Self-supervised Deep Learning Methods for Medical Image Analysis

Aiham Taleb

Universitätsdissertation
zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

in der Wissenschaftsdisziplin
Digital Health - Machine Learning

eingereicht an der
Digital-Engineering-Fakultät
der Universität Potsdam

Datum der Disputation:

Betreuer

Prof. Dr. Christoph Lippert
Hasso Plattner Institute for Digital Engineering, University of Potsdam, Germany
Hasso Plattner Institute for Digital Health at the Icahn School of Medicine at Mount Sinai, NYC, USA

Gutachter

Prof. Dr. Gerard de Melo
Hasso Plattner Institute for Digital Engineering, University of Potsdam, Germany

Prof. Dr. Shadi Albarqouni
University of Bonn, Germany
Helmholtz AI, Munich, Germany

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-64408
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-644089

Abstract

Deep learning has seen widespread application in many domains, mainly for its ability to learn data representations from raw input data. Nevertheless, its success has so far been coupled with the availability of large annotated (labelled) datasets. This is a requirement that is difficult to fulfil in several domains, such as in medical imaging. Annotation costs form a barrier in extending deep learning to clinically-relevant use cases. The labels associated with medical images are scarce, since the generation of expert annotations of multimodal patient data at scale is non-trivial, expensive, and time-consuming. This substantiates the need for algorithms that learn from the increasing amounts of unlabeled data. Self-supervised representation learning algorithms offer a pertinent solution, as they allow solving real-world (downstream) deep learning tasks with fewer annotations. Self-supervised approaches leverage unlabeled samples to acquire generic features about different concepts, enabling annotation-efficient downstream task solving subsequently.

Nevertheless, medical images present multiple unique and inherent challenges for existing self-supervised learning approaches, which we seek to address in this thesis: (i) medical images are multimodal, and their multiple modalities are heterogeneous in nature and imbalanced in quantities, e.g. MRI and CT; (ii) medical scans are multi-dimensional, often in 3D instead of 2D; (iii) disease patterns in medical scans are numerous and their incidence exhibits a long-tail distribution, so it is oftentimes essential to fuse knowledge from different data modalities, e.g. genomics or clinical data, to capture disease traits more comprehensively; (iv) Medical scans usually exhibit more uniform color density distributions, e.g. in dental X-Rays, than natural images. Our proposed self-supervised methods meet these challenges, besides significantly reducing the amounts of required annotations.

We evaluate our self-supervised methods on a wide array of medical imaging applications and tasks. Our experimental results demonstrate the obtained gains in both annotation-efficiency and performance; our proposed methods outperform many approaches from related literature. Additionally, in case of fusion with genetic modalities, our methods also allow for cross-modal interpretability. In this thesis, not only we show that self-supervised learning is capable of mitigating manual annotation costs, but also our proposed solutions demonstrate how to better utilize it in the medical imaging domain. Progress in self-supervised learning has the potential to extend deep learning algorithms application to clinical scenarios.

Zusammenfassung

Deep Learning findet in vielen Bereichen breite Anwendung, vor allem wegen seiner Fähigkeit, Datenrepräsentationen aus rohen Eingabedaten zu lernen. Dennoch war der Erfolg bisher an die Verfügbarkeit großer annotatierter Datensätze geknüpft. Dies ist eine Anforderung, die in verschiedenen Bereichen, z. B. in der medizinischen Bildgebung, schwer zu erfüllen ist. Die Kosten für die Annotation stellen ein Hindernis für die Ausweitung des Deep Learning auf klinisch relevante Anwendungsfälle dar. Die mit medizinischen Bildern verbundenen Annotationen sind rar, da die Erstellung von Experten Annotationen für multimodale Patientendaten in großem Umfang nicht trivial, teuer und zeitaufwändig ist. Dies unterstreicht den Bedarf an Algorithmen, die aus den wachsenden Mengen an unbeschrifteten Daten lernen. Selbstüberwachte Algorithmen für das Repräsentationslernen bieten eine mögliche Lösung, da sie die Lösung realer (nachgelagerter) Deep-Learning-Aufgaben mit weniger Annotationen ermöglichen. Selbstüberwachte Ansätze nutzen unannotierte Stichproben, um generisches Eigenschaften über verschiedene Konzepte zu erlangen und ermöglichen so eine annotationseffiziente Lösung nachgelagerter Aufgaben.

Medizinische Bilder stellen mehrere einzigartige und inhärente Herausforderungen für existierende selbstüberwachte Lernansätze dar, die wir in dieser Arbeit angehen wollen: (i) medizinische Bilder sind multimodal, und ihre verschiedenen Modalitäten sind von Natur aus heterogen und in ihren Mengen unausgewogen, z. B. (ii) medizinische Scans sind mehrdimensional, oft in 3D statt in 2D; (iii) Krankheitsmuster in medizinischen Scans sind zahlreich und ihre Häufigkeit weist eine Long-Tail-Verteilung auf, so dass es oft unerlässlich ist, Wissen aus verschiedenen Datenmodalitäten, z. B. Genomik oder klinische Daten, zu verschmelzen, um Krankheitsmerkmale umfassender zu erfassen; (iv) medizinische Scans weisen in der Regel eine gleichmäßigere Farbdichteverteilung auf, z. B. in zahnmedizinischen Röntgenaufnahmen, als natürliche Bilder. Die von uns vorgeschlagenen selbstüberwachten Methoden adressieren diese Herausforderungen und reduzieren zudem die Menge der erforderlichen Annotationen erheblich.

Wir evaluieren unsere selbstüberwachten Methoden in verschiedenen Anwendungen und Aufgaben der medizinischen Bildgebung. Unsere experimentellen Ergebnisse zeigen, dass die von uns vorgeschlagenen Methoden sowohl die Effizienz der Annotation als auch die Leistung steigern und viele Ansätze aus der

verwandten Literatur übertreffen. Darüber hinaus ermöglichen unsere Methoden im Falle der Fusion mit genetischen Modalitäten auch eine modalübergreifende Interpretierbarkeit. In dieser Arbeit zeigen wir nicht nur, dass selbstüberwachtes Lernen in der Lage ist, die Kosten für manuelle Annotationen zu senken, sondern auch, wie man es in der medizinischen Bildgebung besser nutzen kann. Fortschritte beim selbstüberwachten Lernen haben das Potenzial, die Anwendung von Deep-Learning-Algorithmen auf klinische Szenarien auszuweiten.

Acknowledgments

First and foremost, I want to express my gratitude to my supervisor Prof. Dr. Christoph Lippert for his continuous support and advice throughout my PhD research. Not only am I honored to have been accepted as his advisee in the first place, but also his guidance contributed a lot to my professional development as a researcher in the field. His comments have immensely improved the quality of my scientific publications and this thesis.

I would like to thank both Dr. Moin Nabi and Dr. Tassilo Klein, whom I had the chance to work with in my research internship at SAP SE in Berlin. I am grateful for all the advice and support they provided.

I also consider myself lucky to have worked with many amazing co-authors and colleagues in my papers, from whom I mention Matthias Kirchler, Benjamin Bergner, Remo Monti, Jana Fehr, Shahryar Khorasani, and Dr. Stefan Konigorski. I will never forget the great conversations we had and the fun times we spent together since the beginning of this group at the Max-Delbrück Center.

It is been also a great opportunity to have worked and supervised many talented students throughout my PhD, including Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, and Yamen Ali.

I am extremely fortunate to have met and learned from multiple prominent researchers, from whom I mention Prof. Dr. Jeanette Schulz-Menger, Dr. Falk Schwendicke, and Dr. Joachim Krois. Conversations with each one of you improved my understanding of the biomedical domains considerably.

The research summarized in this thesis has been supported by funding from the German Federal Ministry of Education and Research (BMBF) in the projects KI-LAB-ITSE (project number 01|S19066), SyReal (project number: 01IS21069A), and DAKI-FWS (project number: BMWi 01MK21009E). My position at the university as a research assistant was provided by Hasso-Plattner-Institute (HPI), which is funded by the Hasso Plattner Foundation.

Last but not least, I would love to thank my family, including my beautiful wife Marah Halawa, my loving parents (Zakaria Taleb and Bashira Almona), and my siblings (Humam, Nuha, and Ruba). My wife has been instrumental through all the hard moments in my PhD, and helped keep me alive and motivated. My parents have been supportive and encouraged me all the way even from overseas. I love all of you, without you this day would not have been possible.

Contents

1 Introduction

1.1 Motivation

Medical imaging plays a vital role in patient healthcare, as it aids in disease prevention, early detection, diagnosis, and treatment. As a consequence, unprecedented amounts of medical scans are being conducted on a daily basis. According to the World Health Organization (WHO), an estimated 3.6 billion diagnostic examinations are performed annually across the globe [bus; Org], and these estimations include only most commonly used types of medical imaging modalities. The rapidly rising numbers of imaging studies create a natural workload on human radiologists, justifying the increasing adoption of computer-aided diagnosis (CAD) and detection systems [Doi07]. The latter systems, in turn, can benefit significantly from recent advancements in deep learning methods and the growing evidence of its ability to improve performance on medical imaging applications [AMZ21; Ker+17; Kim+19b; Ma+20].

Nevertheless, machine learning, especially deep learning algorithms, are inherently data hungry, requiring sufficiently large annotated datasets to extract relevant information and learn rich data representations. Several studies show that deep learning methods require large amounts of annotated samples in order to match human diagnostic performance [Ard+19; Est+17; Gul+16; McK+20]. Subsequently, efforts to employ deep learning algorithms to support in clinical settings are often hampered by the high costs of required expert annotations. Generating expert annotations of medical imaging data at scale is non-trivial, expensive, time-consuming, and is associated with risks in privacy leakages. Even semi-automatic software tools may fail to sufficiently reduce annotation expenses [Grü+17]. Consequently, manual annotation of medical images is the main impediment in translating advancements in deep learning methods into clinically useful computer-aided diagnosis (CAD) systems. It is necessary therefore to find solutions for the aforementioned challenge.

1.2 Approach and Contributions

The lack of annotations is also common in other application fields of deep learning. One widely used technique to address this challenge is transfer learning, which

aims to reuse the features of trained neural networks on different, yet related, target tasks, such as adapting the features of networks trained on ImageNet [Den+09] into other visual tasks. To some extent, transfer learning has improved the performance on tasks with limited numbers of samples. However, despite several attempts to leverage ImageNet features in the medical imaging domain [IHA18; Raj+17; Sah+19; Wan+17], the difference in the distributions of natural and medical images seems significant. In other words, generalizing across these domains is questionable and can suffer from dataset bias [TE11]. A recent analysis [Rag+19] has also found that such transfer learning offers limited performance gains, relative to the computational costs it incurs.

All the above reasons, such as the ever-increasing quantities of medical scans, the large costs incurred in their annotation process, and the challenges in transfer learning models pretrained on different imaging domains, paint unsupervised (self-supervised) representation learning as a pertinent solution. Self-supervised approaches leverage unlabeled samples to acquire generic knowledge about different concepts, subsequently enabling annotation-efficient downstream task solving. In other words, pretrained models with self-supervised methods can be transferred (or fine-tuned) into downstream tasks, such as semantic segmentation, and they require fewer annotations therein. Nonetheless, the medical imaging domain presents several unique and inherent challenges when extending self-supervised learning approaches, which we seek to address in this thesis with our proposed methods.

First, medical images are multimodal, e.g. MRI and CT, and their multiple modalities are heterogeneous in nature and their quantities are imbalanced. Our proposed methods in chapter 3 aim to meet this challenge and utilize the multimodality property of medical scans. Second, medical scans are often multi-dimensional, such as in 3D instead of 2D, deeming designing self-supervised tasks for 3D spatial context a necessity. We show in chapter 4 that employing this property can improve downstream performance. Third, disease patterns in medical images are numerous and their incidence exhibits a long tail distribution, resulting in models biased towards the more prevalent disease traits. It is therefore essential to fuse knowledge from additional data modalities, e.g. genomics or clinical data, to capture rare disease traits more comprehensively. We show in chapter 5 that integrating genomic modalities with medical scans in the self-supervised stage can both improve downstream performance and achieve cross-modal explainability as a key byproduct. Finally, medical scans exhibit a more uniform nature than natural images, such as in color density distributions. chapter 6 illustrates how to take advantage of such domain knowledge in improving adoption of existing self-supervised methods in the medical imaging domain. Our proposed self-supervised methods meet these challenges, besides significantly reducing the quantities of required annotations.

1.3 List of published works and own contributions

Some extracts from this thesis appear in the following co-authored publications and preprints.
chapter 3 extends work from:

- Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. "Multimodal self-supervised learning for medical image analysis." *In International Conference on Information Processing in Medical Imaging, pp. 661-673. Springer, Cham, 2021.*

Own contributions in the above publication include source code implementation of the self-supervised multimodal puzzle solving algorithm, of subsequent downstream tasks (Brain Tumor segmentation, Prostate segmentation, Liver segmentation, and survival days prediction), and of the cross-modal generation task. This also includes all related evaluation scripts for trained models in terms of performance and annotation-efficiency. The writing of the manuscript is also an own contribution, including all sections and compilation and analysis of results.
chapter 4 is adapted from:

- Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. "3d self-supervised methods for medical imaging." *Advances in Neural Information Processing Systems 33 (2020): 18158-18172.*

- Yamen Ali, Aiham Taleb, Marina M-C. Höhne, and Christoph Lippert. "Self-Supervised Learning for 3D Medical Image Analysis using 3D SimCLR and Monte Carlo Dropout." *arXiv preprint arXiv:2109.14288 (2021).*

Own contributions in the first publication include source code implementation of the first version of five self-supervised algorithms and of the downstream task of Brain Tumor segmentation. This includes related evaluation scripts for trained models in terms of performance and annotation-efficiency. For the second version of algorithm implementations, own contributions included continuous supervision and review of developed source code. The writing of the complete manuscript is also an own contribution, including all sections and compilation and analysis of results. In the second preprint, own contributions included continuous supervision and review of developed source code as well as writing of the manuscript.
chapter 5 contains work from:

- Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. "ContIG: Self-supervised Multimodal Contrastive Learning for Medical Imaging with

Genetics." *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20908-20921. 2022.*

Own contributions in the above publication include source code implementation of the self-supervised multimodal contrastive loss and of the downstream tasks (Diabetic retinopathy classifiction, Pathological Myopia Segmentation, Retinal Fundus Disease Classification, and Cardiovascular Risk Prediction). This also includes related evaluation scripts for trained models in terms of performance and annotation-efficiency. Writing of manuscript sections and compilation and analysis of results is also a contribution.
And finally, chapter 6 is adapted from:

- Aiham Taleb, Csaba Rohrer, Benjamin Bergner, Guilherme De Leon, Jonas Almeida Rodrigues, Falk Schwendicke, Christoph Lippert, and Joachim Krois. "Self-Supervised Learning Methods for Label-Efficient Dental Caries Classification." *Diagnostics 12, no. 5 (2022): 1237.*

Own contributions in the above publication include source code implementation of the three self-supervised algorithms and of the subsequent downstream task of Dental Caries classification. This includes all related evaluation scripts for trained models in terms of performance and annotation-efficiency. The writing of the manuscript is also an own contribution, across all sections and compilation and analysis of results.

1.4 Thesis Structure

The thesis is outlined as follows. We first start with background information in chapter 2, which includes the necessary definitions and literature reviews on which our contributions in subsequent chapters are built upon. Our proposed novel self-supervised algorithms are then detailed in the following order: chapter 3 concerns learning from multiple modalities of medical images, chapter 4 covers learning from 3D medical scans, chapter 5 is about integrating medical images with genetic modalities, and chapter 6 deals with the homogeneous nature of medical images. In each chapter, we review the most relevant prior art, we provide the detailed formulations for our proposed algorithms and employed deep learning architectures. Finally, in chapter 7, we make our overall conclusions, we discuss our findings and limitations, and we suggest directions for future research.

2 Background and Related Work

In this chapter, we define some requisite concepts, which are used throughout this thesis. Additionally, we review relevant articles from the literature and provide a view on the state-of-the-art in the related domains.

2.1 Definitions

2.1.1 Representation

What is a Representation?

Humans have attempted for centuries to describe how our minds perceive the world around us. Philosophers, for instance, attempted to describe what exactly humans do with visual information sensed by our eyes. The doctrine of Representationalism (or Indirect Realism), in particular, holds the view that the world we see in conscious experience is not the real world itself, but merely a replica of that world in an internal representation [phi]. In other words, the immediate object of knowledge is an idea in the mind distinct from the external object [mer]. Immanuel Kant is viewed as a representationalist by A. B. Dickerson in his book "Kant on Representation and Objectivity" [Dic03], in which Kant's definition of the representation is:

> Representations are the immediate objects of our awareness. However, we cognize objects like trees neither by inferring those objects as the causes of representations nor by constructing them out of representations. Rather, via what Kant calls apperception we are made aware of the object cognized "in" the representation, just as we see a face "in" the lines of a picture. That object is distinct from the matter of the representation, just as the face is distinct from the lines themselves.

