. Feature importance both in negative and positive direction is important in our setting, and therefore we consider the mean absolute value for each feature dimension as a measure of global attribution.

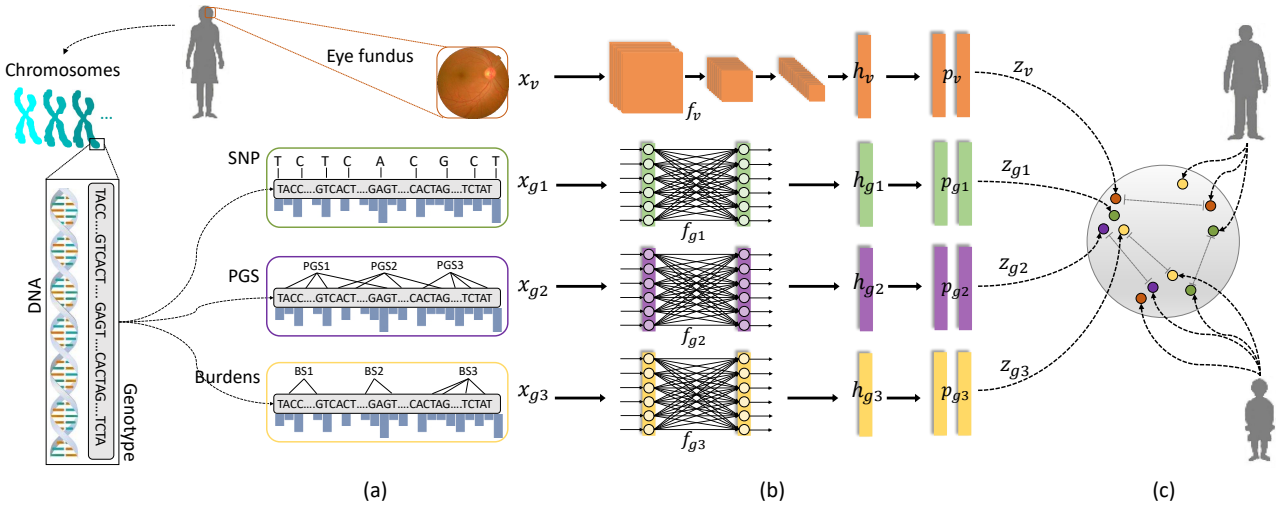

Figure 2. Schematic illustration for the steps of our proposed method. **(a)** Assuming one imaging modality (retinal fundus shown in brown), and three genetic modalities (Single-nucleotide polymorphisms (SNP) in green, polygenic risk scores (PGS) in purple, burden scores in yellow). Note that different genetic modalities exhibit different variant frequencies (denoted by the histogram in blue): SNP and PGS use common variants (high frequency), while burdens use rare variants (low frequency). **(b)** We extract features from each modality with deep neural networks, *i.e*. Convolutional Networks for images and Fully Connected (MLP) networks from genomic data. We use a projection head (MLP) for each modality, which produces equally-sized modality embeddings $z_v, z_{g1}, z_{g2}, z_{g3}$. **(c)** We use these embeddings in the contrastive loss computation. The embeddings of each individual are encouraged to come closer in the feature space (depicted by the gray circle), and farther from other individuals'. The dotted gray lines demonstrate the contrasting mechanism between modalities.

The setting with missing values can be handled analogously to the "outer" aggregation scheme in Sec. 3.2, by just omitting the respective modalities.

4. Experimental Results

In this section, we present the evaluation results of our method. First, we detail the datasets used for pretraining and evaluation (Sec. 4.1). Then, we assess the quality of the learned representations (Sec. 4.2), by: i) fine-tuning (*i.e*. transfer learning) on four downstream tasks, and ii) performing a genome-wide association study (GWAS) on the model features. Finally, we present the genetic feature explanation results (Sec. 4.3), and we analyze the findings to check their relevance with medical literature resources. We provide additional evaluation results in the appendix.

4.1. Datasets

We pretrain our models (and the unsupervised baselines) on data obtained from the UK Biobank (UKB) dataset [96]. This dataset contains multimodal data for almost 500k individuals, although imaging data is only available for a subset of those. The UKB consists of an overwhelming majority of individuals of European descent; we restrict our dataset to those European descent individuals, as including more individuals would likely introduce very large confounding effects [63]. For the purposes of pretraining, we utilize the retinal fundus images, which amount to $155,238$ imaging samples (left and right eyes). The genetic modalities we employ, amount to $155,238$ Raw-SNP samples (all individuals have Raw-SNPs), $145,206$ PGS samples, and $93,216$ burden scores. We holdout a test split (20%) from the UKB dataset, and the remaining data are for training (70%) and validation (10%). Each person only appears in one split.