Another philosopher, David Hume wrote in his "Treatise of Human Nature" in 1740 [Hum03]:

> When I shut my eyes and think of my chamber, the ideas I form are exact representations of the impressions I felt; nor is there any circumstance of the one, which is not to be found in the other. In running over my

> other perceptions, I find still the same resemblance and representation. Ideas and impressions appear always to correspond to each other. ...I observe, that many of our complex ideas never had impressions, that corresponded to them, and that many of our complex impressions never are exactly copied in ideas. ...I have seen Paris; but shall I affirm I can form such an idea of that city, as will perfectly represent all its streets and houses in their real and just proportions? ...The qualities, from which this association arises, and by which the mind is after this manner conveyed from one idea to another, are three, viz. RESEMBLANCE, CONTIGUITY in time or place, and CAUSE and EFFECT. I believe it will not be very necessary to prove, that these qualities produce an association among ideas, and upon the appearance of one idea naturally introduce another.

In the passage above, Hume uses impressions to refer roughly to sensory information, and the representations (ideas) which correspond to the external objects exist purely in the mind. To Hume, such ideas persist in the mind even after the eyes are closed. Hume attempts to explain what is contained in these representations, and it seems that they are not mere copies of the visual world, rather transformed versions of it. Hume uses the example of Paris, where one may be able to find common elements that unite its architecture, even if they are not identical. Hume argues that the purpose of these representations or ideas is that they enable us to form associations. For example, it is completely possible to recognize if a building, unseen by us before, uses the architectural style of Paris. Hence, such associations are necessary to fetch the relevant past memories, and are likely maintained in our mind's representations of the world. In addition, Hume supposes here that a high quality representation resembles (is similar to) the different objects we see.

The above definition of the *representation* in the context of philosophy is a fairly accurate description of its usage in the context of Machine Learning. The term "features" is also alternatively used in this context. Similarly to the definitions above, the "features" or "representations" should resemble the world objects and scenes. In addition, such representations should be persisted in Machine Learning models. Such models should also be able to find associations between unseen object instances to those that exist in the persisted representations. As we will see in next sections, Machine Learning techniques, in general, and Deep Learning, in particular, aim to learn and persist the most relevant features and representations about the input data. Imitating the perception and cognition processes performed by our minds.

Representation Learning

Normally, a Machine Learning (ML) model's purpose is to perform a specific task, or sometimes multiple tasks. Such tasks, e.g. classification, rely on the quality of the data representations being used in this task [GBC16]. Generally, the better the data representations, the higher the efficiency of such tasks. In other words, better data representations would make the task easier. An essential question that arises here, which concerns the comparability of representations: what makes one representation better than another [BCV12]? A commonly used answer is that a good data representation is one that captures the underlying factors of variation relevant to the task, and discards the less relevant. Such factors of variation differ across the tasks, e.g. when analyzing an image of a car these factors may include the position of the car, its color, and the reflection of the sun light on it [GBC16]. Another example is the speaker accent in a speech analysis system.

Early attempts to extract the relevant factors of variation from the data relied on hand-crafted, or engineered, features, e.g. HOG [DT05] in visual inputs. Subsequently, these features are fed into the ML models to perform the actual tasks, e.g. train a classifier. Not only these engineered features required extensive understanding of the application domains, but also they failed to achieve satisfactory results in many tasks, particularly on large visual datasets [Ant+15a; KSH12]. On the other hand, the recent paradigm shift with Deep Learning (DL) algorithms, allowed the model to determine, or learn, which features are most relevant to optimize the task being solved. In other words, DL algorithms extract the important features from the data and store them as representations. DL attempts to mimic how the human brain extracts and persists representations of the world [CD14; Has+17]. Overall, uncovering the relevant factors of variation can be challenging, and deep learning addresses this problem by expressing complex data representations in terms of simpler representations. We briefly explain some DL concepts in the next Section.

Learning a good representation entails priors (clues) about the data domains to guide the model in attempting to uncover the useful factors of variation. Supervised learning approaches make use of strong clues in the form of labels (annotations) for input data samples. Typically these labels specify at least one factor of variation directly. Supervised learning proved to learn good representations [KSH12] from the data [Den+09], and these representations have been repurposed to other tasks [Gir+13]. This comes at the expense of requiring massive amounts of labels [Den+09]. In Unsupervised learning, on the other hand, less direct clues, about the variation factors, are usually used in order to learn more abstract representations of the world. This allows for exploiting the large quantities of unlabelled data samples, and hence reducing the numbers of required labels [DGE15]. Unsuper-

vised representation learning has the potential to learn good data representations while reducing human-expert annotation efforts at the same time, as we will see in subsequent Chapters.

2.1.2 Deep Learning

In order to better understand deep learning models, we will begin with simple linear regression [Gau09; PLA72]. Note that the examples used below are adapted from [Bis06; Gal16]. Given a set of N input-output mapping pairs $\{(x_1, y_1), ..., (x_N, y_N)\}$, e.g. house prices in relation to square areas. Now, we assume a linear function that connects each house area in square meters $x_i \in R^L$ to the house price $y_i \in R^M$. The model thereby is defined as a linear transformation of the inputs $f(x) = (xW + b)$, with W as an $L \times M$ matrix and b is a real vector with M elements. Here W, b are called the model parameters, and their values define different linear functions. Our target is to find the parameter values that would minimize (optimize) the mean squared error (MSE) for the observed data: $\frac{1}{N} \sum_{n=i} \|y_i - (x_i W + b)\|^2$. The MSE error function is called the objective or *loss* function.

Nevertheless, more generally, the relation between x and y may be non-linear, i.e the function $f(x)$ is a non-linear transformation. For this case, linear basis function regression [Bis06] can be used, where the input x is transformed with K scalar-valued non-linear functions $\varphi_k(x)$ to form a feature vector $\Phi(x) = [\varphi_1(x), ..., \varphi_k(x)]$. Then, the regression is performed on the transformed versions of the inputs, rather than the inputs. The basis functions φ_k can be polynomials x^k, sinusoidals, Gaussian basis functions, or even parameterized functions, e.g. $\varphi_k^{w_k, b_k}$. In the latter example, φ_k is applied to the inner-product of $\langle w_k, x \rangle + b_k$. For example, if $\varphi_k(.) = \cos(.)$, then $\varphi_k^{w_k, b_k}(x) = \cos(\langle w_k, x \rangle + b_k)$. From now on, we assume the basis functions φ_k are usually identical for all k. The respective outputs of the basis functions in the feature vector $\Phi(x)$ are again fed as inputs to the linear transformation. Hence, the model output can be rewritten as $f(x) = \varphi_k^{w_1, b_1}(x).W_2 + b_2$, where $\varphi_k^{w_1, b_1} = \varphi(W_1 x + b_1)$, and W_1 is a matrix of $L \times K$ dimensions, b_1 is a vector with K elements, W_2 a matrix of $K \times M$ dimensions, and b_2 a vector with M elements. Now, the regression parameters are W_1, b_1, W_2 and b_2, and our task becomes finding their values that minimize the MSE of $\|y - f(x)\|^2$.

Feed-forward neural networks

The quintessential deep learning model, which is the Multilayer Perceptron (MLP), AKA feed-forward neural network [RHW88] or *fully-connected* layers, can be described as hierarchies of the above parameterized basis functions to form layers,

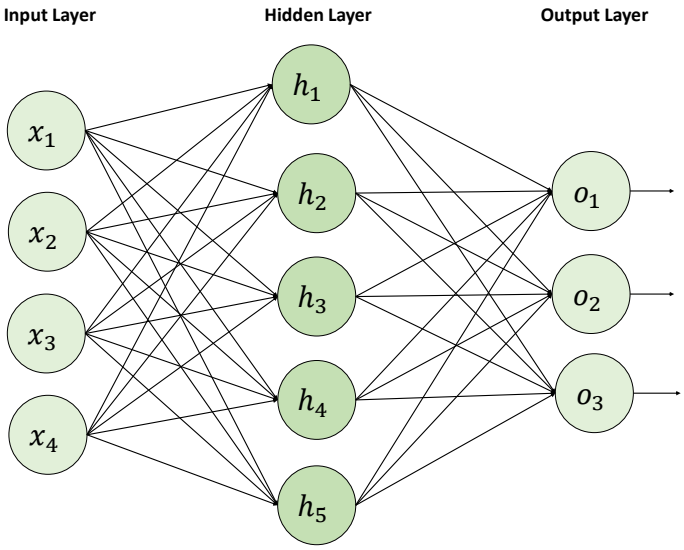


Figure 2.1: An MLP with three layers, including one hidden layer with five units.

as shown in Fig. 2.1. In such hierarchy, each layer outputs a feature vector that is fed to the next layer as input, and so on. They are called feed-forward because information flows through the function being evaluated from input, through the intermediate (hidden) layers in the network, till the output [GBC16]. Each hidden layer contains a weight matrix W and a bias vector b, both of which are used to transform the input x to obtain the layer's output $Wx + b$. Then, this output is transformed with an element-wise non-linear activation function $\sigma(.)$, such as rectified linear units (ReLU), Sigmoid functions, or hyperbolic tangent (tanh). Subsequent hidden layers take the transformed output of previous layers as inputs. This is repeated up to the output layer of the network, whose activation function depends on the task being solved. For regression, a simple linear or ReLU function may be used. For binary classification (binomial output distributions) usually the logistic function (Sigmoid) is used. For multi-class classification (multinomial output distributions), a Softmax function is used. Fig. 2.2 summarizes widely used activation functions.

On the capacity of deep neural networks

In the MLP, or any type of neural network, the number of hidden layers is called the network depth. Such depth is often used as an intuitive measure for model

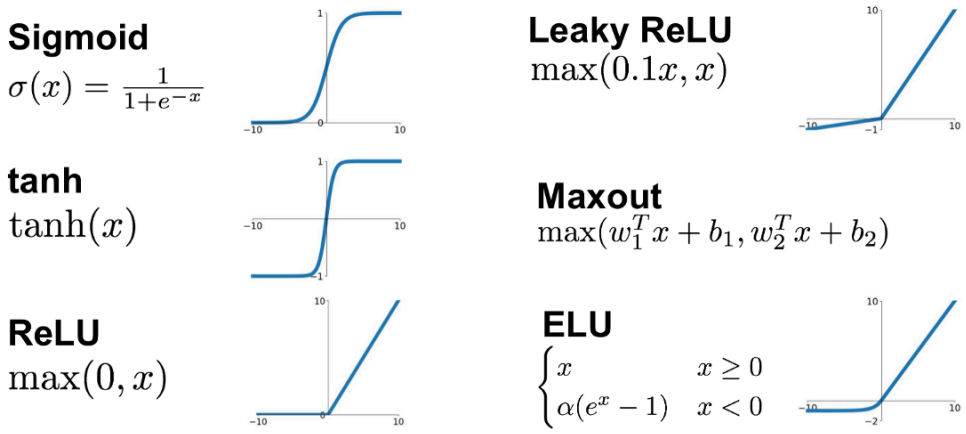


Figure 2.2: Commonly used activation functions in neural networks. Source: [J]

expressiveness, which relates to the complexity of the function the model can learn [BL07]. Normally, deeper models are widely considered to be more powerful, meaning they would be able to learn more complex functions, i.e. the model capacity is larger. Another factor that influences the model capacity is the number of units, or neurons, within each hidden layer. Increasing such number would affect the network width. This certainly comes with a computational expense, and excessively large models may exhibit overfitting. Therefore, many techniques attempt to handle overfitting situations, such as by searching for the optimal model *architecture* or by limiting the magnitudes of model weights, which is usually referred to as *regularization*. In contrast, it may happen that the neural network model is not deep enough, i.e. has limited capacity, which in turn results in underfitting. Many of these aspects are controlled by scalar values, e.g. network depth, which are usually referred to as *hyper-parameters*. In many applications, practitioners often resort to heuristics to reduce the search efforts for the optimal neural network setup.

On the optimization of deep neural networks

As mentioned earlier, training a neural network means finding the values of the its parameters (weights and biases) that would optimize (minimize) the chosen loss function, e.g. MSE. In general, deep neural networks are trained using a form of the stochastic gradient descent (SGD) algorithm accompanied with a form of the Back-propagation algorithm [BDD63]. Here, the SGD algorithm is used to perform

the learning (optimization) using the gradients of the loss function, while the Back-propagation is used to actually calculate the gradients. Technically, a neural network defines a graph, as shown in Fig. 2.1, and computing the gradients of the loss would entail applying the chain rule of differentiation, i.e. the gradients of each layer would require the gradients of subsequent layers. Hence, formulating the gradients may produce many duplicate terms, resulting in redundant gradient calculations if SGD is to be used alone. Therefore, the Back-propagation, which is a dynamic programming algorithm, aims to avoid such re-computations. Many alternatives for the back-propagation algorithm exist, however they use similar differentiation strategies with dynamic programming techniques. In theory, obtaining the optimal weights can be performed by using any optimization algorithm, e.g. evolutionary methods. Nevertheless, most deep learning methods employ some form of the back-propagation algorithm, due to its speed in training the models.

Other techniques of speeding up the training process also exist, such as extensions of the SGD optimization algorithm itself. Here, before we mention examples of these extensions, we should describe briefly the update rule employed by SGD [RM51]:

$$w_{t+1} = w_t - \eta \nabla \ell(w_t) \tag{2.1}$$

Here, the above equation is only the update rule of the weight parameter w_t, which has a specific value at time point t and the aim is to estimate its next value at point $t + 1$. The network loss is $\ell(w_t)$, for which we compute its gradient with respect to $\nabla \ell(w_t)$. Then, this gradient is subtracted from w, after being scaled by a hyper-parameter that is called the learning rate η. This scalar determines how far to move with each gradient in each iteration. The above equation is at the heart of the SGD algorithm, where the gradients of the loss (w.r.t. the model parameters) guide the training process to reach a minimum point in the loss function plane. The several extensions [Cho+19] of the SGD algorithm are usually modifications of, or additions to, the update rule in Eq. (2.1). These are commonly called *optimizers*. For instance, momentum methods [Nes83; Pol64] add a constant multiple of the previous parameter update, to encourage faster training progress. Other more sophisticated optimizers maintain adaptive per-parameter learning rates, such as AdaGrad [DHS11], RMSProp [TH12], and Adam [KB14b]. These latter optimizers mainly differ in how they update the per-parameter learning rates, e.g. based on the recent magnitudes of the gradients for each parameter. In short, as can be already observed, all optimization algorithms in Deep Learning rely on SGD, and it is common to use heuristics in deriving the update rules of such optimizers.

Common types of deep neural networks

While MLPs offered simplicity in their design, they oftentimes become inefficient when processing high-dimensional sparse input data, e.g. images, videos, or natural language sequences. This mainly is caused by their fully-connected nature, where each neuron in each layer is assumed to be connected to every neuron in previous and subsequent layers, as Fig. 2.1 shows. Nowadays, MLPs are still widely-used, e.g. to vectorize their inputs for classifiers or for tabular datasets, but mostly along with other specialized types of neural network layers. A specialized neural network type is called Convolutional Neural Networks (CNN), which is widely used to process signal inputs, e.g. images and audio. Due to their favorable characteristics, we rely on CNNs as model architectures in most of our methods, hence we elaborate on them in the next paragraph. Other specialized neural networks are sequence models, such as Recurrent Neural Networks (RNN) [RHW88; Wer88], which are usually suitable for processing data with sequence nature, e.g. natural language and time-series. RNNs, and their more successful variant LSTMs [HS97], were able to process and learn from sequence inputs, especially in setups called sequence-to-sequence (Seq2Seq in short) [SVL14] such as in language translation. However, developments in attention techniques, namely self-attention in Transformer models [Vas+17], allowed for improved performances in Seq2Seq applications. The limitation in RNN-based models that Transformers overcame was mainly their sequential word-by-word processing. The non-sequential nature of Transformers allowed processing input-output sequences in parallel, and also enabled longer dependencies within sequences. Transformer architectures allowed training large language models, e.g. BERT [Dev+18] and co., which considerably advanced the performance on Natural Language Processing benchmarks. Transformer-based models have also recently found applications in computer vision [Dos+20; Tou+21]. However, admittedly, such architectures require tremendous amounts of input images to match the performances achieved with CNNs on image benchmarks, due to favorable data-efficiency properties inherent to CNNs. More details in the next section.

Convolutional Neural Networks (CNNs)

CNNs [LeC+89a; RHW88] are deep learning models that excel at processing data inputs with a known grid-like topology [GBC16]. Examples include one-dimensional time-series, two-dimensional images, and even three-dimensional images. Despite being invented since the 1980s, their popularity recently increased [KSH12] after enabling solving image classification tasks that were considered beyond our reach. The AlexNet [KSH12] model consisted of consecutively applied convolution and

pooling layers, followed by the fully-connected layers. Since then, several improvements and advancements occurred on CNNs, and they appeared to succeed in several computer vision applications afterwards, such as object detection [Gir+13], semantic segmentation [LSD15], image retrieval [Gon+14], human action recognition from videos [Bac+11; Ji+12], and many more. CNNs have even found success in natural language processing, speech recognition, and time-series forecasting tasks [Gu+18].

In mathematics, the convolution is an operation between two functions of real-values. Assume the two functions x and h for instance, their convolution is expressed by $x * h$, with an asterisk or star. This operation is defined as the integral of the product of the two functions after one is reversed and shifted:

$$y(t) = (x * h)(t) = \int_a^b x(\tau)h(t-\tau)d\tau \tag{2.2}$$

Where x is called the input function or signal, h is called the system function, y is the output signal, and τ denotes the time or count for example. The system function h can be any desired function, and is usually chosen to have properties useful for the application. In the terminology of convolutional neural networks (CNNs), and assuming 2D images as inputs, the input function x is simply the input, the system function h is referred to as the kernel, and the output function y is called the feature map. A simple convolutional layer is illustrated in Fig. 2.3. Here, the convolutional layer is a linear transformation that preserves spatial information in the input image, and it consists of multiple kernels that are stacked together, similar to image color channels. An essential note here is that the values in each convolutional kernel are not fixed in each spatial location, they are in fact learned to capture simple, e.g. edges, or more complex features, e.g. human faces, from input images. Within a CNN, such convolutional layers are stacked consecutively to form a hierarchical architecture, similar to MLPs described before. Hence, it becomes intuitive that the learned features in earlier layers of the CNN are fine-grained edges or corners, and the complex features appear in the higher layers as coarse-grained objects. Illustrations of these features across the layers can be found in the work of Zeiler *et al.* [ZF13], we include an example in Fig. 2.4.

Another essential building block in CNNs is the pooling layer, which operates on the feature maps produced by convolution layers and reduces their dimensionalities. Such dimensionality reduction is a feature selection technique as a matter of fact, and is performed by simply applying summary statistics of nearby values, such as max pooling [ZC88] and average pooling [LeC+89b]. Max pooling computes the maximum of (n, n) input blocks of pixels, whereas average pooling computes

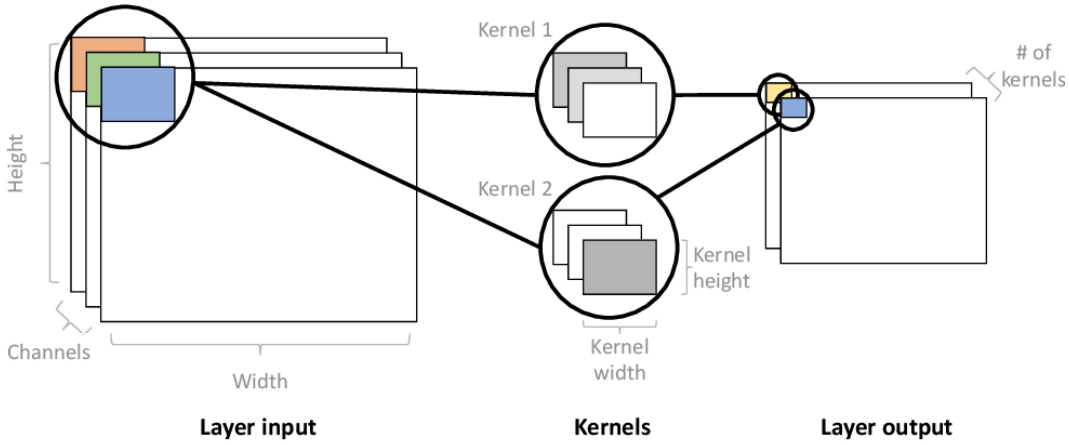


Figure 2.3: A convolution process by a kernel in a convolution layer of a CNN. Source: [Gal16]

the mean of input blocks of pixels. Not only pooling layers are important for feature selection, but also they add Translation Invariance to CNNs, i.e. the features obtained by convolutional layers are described as invariant to translation. This property means that detected objects may be anywhere in input images. Translation Invariance is one of reasons behind the attractiveness of convolutional layers when processing image inputs, as it improves the data-efficiency of CNN models, thus requiring less input images to learn how a certain object looks. In reality, the types of Invariances to different image transformations, see examples in Fig. 2.5, are all meant to inflict the same data-efficiency behavior. The other types of transformation invariances, e.g. to rotation, are usually added to CNNs, or any other architecture for that matter, by artificially creating (augmenting) the training dataset with transformed image versions.

Transfer Learning

The recent wave of interest in Deep Learning models, which was revived mainly in the Computer Vision field with the introduction of AlexNet [KSH12], can be attributed also to the employment of high-performing Graphical Processing Unit (GPU) machines. Efficient implementations of convolution operations and distributed training on GPUs have enabled faster training procedures on large datasets such as ImageNet [Den+09]. However, training such models from scratch requires

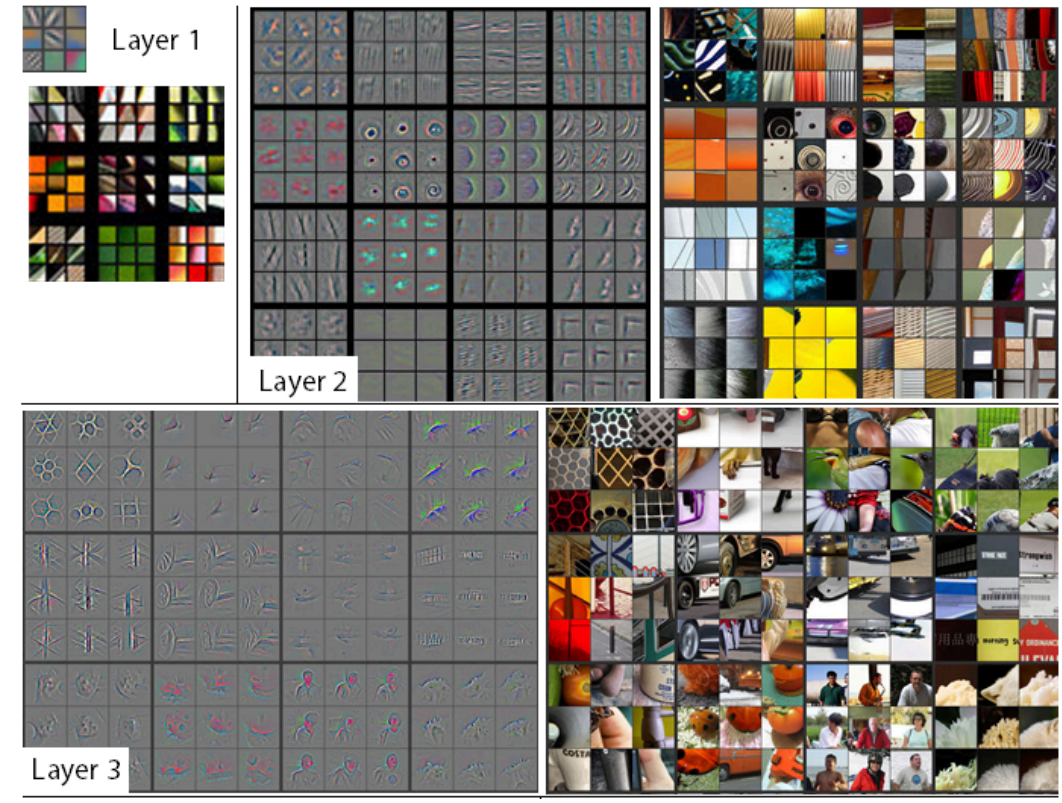


Figure 2.4: Visualizations of CNN layer features. Source: [ZF13]

significant resources as well as long training schedules. In addition, it has been found that the learned hierarchical representations (features) from such large datasets are rich with semantic information about the different classes in the dataset [Gir+13; Gon+14; LSD15], see Fig. 2.4. Therefore, all these reasons, prompt the idea of repurposing, or rusing, the learned representations in these models on different, often smaller, datasets. This is referred to as Transfer learning, in which the model parameters are fine-tuned or refined on other datasets or domains. Fig. 2.6 illustrates a simplified CNN, where the network consists of two main parts: a convolutional base for feature extraction, and a classifier part. Transfer learning would usually mean replacing the latter part, the classifier, and keeping (reusing) the convolutional part. It is also possible to unfreeze the weights in the convolutional part and fine-tune their weights slightly, especially if the new domain is semantically different from the source domain (the one on which the model had been trained on). The story of

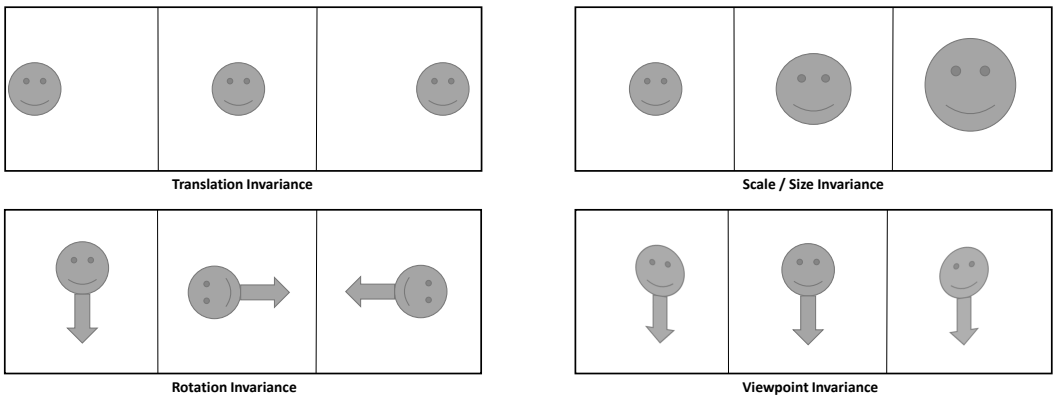


Figure 2.5: Examples of invariance to various transformations.

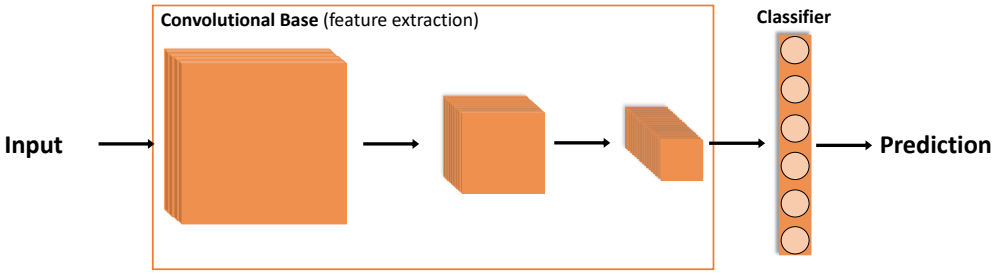


Figure 2.6: Architecture of a model based on Convolutional Neural Network.

transfer learning from ImageNet has sparked an interest in learning representations in an unsupervised manner, from large unlabeled datasets too, as we will see in subsequent sections.

2.1.3 Modality

We, as humans, appear to experience the world in a multimodal manner: we see and observe objects around us, we hear the sounds they make, we smell odors and scents, we touch and feel the textures, and we taste the different flavors. Therefore, a *modality* is a term used to reflect how certain things are experienced. For instance, when attempting to understand what another human is speaking, we most times need both audio and visual information to be able to fully understand them.

Similarly, to improve Machine Learning or Deep Learning methods' understanding of the world, processing such multimodal inputs is required.

In literature, a sub-field in AI that is often referred to as Multimodal ML or DL subsumes the different related aspects. Works in Multimodal DL are numerous [BAM19; Bay+21; Ngi+11; Sum+21]. Multimodal learning comes with several inherent challenges, to name a few:

1. Multimodal fusion: where it is usually asked whether to fuse the information from each modality at a stage before feature extraction (early fusion) or afterwards (late fusion).

2. Alignment: here the correspondences across the modalities are investigated, e.g. aligning the words in a video caption with the visual actions and spoken signals in the audio.

3. Representation: finding the most informative way to represent the knowledge from combined modalities, which are usually heterogeneous in nature, e.g. language is symbolic but vision modalities are signals.

Learning multimodal representations is essential in solving many applications.

Modalities of Medical Imaging

It is necessary to make a distinction between the notion of data modality, which we defined above, and the medical imaging modality. A medical imaging modality is also another view on the objects, here body organs or tissues. However, the difference is that medical imaging modalities are forms of images, and they usually are used depending on the medical use-case. For instance, as shown in Fig. 2.7, X-Ray scans highlight bone structures, but MRI scans are able to capture soft tissues more clearly. Radiologists and physicians request specific scan types depending on the situation, and to have more information on certain tissues and organs, e.g. some cancer tissues only appear in specific variants of MRI. The reader is kindly referred to more specialized resources for additional information about different medical imaging modalities [AYM14; Bey+20; EJ16; EM11].

2.2 Deep Learning in Medical imaging

Medical imaging plays an essential role in patient healthcare, as it allows physicians to view the patient anatomy and therefore diagnose conditions otherwise unfeasible without such medical scans. As explained in previous section, and as shown in

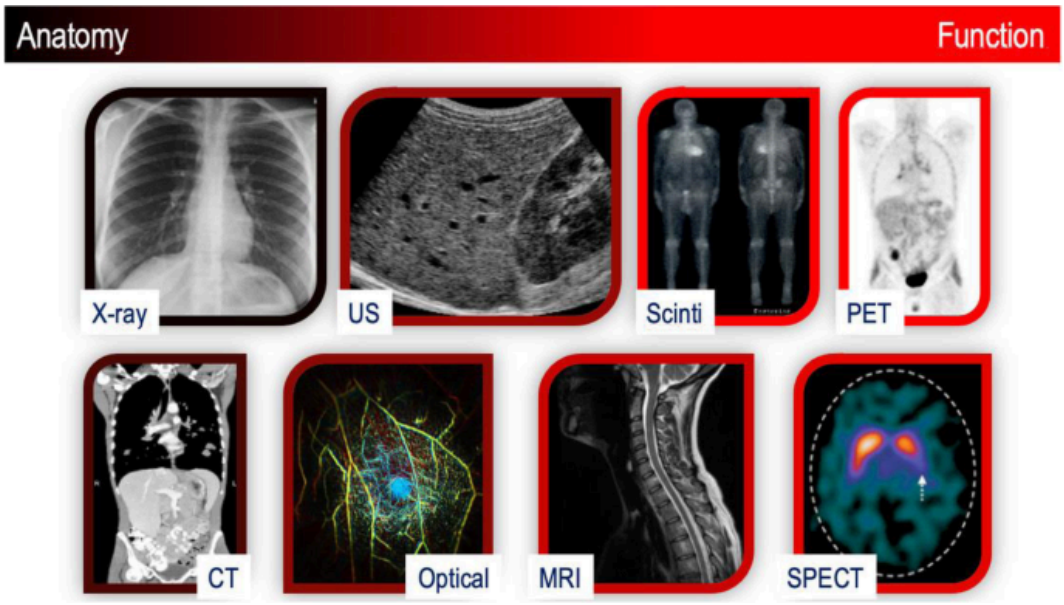


Figure 2.7: Examples of commonly used medical imaging modalities. Source: [Bey+20]

Fig. 2.7, medical imaging technologies are diverse, and are acquired based on the condition type. Such advancements in medical imaging acquisition technologies also comes with an expense: an expert professional is required to interpret the scans. A task that is usually costly, in terms of effort and time. As a result, the success of deep learning in computer vision domains warranted its extension to medical imaging domains. One of the main advantages of applying deep learning on medical scans is that it expects raw images as inputs, and the corresponding labels. In other words, when using deep learning to interpret input scans, it is not a prerequisite to know the scanning technology in detail. Certainly, knowing such information may help in designing a more appropriate model architecture, but it is not a necessary condition.

As a result, due to its appealing properties, deep learning has been receiving a continuing interest in medical imaging applications. A comprehensive review of the studies that investigated deep learning applications in the medical imaging domain is out of the scope of this thesis, thus the reader is referred to more specialized articles [AMZ21; Ker+17; Kim+19b; Ma+20]. However, we summarize below some application categories for deep learning in medical imaging:

1. **Image classification and object detection:** classifying the existence and

the type of a disease from medical scans is a common task for deep learning. It has seen interest in diagnosis of tuberculosis and similar diseases from Chest X-Rays [LS17], diabetic retinopathy from eye Fundus images [Gul+16], skin cancers [Est+17], and many more. Basically classifying lesions and anomalies in almost every medical imaging modality is possible with deep learning methods. To make the task even more useful, such lesions and anomalies are also oftentimes located in the scans with object detection and region proposal techniques [Kim+19b; Liu+17; ZDG19].

2. **Semantic segmentation:** a more useful deep learning task for radiologists is segmenting the lesions pixel-wise. Many studies have attempted to segment brain tumors [Per+16], lung cancer [Gor+18], and many cells and organs [RFB15].

3. **Image generation and translation:** a specific deep learning application, which receives continuous attention, is image generation with generative adversarial networks (GANs) –a specialized architecture for such task–. This application allows for generating almost realistic images from deep networks. In the medical imaging domain, these models have also found use-cases, especially in translating across medical imaging modalities [Arm+19; JRP19; Wol+17; Yan+18; ZYZ18]. This becomes particularly useful when certain modalities are more expensive or harder to obtain.

4. **Image registration:** this application is widely used in the medical imaging domain as a preprocessing step for other tasks, e.g. segmentation. It refers to the spatial matching of images, here medical imaging modalities, based on their contents. A review of such works is in [Fu+20].