For the downstream tasks, we employ: i) APTOS 2019 Blindness Detection [4] dataset for Diabetic Retinopathy detection in Sec. 4.2.1, which has $3,662$ retinal fundus training samples. ii) Retinal Fundus Multi-disease Image Dataset (RFMiD) [79] for disease classification (Sec. 4.2.2), which has $3,200$ training images. iii) $102,219$ images from the UKB [96] training split, but now we predict cardiovascular risk factors (Sec. 4.2.4). iv) Pathologic Myopia challenge dataset [36] for Pathological Myopia Segmentation (Sec. 4.2.3), which has 400 image samples with segmentation masks. More datasets details in the appendix.

4.2. Transfer Learning Results

In this section, we evaluate the quality of representations by fine-tuning to downstream tasks. However, we find that a linear evaluation protocol [23, 106] (encoder weights are kept frozen) behaves similarly to fine-tuning, see appendix.

Models & architectures. Across the following experiments, we employ a Resnet50 [45] architecture as the encoder for image data (f_v in Fig. 2). For the genetic en-

coders (f_{gm}), we vary the number of fully connected layers: "None" hidden layers, one hidden layer "H1", and two hidden layers "H12". In Sec. 3.1 we detail the genetic modalities used in pretraining. Aggregation schemes in Sec. 3.2.

Baselines. We compare to the following baselines:

- Training from scratch (**Baseline**): we train the same model on each downstream task, but initialized from random weights. Comparing to this baseline provides insights about the benefits of pretraining.
- State-of-the-art contrastive methods: we compare to self-supervised (contrastive) methods from literature by training on the same data splits, and using the same experimental setup. Namely, we compare to models pretrained with **SimCLR** [23], **BYOL** [40], **Barlow Twins** [120], **SimSiam** [25], and **NNCLR** [32].

4.2.1 Diabetic Retinopathy Detection (APTOS)

Millions of people suffer from Diabetic Retinopathy, the leading cause of blindness among working aged adults. The APTOS dataset [4] contains 2D fundus images, which were rated by a clinician on a severity scale of 0 to 4. These levels define a five-way classification task. We fine-tune the image encoder of our models and the baselines on this dataset, and then we evaluate on a fixed test split (20% of the data). The metric used in the task, as in the official Kaggle challenge, is the Quadratic Weighted Kappa (QwKappa [27]), which measures the agreement between two ratings. Its values vary from random (0) to complete agreement (1), and if there is less agreement than chance it may become negative. The evaluation results in Tab. 1 support the effectiveness of our proposed contrastive method (ContIG). Our pretrained models outperform all baselines in this task, demonstrating the quality of its learned representations.

4.2.2 Retinal Fundus Disease Classification (RFMiD)

The Retinal Fundus Multi-disease Image Dataset (RFMiD) [79] also contains 2D fundus images, which are captured using three different cameras. It has 46 class labels, which represent disease conditions annotated through adjudicated consensus of two experts. Similarly, to evaluate on this task, we fine-tune the image encoders on this dataset, and we measure the performance on the test set. We should note that this task is solved as a multi-label classification task, since the patients may have multiple conditions at the same time. As an evaluation metric, we compute area under the ROC curve (ROC-AUC), and we use a micro averaging scheme [59]. The results for this task in Tab. 1 also demonstrate the gains in performance obtained by training with ContIG. Our models also outperform the self-supervised baselines in this task.

4.2.3 Pathological Myopia Segmentation (PALM)

Myopia has become a global burden of public health. Pathologic myopia causes irreversible visual impairment to patients, which can be detected by the changes it causes in the optic disc, including peripapillary atrophy, tilting, etc. The PALM dataset [36] contains segmentation masks for these lesions, from which we evaluate on disc and atrophy segmentation tasks. Similar to the above downstream tasks, we fine-tune the image encoder on this dataset and evaluate on the test split. To predict segmentation masks, we add a u-net decoder [87] on top of the ResNet50 encoder. In terms of evaluation metrics, we use the dice score [93]. The results of this task in Tab. 1 demonstrate the quality of the learned representations by ContIG on semantic segmentation.

4.2.4 Cardiovascular Risk Prediction

Previous work has shown that retinal fundus images can predict a range of risk factors for cardiovascular diseases [82]. Namely, retinal fundus images have been found to carry information about age, sex, smoking status, systolic and diastolic blood pressure (SBP, DBP), and body mass index (BMI). We predict these six risk factors using a subset of the UK Biobank [96] dataset, by fine-tuning the image encoder on these values. As evaluation metrics, we use Mean Squared Error (MSE) for the numerical factors (age, BMI, SBP, DBP), and we use the ROC-AUC value for the categorical factors (sex and smoking status). As Tab. 1 shows, models pretrained with ContIG outperform the baseline models in both classification and prediction tasks.