While applications in the medical imaging domain can considerably benefit from deep learning techniques, e.g. to aid in diagnosis, extending them to practical clinical use-cases faces a major impediment, the requirement of the expert annotations [Grü+17; Ker+17], as mentioned earlier. The annotation process in medical imaging application is similar to that used for natural imaging; it means assigning labels to input images in preparation for training datasets in pairs of data and labels. However, natural imaging datasets are usually collected from numerous photos obtained from various sources, such as social media, and annotating such images is often possible by non-experts via crowd-sourcing [Kov+16]. On the other hand, annotating medical images requires expert knowledge and skills, which usually only domain professionals, e.g. radiologists, possess. Therefore, manual medical image annotation is an expensive process, and as a result fewer numbers of labels

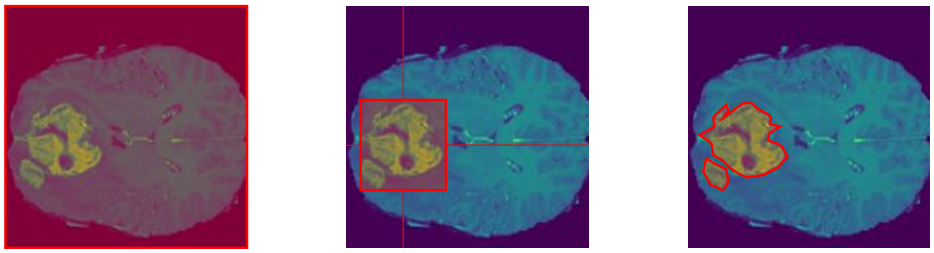


Figure 2.8: The type of annotation relies on the application category (the downstream task), and this in turn influences the associated labelling costs. This figure shows examples of brain tumor annotation for the tasks of classification, detection, and segmentation. The nature of the acquired labels range from the existence of the nodule, its location, and an accurate contour of its boundary (from left to right in the figure). It is apparent that the finer in detail the annotation is, the more expensive it becomes. Scans are from the BraTS [Bak+17; Men+15] dataset.

exist for the medical imaging domain. Fig. 2.8 illustrates the common annotation types in medical imaging.

As we will see throughout this thesis in subsequent sections, we address this challenge in order to improve the applicability of deep learning methods in the medical imaging domain, by relaxing this constraint of annotation.

2.3 Self-supervised Representation Learning

As mentioned earlier when defining transfer learning in the context of deep learning, reusing the learned representations from the ImageNet [Den+09] dataset allowed for both improved performances and reduced annotation quantities across multiple tasks and benchmarks. However, constructing a dataset the scale of ImageNet, which has almost 1.6 million labeled images categorized across 1,000 classes, is certainly not an easy nor an inexpensive job, oftentimes requiring expensive crowd-sourcing platforms [Kov+16] to accomplish the task. Hence, it is become a necessity to turn to the promise of learning data representations from virtually infinite amounts of data available online with unsupervised methods.

Unsupervised representation learning methods are generally of two types: generative and discriminative. Generative approaches build a distribution (a density function), over the data with the aim to generate realistic images, mainly.

These approaches may also learn latent embeddings, which they use as image representations. Depending on the type of the density function learned by generative models, they can learn explicit density such as variational auto-encoders (VAEs) [KW13; Vin+08] or implicit density such as generative adversarial networks (GANs) [DKD16; Goo+14]. With the goal of generating or hallucinating realistic images in mind, generative approaches operate directly in the pixel space. This deems them computationally expensive, and also learning pixel-level details that are mostly unnecessary for learning semantic data representations. That is not to say that generative approaches have not been explored in the representation learning direction, we mention a few works below.

In discriminative approaches for unsupervised representation learning, usually called self-supervised learning (SSL) methods, the aim is to construct (learn) a representation (embedding or feature) space, in which data samples that are semantically similar are encouraged to come closer to each other while moving farther from dissimilar samples, resembling clusters of data. SSL methods learn this embedding space by first deriving and solving a supervised proxy (auxiliary) task from the unlabeled data. A clear illustrative task is colorization [ZIE16], if the colors of the training images were to be removed and the deep learning model were to repaint them with the correct colors. Here, the source of supervision signal is the data itself, no human expert annotation is required. The resulting semantic representations from such task are in the form of neural network weights, and will also be useful for other real-world downstream tasks, afterwards. This two-phase scheme of SSL methods is depicted in Fig. 2.9.

SSL methods have been explored in multiple application fields [JT20], with roots in the natural language processing field in the Word2Vec [Mik+13] algorithm, which uses the context of the word as a source of supervision signal. A comprehensive review of SSL methods can be found in more specialized surveys [JT20; Liu+21; WK]. In this section, we review few relevant works that operate on image inputs mainly, as most of our methods process images as inputs. SSL methods differ mainly in the type of the proxy task used to learn data representations, as illustrated in an example taxonomy in Fig. 2.10. We review some relevant SSL methods below.

Note: In this thesis, we use the terms self-supervised learning and unsupervised learning interchangeably, as both terms usually refer to the same concept. However, self-supervision is not to be confused with self-training. The latter term is used in semi-supervised learning approaches [RHS05;

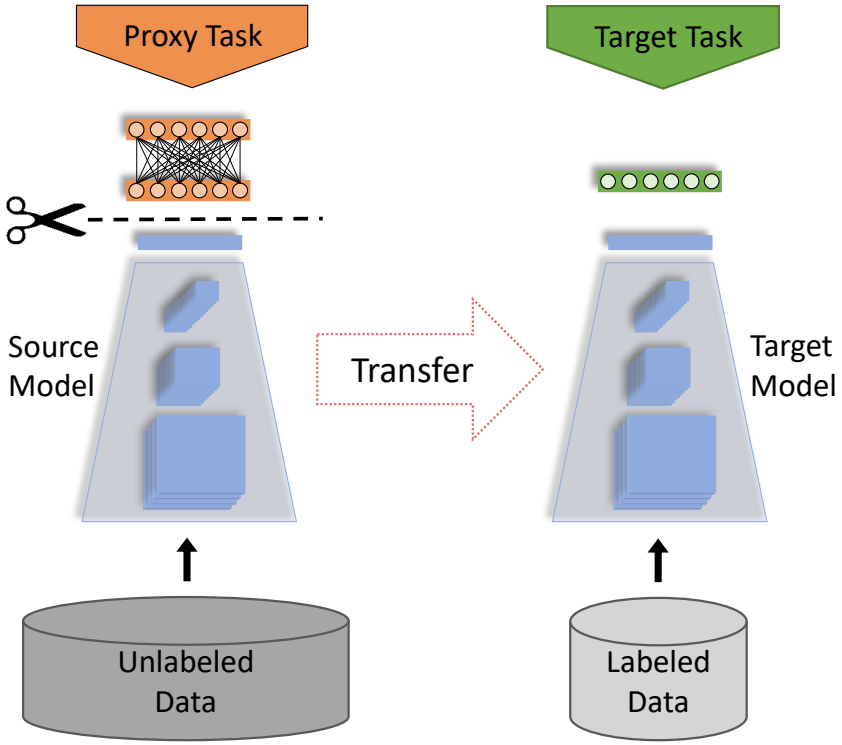


Figure 2.9: Flowchart of self-supervised learning stages. First, a deep learning model is trained on unlabeled data with a proxy task. Then, the obtained representations are transferred into a target downstream task. Figure source [Tal+22b], reprinted with permission.

Xie+20] often to refer to pseudo-labelling [Lee+13].

2.3.1 Pretext Task Learning

Pretext task learning methods rely on auxiliary handcrafted tasks to learn data representations. Disregarding if generative or discriminative, the effective difference across these tasks is the source of supervision signals. A commonly used source is the spatial context in the images, inspired from the skip-gram Word2Vec [Mik+13] algorithm. Doersch *et al.* [DGE15] generalized this idea to computer vision to learn a visual representation by predicting the position of an image patch relative to another (Fig. 2.11 (a)). Noroozi *et al.* [NF16] extended this patch-based approach to solve Jigsaw puzzles on natural images as a proxy task (Fig. 2.11 (b)). Other

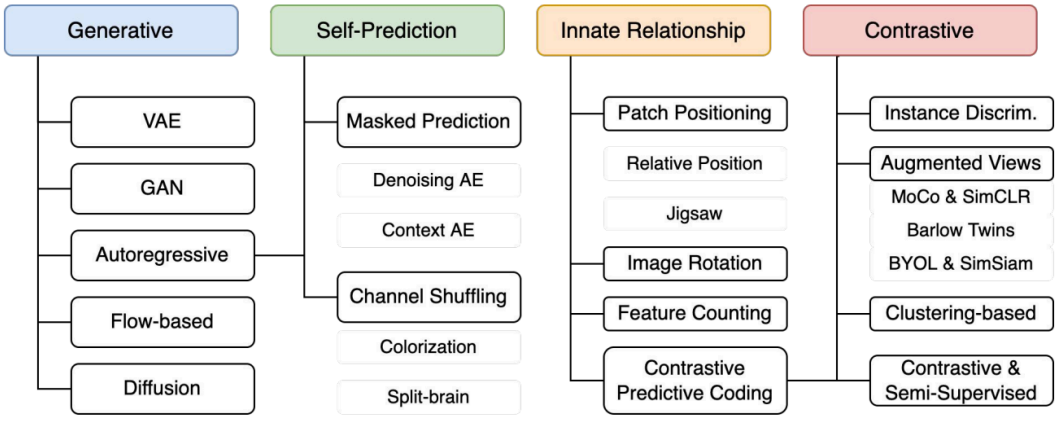


Figure 2.10: Self-supervised proxy tasks taxonomy. Source: [WK]

supervision sources, other than spatial context, were also employed. Examples include image colors [ZIE16] (Fig. 2.11 (c)), image rotation prediction [GSK18] (Fig. 2.11 (d)), masked pixels in-painting [Pat+16] (Fig. 2.11 (e)), clustering with k-means [Car+18] (Fig. 2.11 (f)), and transformed versions of images [Dos+14] (Fig. 2.11 (h)). One challenge that may appear in solving such pretext tasks is the tendency of deep networks to focus on shortcuts related to solving the task successfully, and hence compromizing the generalization of the learned features. Misra *et al.* [MM20] (Fig. 2.11 (g)) attempted to address such limitation by learning pretext-invariant data representations. In chapter 3 and chapter 4 we propose several pretext tasks that prove useful in the medical imaging domain.

2.3.2 Contrastive Learning

Overall, pretext task learning methods have shown evidence that learning semantically rich data representations from unlabeled images is possible. Nevertheless, designing handcrafted auxiliary tasks turned out to be a difficult task, after all, there is only so many types of supervision signals one can derive from still images. In addition, the performance of SSL methods in downstream tasks was still behind supervised counterparts, justifying the search for more informative tasks.

Contrastive Predictive Coding (CPC) approaches in [Hen20; OLV18] follow an auto-regressive way to classify future or next "positive" representations among a set of unrelated "negative" samples (Fig. 2.12 (a)). Here, the negative samples are simply image patches drawn from other locations in the image or even from

other images. The core concept behind CPC, which made them learn better data representations, is the InfoNCE loss [OLV18]. This loss maximizes the mutual information between related signals, here images, in contrast to others, and is inspired from Noise Contrastive Estimation [GH10]. The InfoNCE loss has been at the core of many subsequent contrastive learning methods [Che+20a; Che+20b; Dwi+21; He+20; Hje+18; Wu+18b]. The main difference is that more advanced methods use whole images as positive and negative samples, instead of patches.

The latter contrastive methods have also been inspired by Exemplar networks [Dos+14] (Fig. 2.11 (h)) in learning transformation-invariant representations (see Fig. 2.5). Exemplar networks create an artificial class for each image using image transformations (augmentations), which can become exhaustive if the number of images in the dataset is large. On the other hand, contrastive methods, e.g. SimCLR (Fig. 2.12 (b)), utilize the InfoNCE loss, mentioned above, which simplifies the problem into a binary classification task of positive versus negative samples. Contrastive approaches, particularly [Che+20a; He+20], have improved the performances of SSL on natural imaging benchmarks [Den+09; Eve+10]. Nonetheless, contrastive methods that rely on the InfoNCE loss require sufficient numbers of negative samples. SimCLR [Che+20a] draws these negative samples from the same mini-batch, necessitating a larger batch size. MoCo [He+20] draws negative samples from a memory bank that is internally a FIFO queue, allowing larger numbers of negative samples. Few recent works [Car+20; CH21; Gri+20; Zbo+21] attempt to avoid the negative sampling (mining) mechanism, which can be computationally expensive. We elaborate more on these approaches in chapter 5 and chapter 6, where we create our own contrastive methods that prove effective in the medical imaging domain.

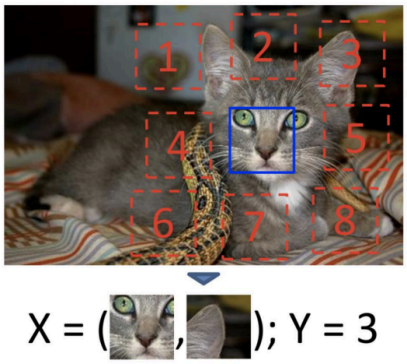
2.4 Self-supervision in the Medical Context

In the medical imaging domain, SSL methods have also witnessed a recent surge of interest [Taj+20; Xu21], due to the high cost incurred in annotating medical scans. Early attempts of extending self-supervision to medical images focused on particular use-cases and made assumptions about the data. Example works include depth estimation in monocular endoscopy [Liu+18], robotic surgeries [Ye+17], medical image registration [LF18], cardiac image segmentation [Bai+19], body part recognition [ZWZ17], disc degeneration using spinal MRIs [JKZ17], body part regression for slice ordering [Yan+19], and many others [Roß+17; Spi+18].

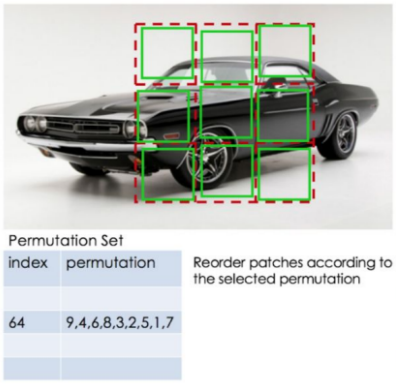
Several other works also proposed pretext tasks for the medical imaging domain, which improved the generalization of learned representations from the data. For instance, Tajbakhsh *et al.* [Taj+19] predicted medical scan's orientation as a proxy

task, Spitzer *et al.* [Spi+18] predicted 3D distances between two 2D patches in brain scans, Zhou *et al.* [Zho+19b] used image reconstruction tasks in 3D scans to learn data representations, Jiao *et al.* [Jia+20] utilized the order of clips in ultrasound videos as a free source of supervision signals, Zhuang *et al.* [Zhu+19] used 3D jigsaw puzzles solving as a proxy task, and many other works that utilize the image spatial context for learning representations [BNH19; Che+19]. In chapter 3 and chapter 4 we present our developed proxy tasks, which show improved results on multiple medical imaging benchmarks.

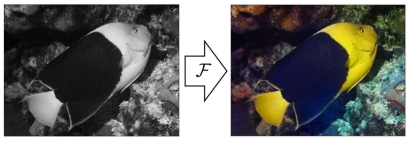
More recently, contrastive learning methods have also been applied to medical images [Cha+20; Hu+19; LAS20], where they also showed promising results on different medical imaging downstream tasks. As we show in chapter 4 and chapter 5, contrastive learning can improve the quality of learned data representations from 3D medical scans and also when integrating with other different modalities, respectively. In chapter 6 we demonstrate a way to take advantage of domain knowledge from medical scans to improve the learned data representations with contrastive methods.



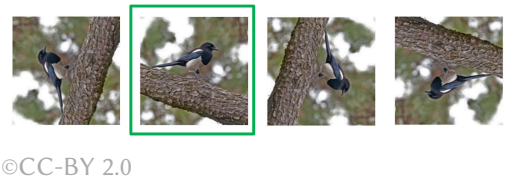
(a) Context prediction [DGE15] of a patch's position (e.g. Y=3) relative to the center.



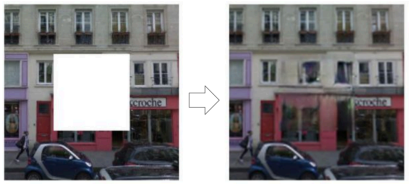
(b) Jigsaw puzzle [NF16] solving by predicting the applied permutation as a proxy task.



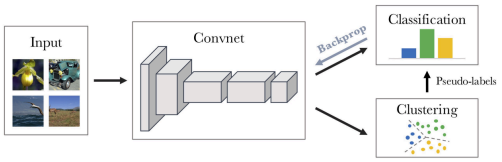
(c) Image colorization [ZIE16] by recovering the original colors of images.



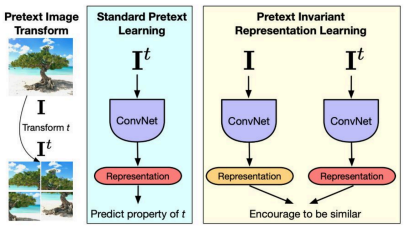
(d) Rotation prediction [GSK18] of artificially applied rotations on input images.



(e) In-painting [Pat+16] of masked pixels, to recover original appearance and texture.