4.2.5 Genome-wide Association Study Results

A GWAS is a statistical analysis that correlates individual genetic markers sampled along the full genome with a trait of interest, such as a specific disease. GWASs usually require a low-dimensional, well-defined trait for association analysis; there is only little work yet on leveraging full medical imaging data in a GWAS setting [7, 57]. Here, we follow the *transferGWAS* [57] framework to evaluate the embeddings learned by ContIG. In this framework, images are projected onto their latent space embeddings and then the dimensionality is further reduced with a Principal Component Analysis. These low dimensional image representations can then be efficiently associated with SNPs using statistical association analysis tools such as PLINK [20, 83]. To compare different training methods, we count how many independent genetic regions each method finds; a more expressive image representation is expected to find more associated regions. We defer the analysis details to the appendix.

Tab. 2 shows the number of found independent regions for each pretraining method. Genetic pretraining increases

Model & Genetics Encoder		APTOS	RFMiD	PALM	Cardio. Risk Pred.	
		QwKappa ↑	ROC-AUC ↑	Dice-Score ↑	MSE ↓	ROC-AUC ↑
Baseline	-	80.47	91.64	77.25	3.440	56.29
SimCLR [23]	-	81.83	91.88	70.41	3.451	59.38
SimSiam [25]	-	75.44	91.28	72.26	3.442	57.37
BYOL [40]	-	71.09	89.88	66.32	3.414	59.73
Barlow Twins [120]	-	72.28	92.03	70.53	3.430	59.05
NNCLR [32]	-	77.93	91.89	72.06	3.426	61.95
ContIG (Raw-SNP)	H1	84.01	93.22	76.98	3.254	70.10
ContIG (PGS)	H1	**85.93**	<u>93.31</u>	**78.47**	<u>3.176</u>	**72.72**
ContIG (Burden)	H1	83.22	93.03	76.49	**3.160**	<u>72.37</u>
ContIG (Inner RPB)	H1	81.52	92.95	<u>77.34</u>	3.202	70.80
ContIG (Outer RPB)	H1	<u>84.22</u>	**93.62**	76.97	3.187	71.80

Table 1. Downstream evaluation results by fine-tuning on each task. **Bold** indicates the best result, <u>underlined</u> is second best. RPB in our method stand for the genetic modalities used: Raw-SNPs, PGS-scores, and Burden-scores. ↑ means higher is better, and ↓ lower is better.

the statistical power of the genetic association study considerably. While BYOL achieves near-competitive results, these results are likely false-positives due to ill-fitted linear models (see Appendix). All other self-supervised methods are outperformed by a large margin. We also looked up the found regions in the GWAS catalog [51] of published association results. Many of the regions were already known to be associated with skin pigmentation. This is not surprising, as the retina is known to be pigmented itself, which again is likely to be correlated with actual skin pigmentation. Besides pigmentation, the GWAS catalog records associations with an array of cardiovascular traits (such as BMI, pulse pressure, large artery stroke, and blood biomarkers), as well as eye-specific associations (cataract and astigmatism). Similar results were found by [57], albeit with a larger sample size.

Model	Found Regions ↑
SimCLR [23]	4
SimSiam [25]	2
BYOL [40]	(17)
Barlow Twins [120]	8
NNCLR [32]	3
ContIG (Raw-SNP)	16
ContIG (PGS)	<u>20</u>
ContIG (Burden)	19
ContIG (Inner RPB)	**22**
ContIG (Outer RPB)	18

Table 2. GWAS results. Indicated is the number of independent regions associated with the image embeddings for each model.

4.3. Genetic Feature Explanation Results

In this section, we investigate the representations learned by ContIG, using the explanation methods developed in Sec. 3.3. We first validate that our explanation approach can distinguish meaningful features from noise features, see appendix. Next, we analyze the models trained with a single genetic modality. Fig. 3 shows the 30 PGS with the strongest attributions, aggregated over 1,000 examples with a reference batch of size 1,000. The most important features are different kinds of skin cancers (basal & squamous cell carcinoma, cutaneous melanoma and melanoma). This can be explained by the fact that the retina is pigmented and skin pigmentation is highly correlated to skin cancer.

Besides that, glaucoma, which is a disease of the optic nerve, is a highly relevant PGS, and many of the other traits are linked to cardiovascular functions (abnormal EKG, HDL cholesterol, blood protein measurements, QT interval), smoking status (lung adenocarcinoma, FEV/FEC ratio, response to bronchodilator) and liver and kidney function (triglyceride & serum urea measurements). This is in line with previous studies which found strong signals with similar biomarkers in retinal fundus images [82]. Interestingly, ContIG also finds correlations with neurological conditions such as Parkinson's disease and autism, which have previously been linked to retinal changes as well [38, 89].