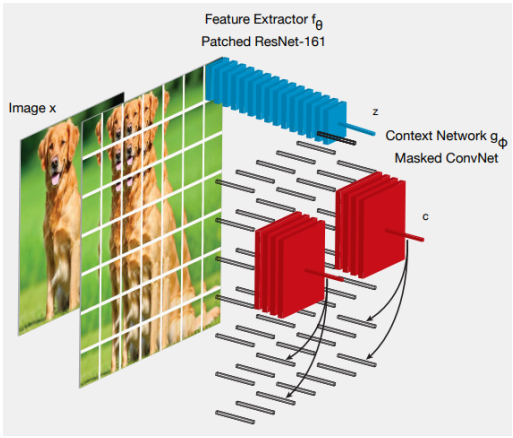
(f) K-Means clustering [Car+18] of image features uses the clusters as classification targets.

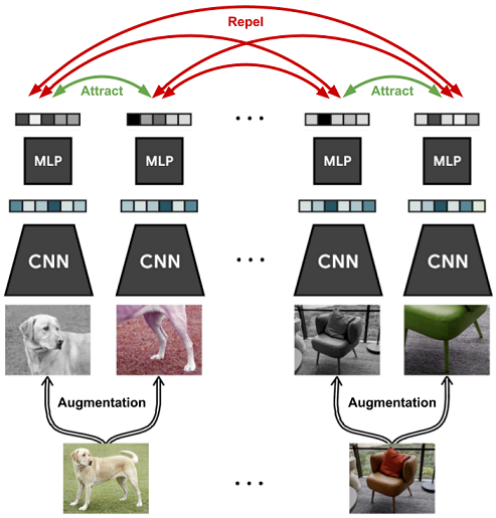


(g) Pretext-invariant learning [MM20] attempts to learn pretext-invariant representations.

(h) Exemplar networks [Dos+14] uses transformed images as classes.

Figure 2.11: Examples of self-supervised pretext (proxy) tasks for representation learning.





(a) Contrastive predictive coding [Hen20] uses an auto-regressive classifier for future or next "positive" representations among a set of unrelated "negative" samples.

(b) SimCLR [Che+20a] utilizes the InfoNCE loss that simplifies the problem into a binary classification task of positive versus negative samples. Figure Source: [Tsa]

Figure 2.12: Examples of contrastive learning approaches for representation learning.

3 Self-supervision from Multimodal Medical Images

This chapter extends own work in [Tal+21] and [Tal+19], presented at the International Conference on Information Processing in Medical Imaging (IPMI 2021) and the Workshops of Neural Information Processing Systems (NeurIPS 2019), respectively.

3.1 Introduction

Modern medical diagnostics rely on the analysis of multiple imaging modalities, such as in differential diagnosis [Lon+12]. As explained earlier in Sec. 2.1.3, medical images are multimodal, e.g. MRI and CT. As can be seen in Fig. 2.7, modalities of medical images have different characteristics in appearance and in what structures, tissues or organs they are able to capture. Multimodality in medical imaging is motivated for exactly that reason, from an anatomical perspective, the physical properties of different organs and tissues are expressed in a complementary fashion. For instance, soft body tissues are better rendered in MRI, whereas CT images capture bone structures more clearly. In addition, certain types of brain tumors or tissues are better seen in specific MRI modalities. Several other examples can be found in more specialized sources [EM11]. The multimodal nature of medical scans and the rich complementary information across these modalities motivate exploiting this property in learning data representations. Integrating the cross-modal complementary information early in the learned representations is necessary for solving subsequent downstream tasks accurately, e.g. semantic segmentation.

As mentioned in earlier chapters, learning representations with supervised methods requires significant amounts of data labels, hence motivating self-supervised learning as a viable alternative when labeled training data is scarce. Some self-supervised methods, e.g. [DGE15; NF16], utilize the spatial context as a supervisory signal to learn effective data representations. However, these methods derive the spatial context from uni-modal image inputs. On the other hand, we extract the spatial context from multiple medical imaging modalities. To that end, we propose to include multiple imaging modalities in the design of our Multimodal Jigsaw puzzle solving task, to integrate the cross-modal complementary information in

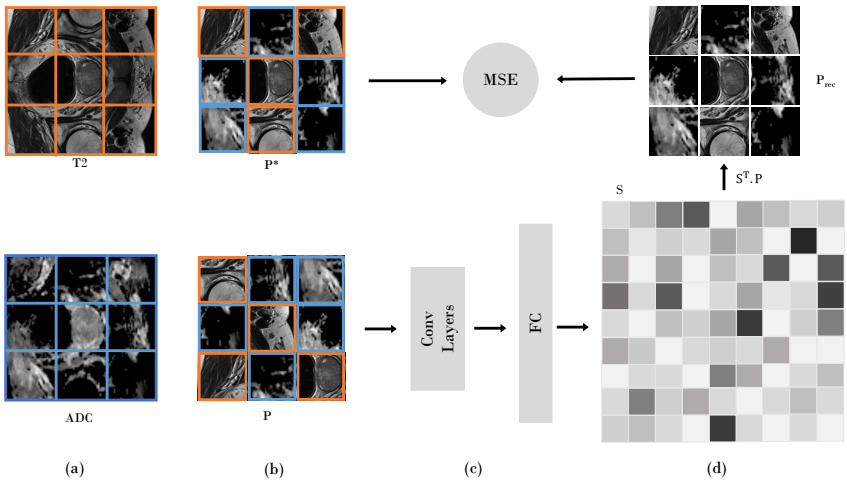


Figure 3.1: Overview of the process pipeline of the proposed method. Assuming two imaging modalities in this figure: T2 and ADC. (a) Multimodal puzzles are created by randomly selecting each patch from a different random modality, (b) yielding ground truth P^* and random puzzle P. (c) Shuffled puzzles P are processed by the model to train the puzzle-solver with the objective of recovering P^*, (d) by applying the learned permutation matrix S to reconstruct P_{rec}. The example scans are from the Prostate [Sim+19a] dataset. Figure source [Tal+21], reprinted with permission.

the learned representations. The proposed multimodal jigsaw puzzles are created by mixing patches from multiple imaging modalities, as depicted in Fig. 3.1.

As explained in Sec. 2.3.1, self-supervised methods with pretext tasks generally aim to recover the applied transformations to input data to learn visual representations. The main intuition behind solving Jigsaw puzzles as a proxy task, in particular, is that performing well on this task requires understanding scenes and objects, i.e. the parts that make an object and their relations to each other. Additionally, our proposed *multimodal* Jigsaw puzzles also encourage deriving modality-agnostic notions (or views) about the data, as the puzzle patches are randomly drawn from several imaging modalities and then mixed in the same puzzle. In other the words, the model confuses the multimodal knowledge it extracts from input imaging modalities in the learned representations. Our multimodal puzzle solving task can be said to define a modality-confusion loss.

As detailed before, modalities of medical scans are diverse, and their respective properties depend on the use-case and the acquisition process. However, in real-world clinical scenarios, the quantities of these modalities can vary, i.e. certain modalities are more abundant than others. This in turn creates a modality

imbalance problem. In addition, the modalities are often non-registered, meaning that the scanned regions that capture tissues or organs can have different view angles or scales. Even though our multimodal Jigsaw puzzles prove effective when trained and fine-tuned using *realistic* non-registered multimodal images, as shown in our experimental results in Sec. 3.4.2 and Sec. 3.4.4. We also propose to address the modality imbalance problem using a cross-modal generation step, with which we enhance the quantities of underrepresented modalities. In other words, the size of the multimodal part in the dataset is increased. This proposed cross-modal generation step is realized by employing a CycleGAN [Zhu+17] image-to-image translation model, which also deems the registration of imaging modalities unnecessary, since it performs this task inherently. By performing this cross-modal translation step prior to our puzzle-solving task, we demonstrate that *synthetic* images can be exploited for self-supervised learning. With the alternative being to directly training downstream task with synthetic data, which may suffer from quality issues. As the results in Sec. 3.4.3 show, the introduced cross-modal translation step alleviates the modality imbalance problem.

In summary, in this chapter, we present two main contributions. First, a novel self-supervised pretext task, namely a multimodal Jigsaw puzzle-solving algorithm that confuses multiple imaging modalities at the data-level. The motivation being that it allows for integrating complementary information across imaging modalities about the various concepts in the data. Second, we employ cross-modal image translation (generation) for self-supervised pretraining, instead of in training downstream tasks directly. This step is meant to address image modality imbalance phenomena, and to circumvent quality concerns associated with synthetic data, while at the same time retaining performance gains in difficult real-world scenarios.

3.2 Related Work

Jigsaw puzzle solving for self-supervised learning. Noroozi *et al.* [NF16] first proposed to solve Jigsaw puzzles as proxy task for learning data representations from natural 2D images, by extending the patch-based approach of [DGE15]. In contrast to our proposed multimodal puzzles, the puzzles created in [NF16]'s are uni-modal by design, thus disregarding the vital cross-modal complementary information from other image modalities. In addition, [NF16]'s method requires expensive memory and compute resources, which explains the limit of creating only small puzzles of 3×3 for which it uses 9 replicas of AlexNet [KSH12], one replica for each patch. On the other hand, our method improves the computational tractability by utilizing the efficient Sinkhorn operator [AZ11; Men+18; Sin64] as

an analog to the Softmax operator for permutation (ranking) tasks. As a result, we are able to solve Jigsaw puzzles with higher levels of complexity, e.g. up to 8×8 patches. The Sinkhorn operator allows for casting the puzzle solving problem as a permutation matrix inference instead of classification. The latter requires choosing a fixed permutation set as classification targets. In other words, choosing a fixed permutation set size limits the classification task complexity, such as in [NF16], and thus the complexity of the self-supervised task is capped. Conversely, defining our task as a permutation matrix inference enforces the model to find the applied permutation among *all* possible permutations.

Algorithms to solve Jigsaw puzzles. Puzzle solving algorithms in the field of artificial intelligence are numerous, and were proposed for different purposes and applications. Greedy algorithms [Gal12; PPT18; SHC14] use sequential pairwise patch matching to solve Jigsaw puzzles, deeming them computationally inefficient. Alternative approaches seek global solutions that observe all the puzzle patches at once, and optimize an objective measure over them. Example methods for global solutions include: Probabilistic methods with Markov Random Fields [CAF10], Genetic (or evolutionary) algorithms [SDN14], consensus agreement algorithms across neighbors [Son+16], and permutation classification with deep learning models [NF16]. The latter work [NF16], which we mentioned earlier, also allows for data representation learning, yet it can only capture subsets of possible puzzle permutations. Thus, algorithms that are able to capture the whole set of permutations [AZ11; Gro+19; Men+18; San+17] are advantageous. Generally, these solutions aim to approximate the applied permutation matrix, which is a doubly stochastic matrix of zeros and ones, and solve an optimization problem to recover the true matrix. As detailed later in the method section, each permutation of puzzle tiles corresponds to a unique permutation matrix, which is the ground-truth when training the models with these solutions, and is approximated by differentiable operators similar to the Sinkhorn [Sin64]. Analogous to the Softmax operator, these differentiable operators, e.g. the Sinkhorn, can be employed in the neural network output, allowing for efficient Jigsaw puzzle solving [AZ11].

Multimodal deep learning works [BAM19; Ngi+11] attempt to tackle many of the inherent challenges in learning from multimodal data, such as multimodal fusion, alignment, and representation. Multimodal representations may improve learning in many tasks, which are otherwise unfeasible using single-modal representations. Prior works that learned deep representations, including self-supervised methods, combined diverse sets of modalities, such as: image and text [Ant+15b; Ayt+18; Chu+17; Joh+16; Ree+16; Xu+15; Zha+19], image and audio [Alb+18; AVT16; AZ17; HTG16; Kar+17; OE18; Owe+18], audio and text [AGG18; YBJ18], and multiview (multimodal) images [Kum+20; PG16a; SBO18; SZ14]. The latter set of

works are most relevant to our proposed multimodal puzzles in terms of employed modality types. In comparison to these works, our approach fuses input imaging modalities at the *data-level* as opposed to the *feature-level* used in these works. In our approach, actually, we perform an early modality fusion [BAM19] when creating a multimodal task by fusing the data of multiple modalities and then solving that task. In the *feature-level* modality fusion, the model is expected to confuse (or combine) the modalities in the learned features, which corresponds to late modality fusion [BAM19]. The feature-level approach is likely to fail when the difference between the characteristics of the modalities is high. On the other hand, our approach is expected to bridge this difference as the integration of modalities occurs at the data-level, as our experimental results confirm.

Image-to-image translation (or conversion) with Generative Adversarial Networks (GANs) [Iso+17; Yi+17; Zhu+17] found multitude of use-cases in the medical imaging domain [Arm+19; JRP19; Wol+17; Yan+18; ZYZ18]. These works aim mainly to improve the quality of cross-modal translation, an objective we deem orthogonal to our goal. In fact, similar to [Fu+18; San+19; Tan+19], we utilize cross-modal translation methods as means of input augmentation to improve the downstream performance. However, especially in clinical applications, the quality of synthetic images may be questionable. Therefore, we circumvent this quality concern by employing synthetic (translated) images for pretraining purposes only (in training multimodal puzzle solver models), and not for the downstream prediction tasks. Furthermore, this fashion of exploiting generated images addresses situations where aligned multimodal samples are limited in quantities, yet their existence is vital to learn the cross-modal information. Here, we refer to the modality imbalance problem mentioned earlier. For example, certain modalities exist in abundance (e.g. X-Ray) in clinical settings, and others are less abundant (e.g. MRI), due to acquisition regimes.

3.3 Methods

Medical imaging modality types are numerous [EM11] and vary in their characteristics and use-cases. For forming our multimodal puzzles and training the models to solve them we assume no prior knowledge about which modalities are employed, i.e. they can vary from one task to another. Even though our method is able to operate on inputs from one modality only, in the following method formulation we assume inputs from two or more imaging modalities. Some passages in this section have been quoted verbatim from own work in [Tal+21], and are only insignificantly changed. Namely, we refer to the passages in Sec. 3.3.1, Sec. 3.3.2, and Sec. 3.3.3.

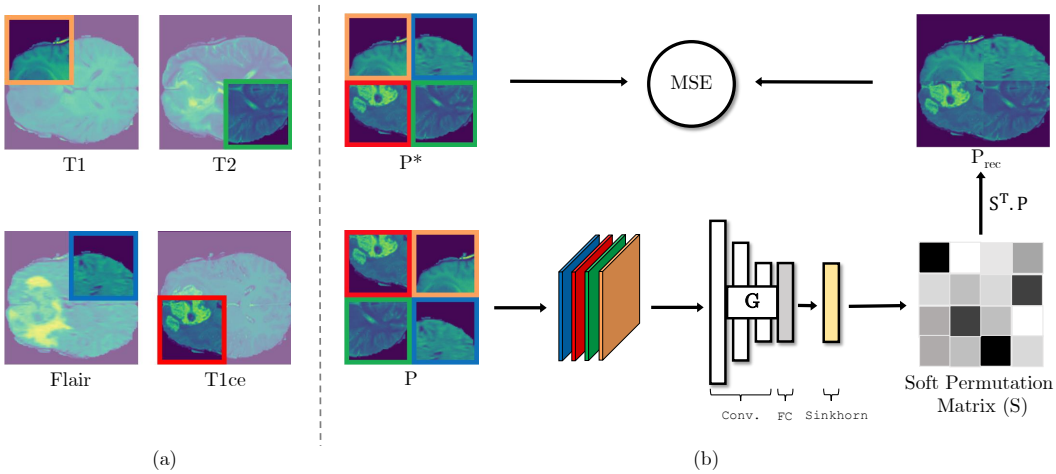


Figure 3.2: Schematic illustration showing the steps of the proposed multimodal puzzles with scans from the BraTS [Bak+17; Men+15] dataset. (a) Assuming four image modalities of Brain MRI in this example, (b) these images are then used to construct multimodal jigsaw puzzles, drawing patches from all the modalities randomly, here we show 2×2 puzzles for simplicity. The neural network model G extracts features from input patches, and the `Sinkhorn` layer outputs a soft permutation matrix S that approximates the true permutation matrix. Then, S is applied to the shuffled puzzle P to reconstruct the puzzle P_{rec}, which is minimized to $P*$ with an MSE loss function. Figure source [Tal+21], reprinted with permission.

3.3.1 Multimodal Puzzle Construction

Solving a jigsaw puzzle involves two main steps. First, the image is cut into puzzle pieces, which are shuffled randomly according to a certain permutation. Second, these shuffled image pieces are assembled such that the original image is restored. If C is the number of puzzle pieces, then there exist $C!$ of possible puzzle piece arrangements. It should be noted that when the puzzle complexity increases, the association of individual puzzle tiles can be ambiguous, e.g. puzzle tiles that originate from uni-colored backgrounds can be tricky to place correctly. Nevertheless, the placement of different puzzle tiles is mutually exclusive. Thus, when all tiles are observed at the same time, the positional ambiguities are alleviated. In a conventional jigsaw puzzle, the puzzle pieces originate from only one image at a time, i.e. the computational complexity for solving such a puzzle is $O(C!)$.

On the other hand, we propose a *multimodal* jigsaw puzzle, where tiles can be from M different modalities. This proposed multimodal puzzle simultaneously

learns the in-depth representation of how the organs compose, along with the spatial relationship across modalities. As a result, the complexity of solving multimodal puzzles is increased to $O(C!^M)$. Consequently, this quickly becomes prohibitively expensive due to two growth factors in the solution space: i) factorial growth in the number of permutations $C!$, ii) exponential growth in the number of modalities M. To reduce the computational burden, we use two tricks. First, we employ the Sinkhorn operator, which allows for an efficient solving of the factorial factor. Second, we employ a feed-forward network G that learns a cross-modal representation, which allows for canceling out the exponential factor M while simultaneously learning a rich representation for downstream tasks.

Algorithm 1: Multimodal jigsaw puzzle creation

```
Algorithm CREATE PUZZLES
    Input: - modality lists (m_1, m_2, ..., m_M), each with L slices
           - number of patches in a puzzle (np)
           - list of possible permutations (perms)
           - # of puzzles to generate per slice (nps)
    Output: list of multimodal puzzles
    for i ← 1 to L do
        for pt ← 1 to np do
            m ← choose random modality
            patches[pt] ← fill patch in position pt from slice with modality
              m
        for p ← 1 to nps do
            perm_patches ← shuffle patches using a random permutation
              from perms
            append perm_patches to puzzles
    return puzzles
```

3.3.2 Puzzle-Solving with Sinkhorn Networks

To efficiently solve the self-supervised jigsaw puzzle task, we train a network that can learn a permutation. A permutation matrix of size $N \times N$ corresponds to some permutation of the numbers 1 to N. Every row and column, therefore, contains precisely a single 1 with 0s everywhere else, and every permutation corresponds to a unique permutation matrix. This permutation matrix is non-differentiable.

However, as shown in [Men+18], the non-differentiable parameterization of a permutation can be approximated in terms of a differentiable relaxation, the so-called Sinkhorn operator. The Sinkhorn operator iteratively normalizes rows and columns of any real-valued matrix to obtain a "soft" permutation matrix, which is doubly stochastic. Formally, for an arbitrary input X, which is an N dimensional square matrix, the Sinkhorn operator $S(X)$ is defined as:

$$\begin{aligned} S^0(X) &= exp(X), \\ S^i(X) &= \mathcal{T}_R(\mathcal{T}_C(S^{i-1}(X))), \\ S(X) &= \lim_{i \to \infty} S^i(X). \end{aligned} \tag{3.1}$$

where $\mathcal{T}_R(X) = X \oslash (X\mathbf{1}_N\mathbf{1}_N^\top)$ and $\mathcal{T}_C(X) = X \oslash (\mathbf{1}_N\mathbf{1}_N^\top X)$ are the row and column normalization operators, respectively. The element-wise division is denoted by $\oslash$, and $\mathbf{1}_N^\top \in \mathbb{N}^N$ is an N dimensional vector of ones.

Assuming an input set of patches $P = \{p_1, p_2, ..., p_N\}$, where $P \in \mathbb{R}^{N \times l \times l}$ represents a puzzle that consists of N square patches, and l is the patch length. We pass each element in P through a network G, which processes every patch independently and produces a single output feature vector with length N. By concatenating together these feature vectors obtained for all region sets, we obtain an $N \times N$ matrix, which is then passed to the Sinkhorn operator to obtain the soft permutation matrix S. Formally, the network G learns the mapping $G : P \longrightarrow S$, where $S \in [0, 1]^{N \times N}$ is the soft permutation matrix, which is applied to the scrambled input P to reconstruct the image P_{rec}. The network G is then trained by minimizing the mean squared error (MSE) between the sorted ground-truth P^* and the reconstructed version P_{rec} of the scrambled input, as in the loss formula below:

$$\mathcal{L}_{puzzle}(\theta, P, P^*) = \sum_{i=1}^{K} \left\| P_i^* - S_{\theta,P_i}^T . P_i \right\|^2, \tag{3.2}$$

where θ corresponds to the network parameters, and K is the total number of training puzzles. After obtaining the network parameters θ, the yielded representations capture different tissue structures across the given modalities as a consequence of the multimodal puzzle solving. Therefore, the learned representations can be employed in downstream tasks by a simple fine-tuning on target domains, in an annotation-efficient regime. Our proposed approach is depicted in Fig. 3.2.

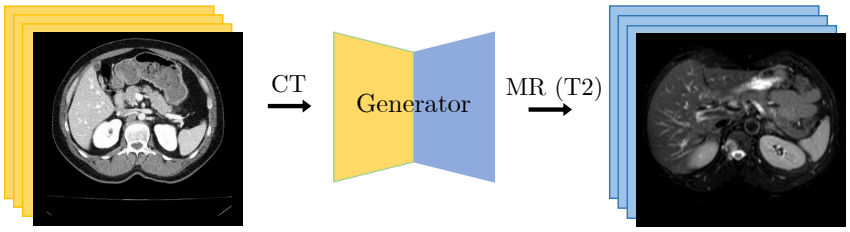


Figure 3.3: A cross-modal generation example on abdominal scans from CT to MR (T2) modalities. Here, G is a generator model between these modalities. The scans are from the CHAOS [Kav+19] dataset. Figure source [Tal+21], reprinted with permission.

3.3.3 Cross-Modal Generation

Multimodal medical images exist in several curated datasets, and in pairs of aligned (or registered) scans. However, as described before, in many real-world scenarios, obtaining such data in large quantities can be challenging. To address this, we add an explicit cross-modal generation step using CycleGAN [Zhu+17], illustrated in Fig. 3.3. This model also uses a cycle-consistency loss, which allows for relaxing the alignment (pairing) constraint across the two modalities. Therefore, CycleGAN can translate between any two imaging modalities, requiring no prior expensive registration steps. This step allows us to leverage the richness of multimodal representations obtained by our proposed puzzle-solving task. In our scenario, after generating data samples of the small (in number of samples) modality using samples from the larger modality, we construct our multimodal puzzles using a mix of real and generated multimodal data. As we show in our experiments, this yields better representations compared to using a single modality only when creating puzzles. We have to highlight that, our multimodal puzzles are capable of operating on realistic multimodal images merely, which are the results shown in Sec. 3.4.2. However, we assess the influence of mixing those realistic multimodal images with synthetic ones in Sec. 3.4.3.

3.4 Experimental Results

In the following sections, we assess the quality of representations learned with our proposed pretraining method on multimodal medical imaging datasets detailed in Sec. 3.4.1. We transfer (and fine-tune) the learned representations to different downstream tasks, and measure their impact in Sec. 3.4.2. Then, we study the effect of integrating generated synthetic data in constructing our multimodal puzzles in

Sec. 3.4.3. Next, we assess how our self-supervised task affects the downstream tasks' data efficiency, i.e. when operating in a low-data regime, in Sec. 3.4.4. Finally, we analyze the effect of many variables (puzzle complexity and permutation list size) in our multimodal puzzles on downstream task performance in a set of ablation studies in Sec. 3.4.5.

3.4.1 Datasets

In our experiments, we consider three multimodal medical imaging datasets. The first is the Multimodal Brain Tumor Image Segmentation Benchmark (**BraTS**) dataset [Bak+17; Men+15]. This dataset is widely used to benchmark different semantic segmentation algorithms in the medical imaging domain. It contains multimodal MRI scans for 285 training cases and for 66 validation cases. All BraTS scans include four MRI modalities per case: a) native (T1), b) post-contrast T1-weighted (T1Gd), c) T2-weighted (T2), and d) T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes. The BraTS challenge involves two different tasks: i) brain tumor segmentation, and ii) number of survival days prediction.

The second benchmark is the **Prostate** segmentation task from the Medical Segmentation Decathlon [Sim+19a]. The prostate dataset consists of 48 multimodal MRI cases, from which 32 cases are for training, and 16 are for testing. Ground-truth segmentations of the whole prostate were produced from T2-weighted scans, and the apparent diffusion coefficient (ADC) maps. The downstream task is segmenting two adjoint prostate regions (the central gland and the peripheral zone).

The third benchmark is the **Liver** segmentation task from the CHAOS [Kav+19] multimodal dataset. The CHAOS dataset contains 40 multimodal cases, from which 20 cases are used for training, and 20 for testing. This dataset consists of CT and MRI multimodal data, where each case (patient) has one CT and one MRI scans. The CT and MR modalities in this benchmark are not only different in appearance, but also they are non-registered, making this dataset a pertinent test-bed for our multimodal puzzles.

3.4.2 Transfer Learning Results

We evaluate the quality of the learned representations from our task of multimodal puzzle-solving by transferring them in downstream tasks. Then, we assess their impact on downstream performance. As mentioned before, we only use *realistic* data in the experiments presented in this section.

Brain Tumor Segmentation

The goal of this task is to segment 3 different regions of brain tumor: a) the whole tumor (WT), b) the tumor core (TC), and c) the enhanced tumor (ET). Each of these regions has different characteristics, and each may appear more clearly on specific MRI modalities than others, justifying the need for multiple modalities.

Baselines In order to assess the quality of our representations, we establish the following baselines. Apart from the **Single-modal** baseline, all of the following baselines use multimodal data. To average out random variations and compute p-values, we used a 5-fold cross validation evaluation approach. For each fold, we used the same training and test datasets for our method and all baselines to perform a one-sided two-sample t-test on the dice scores, following [Kum+20].

From Scratch: The first sensible baseline for all self-supervised methods is to compare with the model when trained on the downstream task from scratch. This baseline provides an insight into the benefits of self-supervised pretraining (initialization), as opposed to learning the target task directly.

Single-modal: We study the impact of our pretraining method on this task when processing only a single modality as input. This experiment aims at simulating the realistic situation when human experts examine brain scans, as some modalities highlight certain aspects of the tumor more than others. For instance, Flair is typically used to examine the whole tumor area, while T2 is used for tumor core, and the T1ce highlights the enhanced tumor region. We select the best modality for each task when comparing to these results.

Isensee *et al.* [Ise+18]: This work ranks among the tops in the BraTS 2018 challenge. It uses additional datasets next to the challenge data, and it performs multiple types of augmentation techniques. The model architecture is a 3D U-Net [RFB15]. We only fine-tune our learned representations from the self-supervised task, thus requiring much less data and augmentation methods. **2D Isensee** is a 2D version of their network, which we implement for better comparability.

Chang *et al.* [Cha+18b]: Trained multiple versions of the 2D U-Net models, and used them as an ensemble to predict segmentation masks. This requires significantly more computing time and resources than training a single model that performs the task with higher performance in many cases.

Li [Li18]: Implemented a 3-stage cascaded segmentation network that combines whole-tumor, tumor-core and enhanced-tumor masks. For the whole-tumor stage, they utilize a modified multi-view 2D U-Net architecture, which processes three slices at a time from input 3D scans: axial, sagittal, and coronal views. Our method produces better results while requiring less computations using a smaller network.

JiGen [Car+19]: This method is a multi-tasking approach called JiGen [Car+19]. JiGen solves jigsaw puzzles as a secondary task for domain generalization. We implemented their model and treated the multiple modalities as if they were other domains. This baseline aims to analyze the benefits of performing modality confusion on the data-level (our approach), as opposed to the feature-level (JiGen).

Models Genesis [Zho+19b]: This is a self-supervised method that relies on multiple image reconstruction tasks to learn from unlabeled scans. Even though their method is mainly implemented in 3D, we employ the 2D version (**2D MG**) of their model[1] (pretrained on Chest-CT), for better comparability.

Rubik Cube [Zhu+19]: A self-supervised method that relies on solving 3D jigsaw puzzles as a proxy task, and also applies random rotations on puzzle cubes. Similarly, we compare to the 2D version (**2D RC**) for better comparability.

Evaluation Metrics The reported metrics are the average dice scores for the Whole Tumor (WT), the Tumor Core (TC), and the Enhanced Tumor (ET).

Discussion The results of our **multi-modal** method compared to the above baselines are shown in Tab. 3.1. Our proposed method outperforms both the "from scratch" and "single-modal" baselines, confirming the benefits of pretraining using our multimodal approach. In addition, our method achieves comparable results to methods from literature. We outperform these baselines in most cases, such as the methods Chang *et al.* [Cha+18b], and Li [Li18], in terms of all reported dice scores. We also report the result of a 2D version (for better comparability) of Isensee *et al.* [Ise+18], which ranks among the best results on the BraTS 2018 benchmark. Even though their method uses co-training with additional datasets and several augmentation techniques, we outperform their results in this task. This supports the performance benefits of initializing CNNs with our multimodal puzzles. We also compare with 2D Models Genesis (2D MG) [Zho+19b], which we outperform in this downstream task, supporting the effectiveness of our pretraining method. Compared to 2D Rubik Cube [Zhu+19], which we implement in 2D for comparability, we observe that we outperform this method in this task. This confirms the higher quality of the representations learned by our method, compared to this jigsaw puzzle solving method (including random rotations). Compared to the work of [Car+19] (JiGen), we also find that our results outperform this baseline, confirming that our approach of performing the modality confusion in the data-level is superior to modality confusion in the feature-level.

1 It uses Resnet18 as the encoder of the U-Net architecture implemented here: https://github.com/qubvel/segmentation_models

Table 3.1: Results on the BraTS segmentation task

Model	ET	WT	TC	p-value
From scratch	68.12	80.54	77.29	3.9e-5
Single-modal	78.26	87.01	82.52	6.0e-4
Li [Li18]	73.65	88.24	78.56	9.0e-4
Chang *et al.* [Cha+18b]	75.90	89.05	82.24	2.6e-3
2D MG [Zho+19b]	79.21	88.82	83.60	7.6e-2
2D Isensee *et al.* [Ise+18]	78.92	88.42	83.73	4.6e-2
2D RC [Zhu+19]	78.38	87.16	82.92	1.4e-2
JiGen [Car+19]	78.75	88.15	83.32	5.0e-4
Ours (Multi-modal)	**79.65**	**89.74**	**84.48**	-

Prostate Segmentation

The target of this task is to segment two regions of the prostate: central gland, and peripheral zone. This task utilizes two available MRI modalities.

Baselines We establish the following baselines, which, apart from **Single-modal**, all use multimodal data. To average out random variations and compute p-values, we used a 5-fold cross validation evaluation approach. For each fold, we used the same training and test datasets for our method and all baselines to perform a one-sided two-sample t-test on the dice scores, following [Kum+20].

From Scratch: Similar to the first downstream task, we compare to the same model architecture when training on the prostate segmentation task from scratch.

Single-modal: We also study the impact of our method when using only a single modality (T2) to create the puzzles.

JiGen [Car+19]: Similar to the first downstream task, we compare our method to the multi-tasking approach JiGen.

2D Models Genesis [Zho+19b] (2D MG): Similar to the first task, we fine-tune this model on multimodal Prostate data.

2D Rubik Cube [Zhu+19] (2D RC): Similar to the first task, we fine-tune the 2D version of this method on multimodal prostate data.

Evaluation Metrics We report the values of two evaluation metrics in this task, the average dice score (Dice) and the normalized surface distance (NSD). These metrics are used on the official challenge. The metrics are computed for the two prostate regions (**C**entral and **P**eripheral).

Table 3.2: Results on the Prostate segmentation task

Model	Dice		NSD		p-value
	C	P	C	P	
From scratch	68.98	86.15	94.57	97.84	4.9e-3
Single-modal	69.48	87.42	92.97	97.21	9.3e-5
2D MG [Zho+19b]	73.85	87.77	94.61	98.59	4.3e-2
2D RC [Zhu+19]	73.11	86.14	93.65	97.47	3.1e-2
JiGen [Car+19]	69.98	86.82	92.67	96.13	9.1e-3
Ours (Multi-modal)	**75.11**	**88.79**	**94.95**	**98.65**	-

Discussion The results of our **multi-modal** method compared to the above baselines are shown in Tab. 3.2. Our proposed method outperforms both "from scratch" and "single-modal" baselines in this task, too, supporting the advantages of pretraining the segmentation model using our multimodal approach. We also compare with 2D Models Genesis (2D MG) [Zho+19b], which we outperform in this downstream task, supporting the effectiveness of our pretraining method. Also, our method outperforms the multitasking method JiGen [Car+19], when trained on this task too. We observe a more significant gap in performance between our approach and JiGen in this task, compared to the first downstream task of brain tumor segmentation. We posit that this can be attributed to the more significant difference between the imaging modalities used in this prostate segmentation task, as opposed to those in the brain tumor segmentation task. The Fig. 3.4 shows this difference more clearly. It can be noted that the imaging modalities of the prostate dataset, are more different in appearance than those of the brain tumor dataset. This difference in appearance among the modalities can be explained by understanding the physics from which these MRI modalities are created. All of the brain MRI sequences in the BraTS dataset are variants of T1- and T2-weighted scans, they only differ in configurations of the MRI scanner. These different configurations cause the contrast and brightness of some brain areas to vary among these MRI sequences. The ADC map, on the other hand, is a measure of the magnitude of diffusion (of water molecules) within the organ tissue. This requires a specific type of MRI imaging called Diffusion Weighted Imaging (DWI). In general, highly cellular tissues or cellular swellings exhibit lower diffusion coefficients, e.g. a tumor, a stroke, or in our case, the prostate. Compared to 2D Rubik Cube [Zhu+19], similarly, we outperform this method on this downstream task.

Liver Segmentation

The target of this task is to segment the liver from multimodal abdominal scans, which include CT and MRI modalities.

Baselines Apart from the **Single-modal** baseline, all of the following baselines use multimodal data. To average out random variations and compute p-values, we used a 5-fold cross validation evaluation approach. For each fold, we used the same training and test datasets for our method and all baselines to perform a one-sided two-sample t-test on the dice scores, following [Kum+20].