Similarly, among the 15 strongest associations for raw SNPs, these SNPs were previously associated with cardiovascular traits (rs10807207, rs228416, rs1886785, rs10415889, rs3851381), pigmentation (rs228416), neurological and psychological conditions (rs1886785, rs1738895, rs6533374), and smoking status (rs6533374).

In addition to the global attributions, Fig. 4 shows the local attributions for one image/PGS pairing. The retinal fundus image shows strong signs of vascular tortuosity, a known and important biomarker for cardiovascular conditions [26]. Analogously, for this instance there is a large number of PGSs very strongly related to cardiovascular

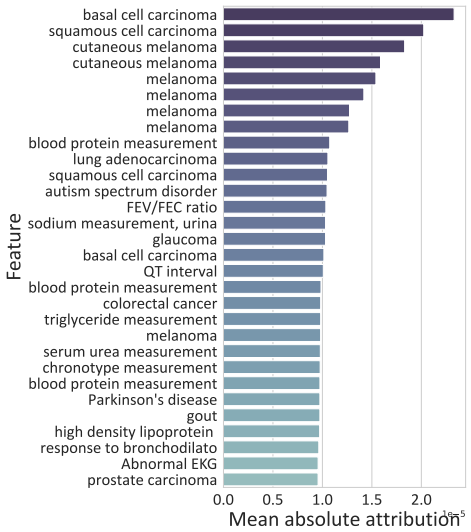

Figure 3. Global explanations for genetic features in ContIG (PGS only). Recorded is the mean absolute attribution per feature, aggregated over 1000 individuals, and the 30 PGS with highest associations are shown. Repeated traits (*e.g.* Melanoma) are due to multiple different risk scores published in the PGS catalog.

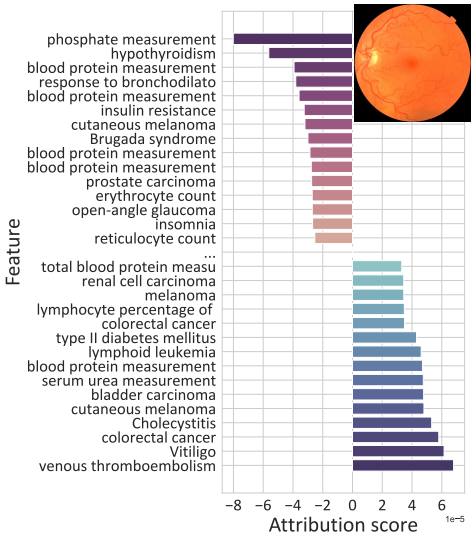

Figure 4. Local explanation attributions (signed) of genetic features for one image-PGS pair. Only the risk scores with highest values in either positive direction are shown. Retinal fundus image reproduced by kind permission of UK Biobank ©.

health (insulin resistance, many blood biomarkers, type II diabetes, Brugada syndrome, thromboembolism).

These local and global explanations together provide further evidence that self-supervised pretraining with ContIG is able to learn semantically meaningful image representations without the need for manual annotations. We provide additional explanatory results in the appendix.

5. Discussion & Limitations

We presented ContIG, a self-supervised representation learning algorithm for imaging-genetics datasets. Our evaluation results show that including genetic information in the pretraining process can considerably boost performance of image models in a variety of downstream tasks relevant for clinical practice and genetic research. We additionally conjecture that the self-supervised baseline methods' reliance on image augmentations alone may be disadvantageous in medical applications due to the more uniform nature (*e.g.* color distributions) of medical images compared to in natural images. We also leveraged interpretability methods to understand the relationship between imaging and genetic modalities in more detail and find interesting associations. Naturally, there are a number of limitations for our proposed approach. First, ContIG requires datasets that capture both imaging and genetics data, and is thus not applicable to pure-imaging datasets. In recent years, however, an increasing number of imaging-genetics studies have started, and proprietary datasets of joint imaging and genetics data are available in some large-scale health systems. With the decreasing prices in both imaging and genotyping technology, this trend is likely to continue further. A second limitation lies in the potentially limited application fields of our method. ContIG is not applicable to standard natural images, as there are no corresponding genetic features. On the other hand, large-scale biobanks often include multiple imaging modalities, such as different MRI and histopathology images. Our method is also applicable to imaging-genetics applications in live-stock and plant breeding, and may also be useful in basic science studies.