From Scratch: Similar to the first downstream task, we compare our model with the same architecture when training on liver segmentation from scratch.

Single-modal: We also study the impact of our pretraining method when using only a single modality to create the puzzles. Here, CT, as it is more abundant.

JiGen [Car+19]: Similar to the first downstream task, we compare our method to the multi-tasking approach JiGen.

Registered: Because the CT and MR modalities are not registered in this benchmark, we register them when training this baseline. This aims to assess the influence of registration on learned representations by our multimodal puzzles. We employ VoxelMorph [Bal+19][2] for multimodal image registration.

2D Models Genesis [Zho+19b] (2D MG): Similar to the first task, we fine-tune this model on multimodal Liver data.

2D Rubik Cube [Zhu+19] (2D RC): Similar to the first task, we fine-tune this method on multimodal liver data.

Evaluation Metrics We report the results of liver segmentation using the average dice score (Dice). This metric is used on the official challenge.

Discussion The results of our **multimodal** method compared to the above baselines are shown in Tab. 3.3. Our method outperforms both "from scratch" and "single-modal" baselines in this task too, supporting the advantages of initializing the model using our multimodal puzzle solving task. We also compare with 2D Models Genesis (2D MG) [Zho+19b], which we outperform in this downstream task, supporting the effectiveness of our pretraining method. However, we observe that our method only marginally outperforms this method, and we believe this is because Models Genesis was pretrained on Chest CT data. Also, our method

2 Our aim is to benchmark our method against a proven image registration method (VoxelMorph takes structural information into consideration)

Table 3.3: Results on the Liver segmentation task

Model	Avg. Dice	p-value
From scratch	89.98	1.2e-5
Registered	95.09	9.6e-1
Single-modal	92.01	6.3e-3
JiGen [Car+19]	93.18	2.1e-2
2D MG [Zho+19b]	95.01	3.8e-1
2D RC [Zhu+19]	94.15	4.9e-2
Ours (Multi-modal)	**95.10**	-

outperforms the multitasking method JiGen [Car+19], when trained on this task too. The results against the "Registered" baseline are almost on par with the results of our "multimodal" method using non-registered data. This result is significant because it highlights our multimodal puzzles' ability to operate on non-registered imaging modalities. Compared to 2D Rubik Cube [Zhu+19], we outperform this method on this downstream task too. This confirms that our method is able to learn better multimodal representations, given the same input modalities.

Survival Days Prediction (Regression)

The BraTS challenge involves a second downstream task, which is the prediction of survival days. The number of training samples is 60 cases, and the validation set contains 28 cases. Similar to what we did for the other downstream tasks, we transfer the learned weights of our multimodal puzzle solver model. The downstream task performed here is regression, hence the output of our trained model here is a single scalar that represents the expected days of survival. We reuse the convolutional features, and we add a fully connected layer with five features in it, and then a single output layer on top. We also include the age as a feature for each subject right before the output layer. The size of the fully connected layer, was determined based on the obtained performance, i.e. by hyperparameter tuning.

In Tab. 3.4, we compare our results to the baselines of Suter *et al.* [Sut+18]. In their work, they compared deep learning-based methods performances with multiple other classical machine learning methods on the task of survival prediction. The first experiment they report is **(CNN + age)**, which uses a 3D CNN. The second is a **random forest regressor**, the third is a multi-layer perceptron (MLP) network that uses a set of hand-crafted features called **FeatNet**, and finally, a **linear regression** model with a set of 16 engineered features. We outperform their results in all cases

Table 3.4: BraTS survival prediction (regression). The baselines (except for "From scratch") are taken from [Sut+18]

Model	MSE
From scratch	112.841
CNN + age	137.912
Random Forest Regression	152.130
FeatNet + all features	103.878
Lin. Reg. + top 16 features	99.370
Ours (Multi-modal)	**97.291**

when fine-tuning our puzzle solver model on this task. The reported evaluation metric is the Mean Squared Error (MSE).

3.4.3 Cross-Modal Generation Results

We investigate in this set of experiments the effect of the cross-modal generation step. As mentioned earlier, obtaining large multimodal medical imaging datasets can be sometimes challenging. Therefore, we investigate in this set of experiments, the effect of the explicit cross-modal generation step. This step allows for better adoption of our multimodal puzzle-solving, even in the case of a few multimodal samples only. It is common that some imaging modalities exist in larger quantities than others. Hence, in this set of experiments, we perform this step in a semi-supervised fashion, assuming small multimodal and large single-modal data subsets. The size of the multimodal subset influences the downstream task performance, and the quality of generated data. We evaluate the generation process at data subset sizes of 1%, 10%, and 50% of the total number of patients in each benchmark. We assume a reference modality, which is often abundant in practice, to generate the other modalities[3]. In the BraTS and Prostate benchmarks, we use T2-weighted MRI. In the Prostate dataset, we use T2-weighted MRI scans to generate the ADC diffusion-weighted scans. In BraTS, since we have four MRI modalities, we train three GANs and convert T2 MRI to the other MRI modalities (T1, T1CE, FLAIR). In the CHAOS liver benchmark, we use the CT modality to generate T2 MRI.

3 When there is no clear reference modality, it is also possible to generate all modalities from each other, which results in an increased number of trained GAN models.

Table 3.5: Results on segmentation in avg. dice scores. The percentages are sizes of multimodal subsets used to train CycleGAN

Model	BraTS			Prostate		CHAOS
	ET	WT	TC	C	P	Liver
Single-modal	72.12	82.72	79.61	69.48	87.42	92.01
Downstream training (1%)	65.40	74.11	69.24	55.24	71.23	80.31
Downstream training (10%)	69.28	78.72	71.58	62.65	76.18	83.65
Downstream training (50%)	72.92	81.20	78.36	66.34	80.24	87.58
Our method (1%)	73.12	82.42	80.01	61.87	82.67	82.71
Our method (10%)	74.19	85.71	81.33	67.67	84.37	86.26
Our method (50%)	**76.23**	**87.04**	**82.38**	**73.45**	**87.92**	**93.85**

Discussion. This step is only justified if it provides a performance boost over the **single-modal** puzzle solving baseline, i.e. training our model on puzzles that originate from one modality. We measure the performance on the three downstream tasks, by fine-tuning these models and then evaluating them on segmentation. The presented results in Tab. 3.5 clearly show an improvement on all benchmarks, when training our puzzle solver on the mixture of synthetic and realistic multimodal data. Even when we use only 1% of the total dataset sizes, the generator appears to capture the important characteristics of the generated modality. The qualitative results in Fig. 3.4, confirm the quality of generated images. In addition, we study the benefits of using the synthetic data for self-supervised pretraining, instead of training the downstream task directly on them. The results of **Our method** in Tab. 3.5 support this approach, as opposed to direct **Downstream training**.

3.4.4 Low-Shot Learning Results

In this set of experiments, we assess how our self-supervised task benefits the data-efficiency of the trained models, by measuring the performance on both downstream segmentation tasks at different labeling rates by fine-tuning our pre-trained model with corresponding sample sizes. We randomly select subsets of patients at 1%, 10%, 50%, and 100% of the total segmentation training set size. Then, we fine-tune our model on these subsets for a fixed number of epochs (50 epochs each). Finally, for each subset, we compare the performance of our fine-tuned **multimodal** model to the baselines trained **from scratch** and **single-modal**. As shown in Fig. 3.5, our method outperforms both baselines with a significant margin when using few

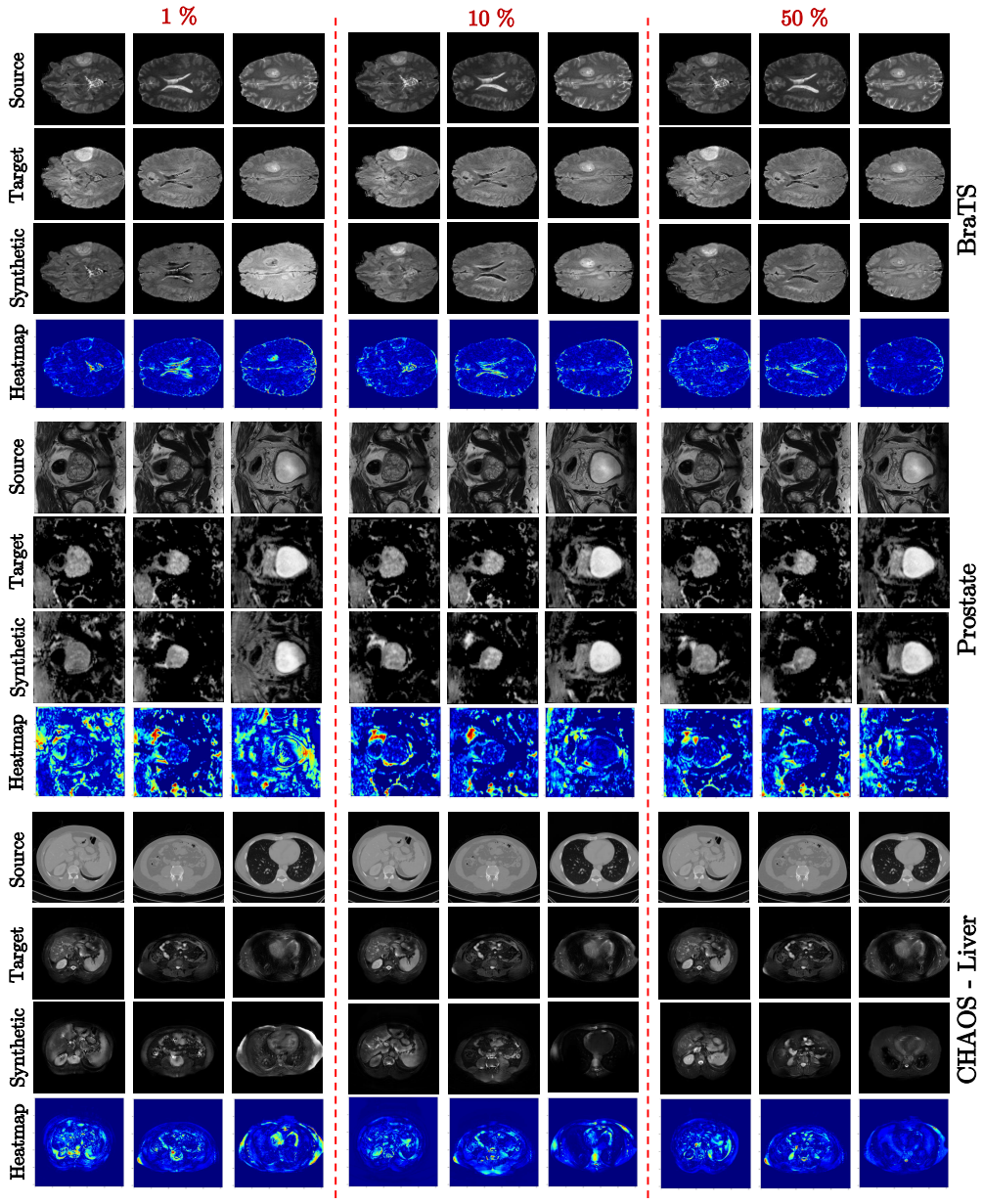


Figure 3.4: Qualitative results of the trained CycleGAN at different multimodal subset sizes (at 1%, 10%, and 50% of the full dataset). For BraTS, we convert scans from T2 to FLAIR, T2 to ADC for Prostate, and CT to MR-T2 for CHAOS liver. Despite using small multimodal subsets, the quality of synthetic images is high. The target images of the CHAOS liver dataset are obtained by registration. Source [Tal+21], reprinted with permission.

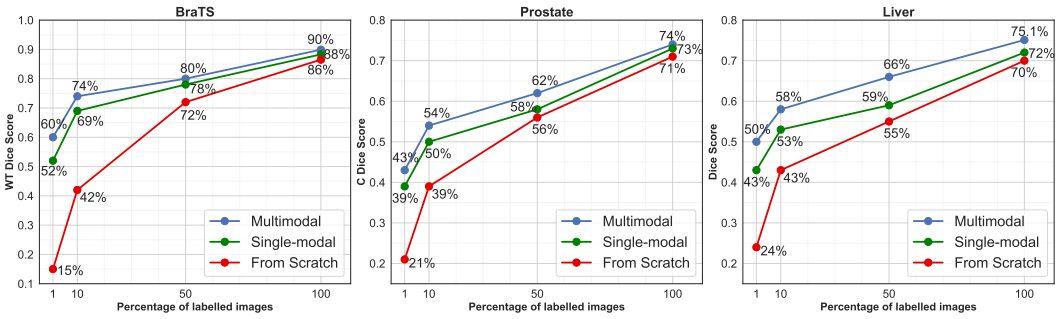


Figure 3.5: Results in the low-shot scenario. **Our method** outperforms both **single-modal** and **from scratch** baselines, confirming the benefits in data-efficiency when pretraining using our multimodal puzzles. Figure source [Tal+21], reprinted with permission.

training samples. This gap to the **single-modal** baseline confirms the benefits of using our multimodal puzzles instead of traditional single-modal puzzles. In a low-data regime of as few samples as 1% of the overall dataset size, the margin to the **from scratch** baseline appears larger. This case, in particular, suggests the potential for generic unsupervised features applicable to relevant medical imaging tasks. Such results have consequences on annotation efficiency, i.e. only a fraction of data annotations is required. It is worth noting that we report these low-shot results on *realistic* multimodal data. The **single-modal** baseline uses these modalities for each task: FLAIR for BraTS, T2 for Prostate, and CT for Liver.

3.4.5 Ablation Study

We use realistic data only in the experiments presented in this section.

Puzzle Complexity

In this set of experiments, we analyze the impact of the complexity of our multimodal jigsaw puzzles in the pretraining stage, on the performance of downstream tasks. This aims to evaluate whether the added complexity in this task can result in more informative data representations; as the model is enforced to work harder to solve the more complex tasks. Our results confirm this intuition, as shown in Fig. 3.6 (Left), where, in general, the downstream task performance (measured in Dice Score) increases as the puzzle complexity rises. This is true up to a certain point, at which we observe that the downstream performance almost flattens. This complexity turning point is different across the three downstream tasks, as can be

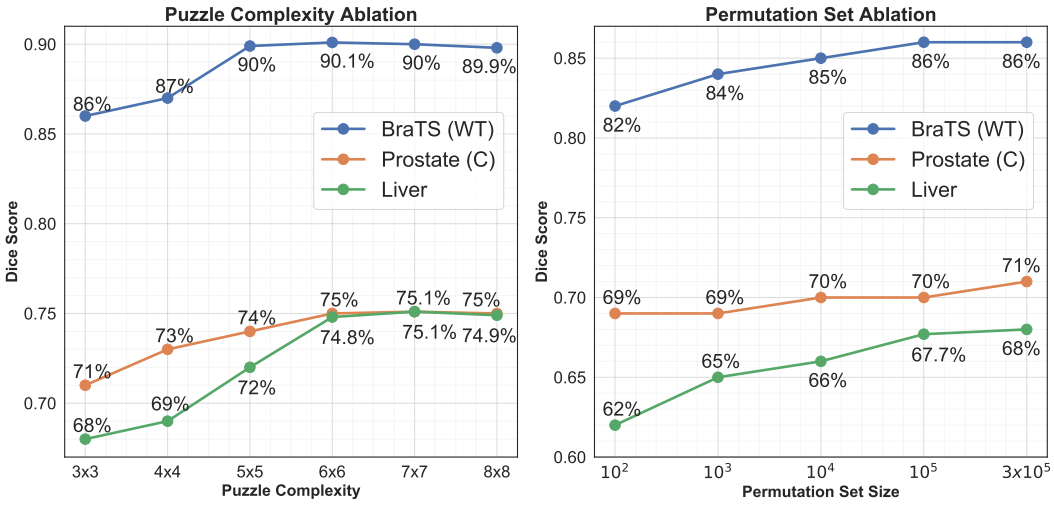


Figure 3.6: (Left) Puzzle complexity vs. downstream performance. The trendlines suggest more complex jigsaw puzzles improve downstream performance, up to a certain point where added complexity has little effect. (Right) Permutation set size vs. downstream performance. The plot shows that a larger set than a certain size has little influence on downstream performance, but a small set affects the results negatively. Figure source [Tal+21], reprinted with permission.

seen in Fig. 3.6 (Left). We should note that in all previous experiments, we use 5×5 puzzle complexity for BraTS tasks, and 7×7 for Prostate and Liver.