Unfortunately, most large-scale imaging-genetics datasets to date are conducted in European and Northern American countries. Therefore, one limitation of the presented results is that the UKB mostly consists of populations with European ancestry, and may carry a biased representation. We have shown that ContIG nevertheless improves downstream tasks in other populations, *e.g.* in APTOS (collected in India), RFMiD (collected in India), and PALM (collected in China). We deem extending ContIG to other medical imaging datasets and genetic populations a future work.

Acknowledgements

This project has received funding by the European Commission in the Horizon 2020 project INTERVENE (Grant agreement ID: 101016775), and the Federal Ministry of Education and Research - Germany (Bundesministerium für Bildung und Forschung - BMBF) [project: SyReal number: 01IS21069A]. This research has been conducted using the UK Biobank Resource.

References

[1] Tuka Al Hanai, Mohammad Ghassemi, and James Glass. Detecting depression with audio/text sequence modeling of interviews. In *Proc. Interspeech 2018*, pages 1716–1720, Graz, Austria, 2018. ISCOM. 2

[2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2(6):7, 2020. 2

[3] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019. 2

[4] APTOS. Aptos 2019 blindness detection. https://www.kaggle.com/c/aptos2019-blindness-detection/. Accessed: 2021-11-04. 5, 6

[5] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, Oct 2017. 2

[6] Yuki M Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *arXiv preprint arXiv:2006.13662*, 2020. 2

[7] Jordan T Ash, Gregory Darnell, Daniel Munro, and Barbara E Engelhardt. Joint analysis of expression levels and histological images identifies genes associated with tissue morphology. *Nature communications*, 12(1):1–12, 2021. 3, 6

[8] Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Cross-modal scene networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2303–2314, Oct. 2018. 2

[9] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NeurIPS*, NIPS'16, page 892–900, Red Hook, NY, USA, 2016. Curran Associates Inc. 2

[10] Liviu Badea and Emil Stănescu. Identifying transcriptomic correlates of histology using deep learning. *PloS one*, 15(11):e0242858, 2020. 3

[11] Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E. Petersen, Yike Guo, Paul M. Matthews, and Daniel Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *MICCAI*, pages 541–549. Springer International Publishing, 2019. 2

[12] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. 2

[13] BioMe Biobank. Biome biobank. https://icahn.mssm.edu/research/ipm/programs/biome-biobank. Accessed: 2021-11-09. 1

[14] Estonia Biobank. Estonia biobank. https://genomics.ut.ee/en. Accessed: 2021-11-09. 1

[15] Nako Biobank. Nako biobank. https://nako.de/. Accessed: 2021-11-09. 1

[16] Maximilian Blendowski, Hannes Nickisch, and Mattias P. Heinrich. How to learn from unlabeled volume data: Self-supervised 3d context feature learning. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *MICCAI*, pages 649–657. Springer International Publishing, 2019. 3

[17] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, Munich, Germany, September 2018. Springer. 2

[18] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 2

[19] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations, 2020. 3

[20] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015, 2015. 6

[21] P Chang, J Grinband, BD Weinberg, M Bardis, M Khy, G Cadena, M-Y Su, S Cha, CG Filippi, D Bota, et al. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *American Journal of Neuroradiology*, 39(7):1201–1207, 2018. 3

[22] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58:101539, 2019. 3

[23] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Int. Conf. on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 2, 4, 5, 6, 7

[24] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[25] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 2, 6, 7

[26] Carol Yim-lui Cheung, Yingfeng Zheng, Wynne Hsu, Mong Li Lee, Qiangfeng Peter Lau, Paul Mitchell, Jie Jin Wang, Ronald Klein, and Tien Yin Wong. Retinal vascular tortuosity, blood pressure, and cardiovascular risk factors. *Ophthalmology*, 118(5):812–818, 2011. 7

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

[28] Karren Dai Yang, Anastasiya Belyaeva, Saradha Venkatachalapathy, Karthik Damodaran, Abigail Katcoff, Adityanarayanan Radhakrishnan, GV Shivashankar, and Caroline Uhler. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nature Communications*, 12(1):1–10, 2021. 3

[29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, Miami, FL, USA, 2009. IEEE. 2

[30] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, ICCV '15, page 1422–1430, USA, 2015. IEEE Computer Society. 2

[31] Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, 9(3):e1003348, 2013. 3

[32] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *ICCV*, pages 9588–9597, October 2021. 2, 6, 7

[33] Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J Theis. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, 2019. 3

[34] FinnGen. Finngen. https://www.finngen.fi/en. Accessed: 2021-11-09. 1

[35] Kelly A Frazer, Sarah S Murray, Nicholas J Schork, and Eric J Topol. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–251, 2009. 3

[36] Huazhu Fu, Fei Li, José Ignacio Orlando, Hrvoje Bogunović, Xu Sun, Jingan Liao, Yanwu Xu, Shaochong Zhang, and Xiulan Zhang. Palm: Pathologic myopia challenge, 2019. 5, 6

[37] Yu Fujinami-Yokokawa, Hideki Ninomiya, Xiao Liu, Lizhu Yang, Nikolas Pontikos, Kazutoshi Yoshitake, Takeshi Iwata, Yasunori Sato, Takeshi Hashimoto, Kazushige Tsunoda, et al. Prediction of causative genes in inherited retinal disorder from fundus photography and autofluorescence imaging using deep learning techniques. *British Journal of Ophthalmology*, 2021. 3

[38] Leonardo Emberti Gialloreti, Matteo Pardini, Francesca Benassi, Sara Marciano, Mario Amore, Maria Giulia Mutolo, Maria Cristina Porfirio, and Paolo Curatolo. Reduction in retinal nerve fiber layer thickness in young adults with autism spectrum disorders. *Journal of autism and developmental disorders*, 44(4):873–882, 2014. 7

[39] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *CoRR*, abs/1803.07728, 2018. 2

[40] 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, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. 2, 6, 7

[41] Katharina Grünberg, Oscar Jimenez-del Toro, Andras Jakab, Georg Langs, Tomàs Salas Fernandez, Marianne Winterstein, Marc-André Weber, and Markus Krenn. *Annotating Medical Image Data*, pages 45–67. Springer International Publishing, 2017. 1

[42] Gregory Gundersen, Bianca Dumitrascu, Jordan T Ash, and Barbara E Engelhardt. End-to-end training of deep probabilistic cca on paired biomedical observations. In *Int. Conf. on AI and Stats.*, 2020. 3

[43] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterington, editors, *Int. Conf. on AI and Stats.*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. 2

[44] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 2

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[46] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020. 2

[47] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 2

[48] Jinrong Hu, Shanhui Sun, Xiaodong Yang, Shuang Zhou, Xin Wang, Ying Fu, Jiliu Zhou, Youbing Yin, Kunlin Cao, Qi Song, et al. Towards accurate and robust multi-modal medical image registration using contrastive metric learning. *IEEE Access*, 7:132816–132827, 2019. 3

[49] National Human Genome Research Institute. Dna sequencing. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet. Accessed: 2021-10-31. 3

[50] National Human Genome Research Institute. Genome-wide association studies fact sheet. https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet. Accessed: 2021-11-03. 2

[51] National Human Genome Research Institute. Genome-wide association studies fact sheet. https://www.ebi.ac.uk/gwas/. Accessed: 2021-11-03. 2, 7

[52] Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Self-supervised learning for spinal mris. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 294–302. Springer, 09 2017. 2

[53] Jianbo Jiao, Richard Droste, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. Self-supervised representation learning for ultrasound video. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1847–1850, 2020. 3

[54] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *CoRR*, abs/1902.06162, 2019. 2

[55] Justin Johnson, Andrej Karpathy, Fei Fei Li, and Djabeur Mohamed Seifeddine Zekrifa. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, pages 4565–4574, Las Vegas, NV, USA, 06 2016. IEEE. 2

[56] Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics*, 50(9):1219–1224, 2018. 3

[57] Matthias Kirchler, Stefan Konigorski, Matthias Norden, Christian Meltendorf, Marius Kloft, Claudia Schurmann, and Christoph Lippert. transfergwas: Gwas of images using deep transfer learning. *bioRxiv*, 2021. 3, 6, 7

[58] Lefteris Koumakis. Deep learning models in genomics; are we there yet? *Computational and Structural Biotechnology Journal*, 18:1466–1473, 2020. 3

[59] Scikit Learn. Roc auc micro averaging. https://scikit-learn.org/stable/modules/model_evaluation.html. Accessed: 2021-11-04. 6

[60] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, ESP Lung Project Team, David C Christiani, Mark M Wurfel, Xihong Lin, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, 2012. 3

[61] Hongming Li and Yong Fan. Non-rigid image registration using self-supervised fully convolutional networks without training data. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1075–1078, Washington, DC, USA, April 2018. IEEE. 2

[62] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[63] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–835, 2011. 3, 5

[64] Christoph Lippert, Riccardo Sabatini, M Cyrus Maher, Eun Yong Kang, Seunghak Lee, Okan Arikan, Alena Harley, Axel Bernal, Peter Garst, Victor Lavrenko, et al. Identification of individuals by trait prediction using whole-genome sequencing data. *Proceedings of the National Academy of Sciences*, 114(38):10166–10171, 2017. 2

[65] Lihao Liu, Angelica I Aviles-Rivero, and Carola-Bibiane Schönlieb. Contrastive registration for unsupervised medical image segmentation. *arXiv preprint arXiv:2011.08894*, 2020. 3

[66] Xingtong Liu, Ayushi Sinha, Mathias Unberath, Masaru Ishii, Gregory D. Hager, Russell H. Taylor, and Austin Reiter. Self-supervised learning for dense depth estimation in monocular endoscopy. *CoRR*, abs/1806.09521, 2018. 2

[67] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 2

[68] Jiechao Ma, Yang Song, Xi Tian, Yiting Hua, Rongguo Zhang, and Jianlin Wu. Survey on deep learning for pulmonary medical imaging. *Frontiers of medicine*, 14(4):450–469, 2020. 3

[69] Teri A Manolio. Genomewide association studies and assessment of the risk of disease. *New England journal of medicine*, 363(2):166–176, 2010. 2

[70] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009. 2

[71] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, Scottsdale, Arizona, USA, 2013. OpenReview. 2

[72] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, pages 6707–6717, 2020. 2

[73] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Int. Conf. on Machine Learning*, ICML'11, page 689–696, Madison, WI, USA, 2011. Omnipress. 2

[74] Kaida Ning, Bo Chen, Fengzhu Sun, Zachary Hobel, Lu Zhao, Will Matloff, Arthur W Toga, Alzheimer's Disease Neuroimaging Initiative, et al. Classifying alzheimer's disease with brain imaging and genetic data using a neural network framework. *Neurobiology of aging*, 68:151–158, 2018. 3

[75] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 69–84. Springer International Publishing, 2016. 2

[76] Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639, 2005. 1

[77] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, September 2018. 2

[78] Andrew Owens, Jiajun Wu, Josh H. Mcdermott, William T. Freeman, and Antonio Torralba. Learning sight from sound: Ambient sound provides supervision for visual learning. *Int. J. Comput. Vision*, 126(10):1120–1137, Oct. 2018. 2

[79] 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), 2021. 5, 6

[80] Sang Tae Park and Jayoung Kim. Trends in next-generation sequencing and a new era for whole genome sequencing. *International neurourology journal*, 20:S76, 2016. 1

[81] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *ICCV*, pages 9577–9587, 2021. 2

[82] 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. *Nature Biomedical Engineering*, 2(3):158–164, 2018. 6, 7

[83] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007. 6

[84] Senthil Purushwalkam and Abhinav Gupta. Pose from action: Unsupervised learning of pose features based on motion. *CoRR*, abs/1609.05420, 2016. 2

[85] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2

[86] Ralph Rapley and Stuart Harbron. *Molecular analysis and genome discovery*. John Wiley & Sons, 2005. 3

[87] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *MICCAI*, pages 234–241. Springer International Publishing, 2015. 6

[88] Tobias Roß, David Zimmerer, Anant Vemuri, Fabian Isensee, Sebastian Bodenstedt, Fabian Both, Philip Kessler, Martin Wagner, Beat Müller, Hannes Kenngott, Stefanie Speidel, Klaus Maier-Hein, and Lena Maier-Hein. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International Journal of Computer Assisted Radiology and Surgery*, 13, 11 2017. 2

[89] M Satue, M Seral, S Otin, R Alarcia, R Herrero, MP Bambo, MI Fuertes, LE Pablo, and E Garcia-Martin. Retinal thinning and correlation with functional disability in patients with parkinson's disease. *British Journal of Ophthalmology*, 98(3):350–355, 2014. 7

[90] Nawid Sayed, Biagio Brattoli, and Björn Ommer. Cross and learn: Cross-modal self-supervision. *CoRR*, abs/1811.03879, 2018. 2

[91] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Int. Conf. on Machine Learning*, pages 3145–3153. PMLR, 2017. 4

[92] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, NIPS'14, page 568–576, Cambridge, MA, USA, 2014. MIT Press. 2

[93] Sorensen-Dice. Dice score. https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient. Accessed: 2021-11-05. 6

[94] Hannah Spitzer, Kai Kiwitz, Katrin Amunts, Stefan Harmeling, and Timo Dickscheid. Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *MICCAI*, pages 663–671. Springer International Publishing, 2018. 2

[95] Hannah Spitzer, Kai Kiwitz, Katrin Amunts, Stefan Harmeling, and Timo Dickscheid. Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *MICCAI*, pages 663–671. Springer International Publishing, 2018. 3

[96] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):1–10, 03 2015. 1, 5, 6

[97] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019. 2

[98] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473, 2019. 2

[99] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Int. Conf. on Machine Learning*, pages 3319–3328. PMLR, 2017. 4

[100] N. Tajbakhsh, Y. Hu, J. Cao, X. Yan, Y. Xiao, Y. Lu, J. Liang, D. Terzopoulos, and X. Ding. Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1251–1255, 2019. 3

[101] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N. Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020. 3

[102] Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal self-supervised learning for medical image analysis. In Aasa Feragen, Stefan Sommer, Julia Schnabel, and Mads Nielsen, editors, *Information Processing in Medical Imaging (IPMI)*, pages 661–673. Springer International Publishing, 2021. 3

[103] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. In *NeurIPS*, Vancouver, Canada, 2020. 3

[104] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[105] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, pages 776–794. Springer, 2020. 2

[106] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 2, 5

[107] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports*, 11(1):1–13, 2021. 3

[108] Joost AM Verlouw, Eva Clemens, Jard H de Vries, Oliver Zolk, Annemieke JMH Verkerk, Antoinette am Zehnhoff-Dinnesen, Carolina Medina-Gomez, Claudia Lanvers-Kaminsky, Fernando Rivadeneira, Thorsten Langer, et al. A comparison of genotyping arrays. *European Journal of Human Genetics*, pages 1–14, 2021. 1

[109] Pierrick Wainschtein, Deepti Jain, Zhili Zheng, L Adrienne Cupples, Aladdin H Shadyab, Barbara McKnight, Benjamin M Shoemaker, Braxton D Mitchell, Bruce M Psaty, Charles Kooperberg, et al. Recovery of trait heritability from whole genome sequence data. *BioRxiv*, page 588020, 2021. 3

[110] John S Witte, Peter M Visscher, and Naomi R Wray. The contribution of genetic variants to disease depends on the ruler. *Nature Reviews Genetics*, 15(11):765–776, 2014. 2

[111] Di Wu, Deepti S. Karhade, Malvika Pillai, Min-Zhi Jiang, Le Huang, Gang Li, Hunyong Cho, Jeff Roach, Yun Li, and Kimon Divaris. *Machine Learning and Deep Learning in Genetics and Genomics*, pages 163–181. Springer International Publishing, 2021. 3

[112] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018. 2

[113] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Int. Conf. on Machine Learning*, ICML'15, page 2048–2057, Lille, France, 2015. JMLR.org. 2

[114] Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam P. Harrison, Mohammadhadi Bagheri, and Ronald M. Summers. *Deep Lesion Graph in the Wild: Relationship Learning and Organization of Significant Radiology Image Findings in a Diverse Large-Scale Lesion Database*, pages 413–435. Springer International Publishing, 2019. 2

[115] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010. 3

[116] Wanneng Yang, Hui Feng, Xuehai Zhang, Jian Zhang, John H Doonan, William David Batchelor, Lizhong Xiong, and Jianbing Yan. Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Molecular Plant*, 13(2):187–214, 2020. 2

[117] Menglong Ye, Edward Johns, Ankur Handa, Lin Zhang, Philip Pratt, and Guang Yang. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. In *The Hamlyn Symposium on Medical Robotics*, pages 27–28, 06 2017. 2

[118] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118, Athens, Greece, 12 2018. IEEE. 2

[119] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *CVPR*, pages 6995–7004, 2021. 2

[120] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 2, 6, 7

[121] Pengyue Zhang, Fusheng Wang, and Yefeng Zheng. Self supervised deep representation learning for fine-grained body part recognition. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 578–582, Melbourne, Australia, April 2017. IEEE. 2

[122] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 649–666. Springer International Publishing, 2016. 2

[123] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 2

[124] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *NeurIPS*, 33:18123–18134, 2020. 2

[125] Tao Zhou, Kim-Han Thung, Xiaofeng Zhu, and Dinggang Shen. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. *Human brain mapping*, 40(3):1001–1016, 2019. 3

[126] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B. Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *MICCAI*, pages 384–393. Springer International Publishing, 2019. 3

[127] Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng. Self-supervised feature learning for 3d medical images by playing a rubik's cube. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *MICCAI*, pages 420–428. Springer International Publishing, 2019. 3

[128] Marinka Zitnik, Francis Nguyen, Bo Wang, Jure Leskovec, Anna Goldenberg, and Michael M Hoffman. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50:71–91, 2019. 3

[129] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. A primer on deep learning in genomics. *Nature genetics*, 51(1):12–18, 2019. 3

[130] Igor Zwir, Javier Arnedo, Coral Del-Val, Laura Pulkki-Råback, Bettina Konte, Sarah S Yang, Rocio Romero-Zaliz, Mirka Hintsanen, Kevin M Cloninger, Danilo Garcia, et al. Uncovering the complex genetics of human character. *Molecular psychiatry*, 25(10):2295–2312, 2020. 2