Permutation Set Size

In this set of experiments, we analyze the impact of the permutation list size used in creating multimodal jigsaw puzzles, on the performance of downstream tasks. This hyperparameter can be viewed as another form of puzzle complexity [NF16]. However, as mentioned before, our design choice to utilize the efficient Sinkhorn operator for puzzle solving captures the *whole* set of possible permutations. The main reason lies in solving the task as permutation matrix inference (regression), instead of permutation classification used in other solutions [Car+19; NF16; Zhu+19]. Consequently, our puzzle solver is expected to reduce the influence of the permutation list size on trained models. Nevertheless, in practice, we sample from a finite permutation set, for feasibility reasons, and we attempt to employ a permutation set as large as possible. Our intuition here is that a larger permutation set can facilitate learning a better semantic representation. Hence, in this set of experiments, we

analyze the effect of the chosen permutation set size. Fig. 3.6 (Right) shows the downstream performance for every permutation set size. It is worth noting that in this plot, we use a 3-by-3 puzzle complexity, which results with 9! total permutations. We vary the permutation set size from 10^2 to 3×10^5, which is almost equal to 9!. The trendlines in the plot show that a small permutation set, e.g. 100 or 1000, can harm the learned representations, as the model may overfit this small set of permutations. It also shows that an increase in permutation set size has a positive effect on downstream performance, similar to puzzle complexity. Nevertheless, the influence of permutation set size becomes negligible after a certain limit, as appears in the Fig. 3.6. This is consistent with the effect of varying puzzle complexity. We should highlight that the actual number of permutations grows with the number of modalities in each downstream task, as explained in Sec. 3.3.1. However, since our method learns cross-modal representations, it allows for cancelling out this exponential growth with more modalities.

Across both ablations presented in this section (for puzzle complexity and permutation set size), we observe a consistent behavior, even though the actual numbers may differ. This behavior is that both hyperparameters improve downstream performance as they increase, up to a certain limit, where their variation causes a negligible change in performance then. We believe that the model capacity here is the cause of such saturation, which is consistent with the findings of Goyal *et al.* [Goy+19]. Employing a larger model architecture may benefit more from increased complexity in our multimodal puzzles.

3.5 Discussion

In this chapter, we proposed a *multimodal* self-supervised Jigsaw puzzle-solving task. This approach allows for learning rich semantic representations that facilitate downstream task solving in the medical imaging context. In this regard, we showed competitive results to the state-of-the-art results in three medical imaging benchmarks. One of which has unregistered modalities, further supporting the effectiveness of our approach in producing rich data representations. The proposed multimodal puzzles outperform their single-modal counterparts, confirming the advantages of including multiple modalities when constructing jigsaw puzzles. It is also noteworthy that the efficient Sinkhorn operator enabled large permutation sets and puzzle complexities, an aspect commonly used puzzle solvers do not offer. In addition, our approach further reduces the cost of manual annotation required for downstream tasks, and our results in the low-data regime support this benefit. We also evaluated a cross-modal translation method as part of our framework, which

when used in conjunction with our method, it showed performance gains even when using few multimodal samples to train the generative model. We provide additional training details in Appendix A.

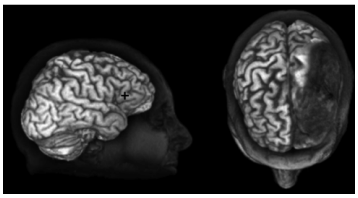
4 Self-supervision from 3D Medical Scans

This chapter extends own work in [Tal+20], published in the Proceedings of the Neural Information Processing Systems Conference (NeurIPS 2020). It also includes results from a supervised master's thesis, whose results are summarized in the following pre-print [Ali+21].

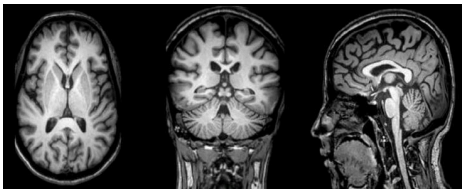
4.1 Introduction

Advancements in 3D sensing technologies have improved their adoption in multiple application fields, such as in Autonomous Driving, Robotic Navigation, CAD imaging, Virtual/Augmented Reality, and Medical Imaging. This wide spread of 3D technologies have prompted the development of Computer Vision algorithms that can extract meaningful information from 3D data. As a result, many works have relied on Deep Learning methods to extract rich semantic representations from 3D inputs [GB19; Ioa+17; Su+17], which have been successfully deployed in self-driving cars [Li+20]. In this chapter, we develop deep self-supervised learning algorithms for 3D input data, for which we use Medical Imaging as a test-bed. Nevertheless, we ensure their design allows generalization to other 3D domains.

As described in earlier chapters, medical imaging plays a vital role in patient healthcare, yet efforts to utilize advancements in Deep Learning are often hampered by the sheer expense of the expert annotation required. This becomes more apparent with 3D medical scans, where expert annotation at scale is even more expensive



(a) Example scan of 3D Brain MRI. Source: [Bar+11]



(b) View axes (Axial, Coronal, and Sagittal, from left to right). Source: [Pad+20]

Figure 4.1: Showing how a 3D scan **(a)** is projected on three view axes **(b)**.

and time-consuming. Hence, learning from unlabeled 3D medical scans with self-supervised learning methods is expected to alleviate the manual annotation burden. Unlabelled medical scans carry valuable information about organ and tissue structures, and self-supervised methods enable the models to derive notions about these structures.

Despite the surge of interest in self-supervised learning algorithms for representation learning, only little attention has been paid to 3D nature of many medical imaging modalities [EM11]. Typically, 3D imaging tasks are cast in 2D by extracting slices along an arbitrary axis, e.g. the axial, which is a sub-optimal solution that compromises the performance on downstream tasks. Fig. 4.1 (a) shows an example 3D brain scan, and Fig. 4.1 (b) shows the different 2D planes (axes or views) on which this scan may be projected. No matter which axis is used to project the scan in the form of 2D slices, it will incur information loss compared to when the full 3D context is used. Therefore, we believe that employing the full 3D spatial context can substantially benefit the performance in downstream tasks, as it captures the anatomical information better.

In this chapter, we present six different self-supervised tasks that aim to take advantage of the full 3D spatial context from input medical scans. Our methods facilitate neural network representation learning from *unlabeled* 3D images, hence reducing the required cost for manual expert annotation. These proposed proxy tasks are: 3D Contrastive Predictive Coding, 3D SimCLR, 3D Rotation prediction, 3D Jigsaw puzzles, Relative 3D patch location, and 3D Exemplar networks. To the best of our knowledge, these algorithms have never been applied on 3D inputs, including on 3D medical scans. Nevertheless, as detailed earlier in Sec. 2.3, the 2D versions of these algorithms are successful in learning data representations from 2D natural images. We perform extensive experiments to evaluate the quality of the learned semantic representations in three different downstream tasks from four dataset benchmarks. The experimental results show that our 3D tasks learn rich data representations by improving the data-efficiency and performance on downstream tasks. More importantly, our experimental results support that pretraining with our proposed 3D algorithms yields more powerful semantic representations compared to pretraining on 2D slices.

Naturally, few computational and methodological challenges arise when operating on 3D inputs, due to the increased data dimensionality, which we make sure to address when designing and implementing[4] our self-supervised tasks. We provide the details in each respective algorithm in the methods section.

4 https://github.com/HealthML/self-supervised-3d-tasks

4.2 Related work

Works that employ Deep Learning algorithms on 3D input data are numerous [GB19; Guo+20; Ioa+17; Li+20; Su+17], and they have been applied to several application fields. Examples applications include Autonomous Driving with Lidar inputs, Robotics with 3D point cloud inputs, Video data with temporal information as a third dimension, Medical domain with 3D scans, and many more. In this section, however, we focus on self-supervised tasks that learn from volumetric 3D data, similar to medical scans.

Self-supervision from Videos. Videos are rich with several supervisory signals [PG16b; Von+18; VPT15; WG15a; WGH15] for self-supervised tasks. For instance, one may employ the temporal information across the scene frames besides the spatial context in each frame (image). In this paragraph, we discuss a subset of these works that use 3D-CNNs to process the videos. To that end, 3D-CNNs are usually employed to simultaneously extract the spatial features from each frame, and the temporal features across multiple frames. Here, the frames are typically stacked along the 3^{rd} (depth) dimension of the 3D-CNN. This scheme of exploiting 3D convolutions for video inputs was first proposed in [Ji+13] for human action recognition, and was later extended to other application types [JT20]. In self-supervised learning, however, the number of tasks that use 3D convolutions is limited. Kim *et al.* [DCK19] proposed a task that solves "cubic" puzzles of $2 \times 2 \times 1$ created from videos, yet the 3^{rd} depth dimension is not actually employed in puzzle creation. Jing *et al.* [JT18] extended the rotation prediction task [GSK18] to video inputs, by stacking the frames along the depth dimension. Nevertheless, the latter dimension is essentially not used in rotation operations, since only spatial rotations are considered. On the other hand, in our proposed versions of 3D Jigsaw puzzles and 3D Rotation prediction, respectively, we exploit the depth (3^{rd}) dimension effectively. For instance, we solve larger 3D puzzles (up to $3 \times 3 \times 3$), and we also predict more rotations along all axes in the 3D space. Generally, we believe that the different nature of the data, i.e. stacked video frames versus 3D volumetric scans in our case, affects the design of the respective proxy tasks. In other words, the 3^{rd} depth dimension has an actual meaning in volumetric scans. Therefore, we consider the full 3D spatial context in the design of all our tasks in order to learn the rich anatomical information from unlabeled volumetric scans.

Self-supervision from 3D medical scans. In a more related set of works, Zhou *et al.* [Zho+19b] proposed to employ image reconstruction on 3D medical scans as a source of self-supervision. Zhuang *et al.* [Zhu+19] and Zhu *et al.* [Zhu+20a] developed a proxy task that solves small $2 \times 2 \times 2$ jigsaw puzzles. Contrarily, our

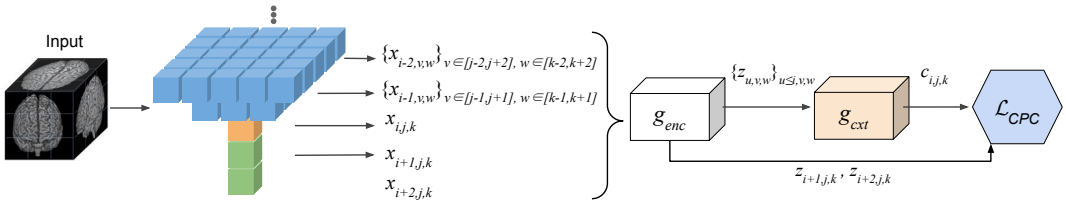


Figure 4.2: 3D-CPC: each input image is split into 3D patches, and the latent representations $z_{i+1,j,k}, z_{i+2,j,k}$ of **next patches** $x_{i+1,j,k}, x_{i+2,j,k}$ are predicted using the context vector $c_{i,j,k}$. The considered context is **the current patch** $x_{i,j,k}$, plus **the above patches** that form an inverted pyramid.

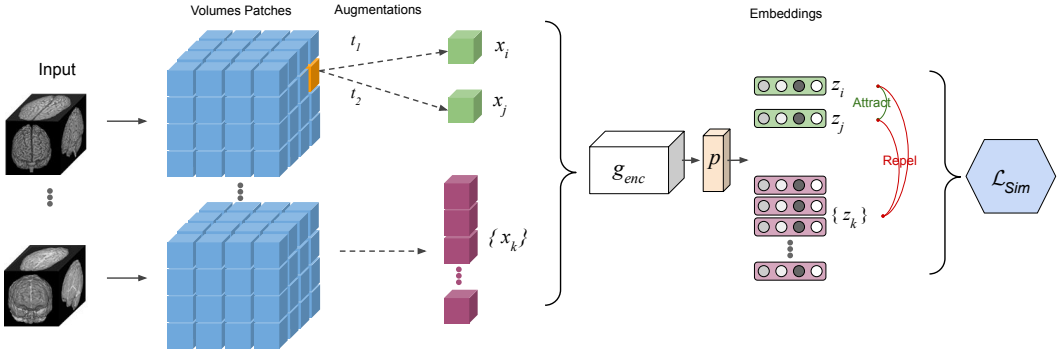


Figure 4.3: 3D-SimCLR: input volumes are split into 3D patches. Then, **positive samples** are created for each **anchor patch**. The remaining patches from other volumes are used as **negative samples**. The latent representations of positive samples z_i, z_j are attracted together in the embedding space, and repelled from the embeddings of negative samples $\{z_k\}$.

version of 3D Jigsaw puzzles may efficiently solve larger puzzles of $3 \times 3 \times 3$, and hence it can improve the performance on downstream tasks.

4.3 Methods

The methods we present in this section are all 3D self-supervised tasks, which learn semantic data representations from unlabeled 3D scans. The learned representations are all stored in the form of neural network weights in the resulting encoder model g_{enc}, allowing for subsequent label-efficient fine-tuning on downstream benchmarks.

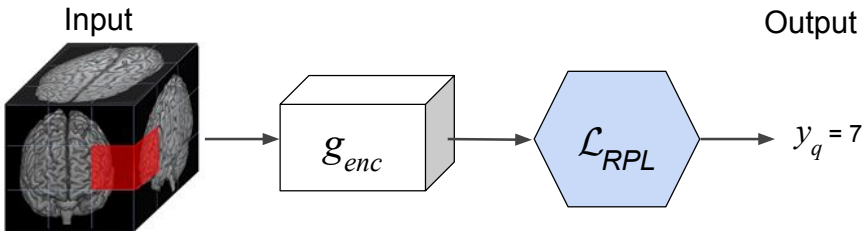


Figure 4.4: 3D-RPL: assuming a 3D grid of 27 patches ($3 \times 3 \times 3$), the model is trained to predict the location y_q of **the query patch** x_q, relative to the central patch x_c (whose location is 13).

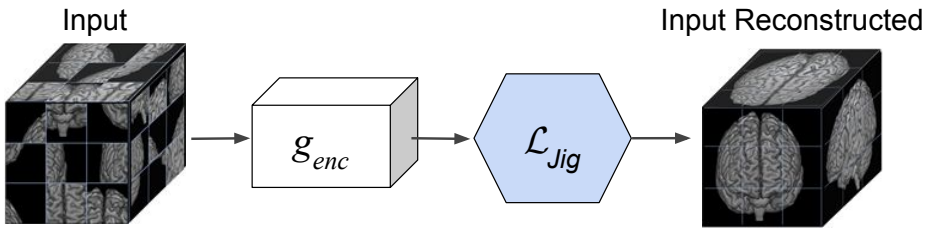


Figure 4.5: 3D-Jig: by predicting the permutation applied to the 3D image when creating a $3 \times 3 \times 3$ puzzle, we are able to reconstruct the scrambled input.

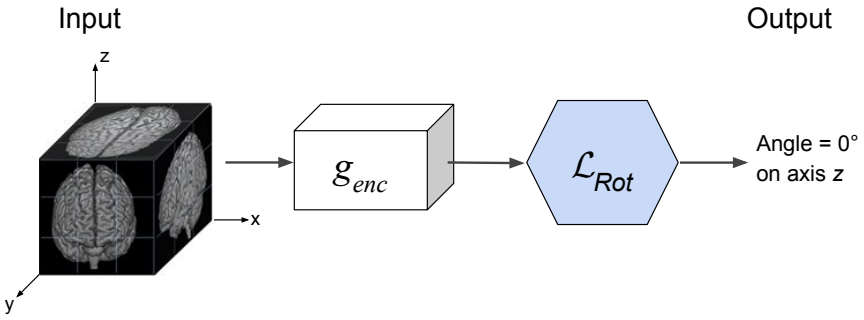


Figure 4.6: 3D-Rot: the network is trained to predict the rotation degree (out of the 10 possible degrees) applied on input scans.

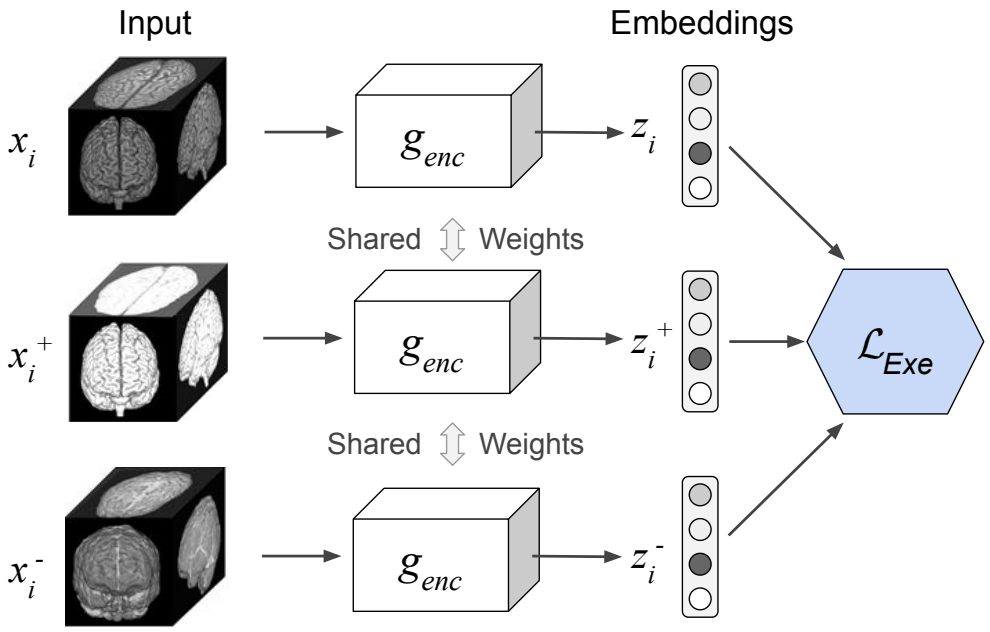


Figure 4.7: 3D-Exe: the network is trained with a triplet loss, which drives positive samples closer in the embedding space (x_i^+ to x_i), and the negative samples (x_i^-) farther apart.

4.3.1 3D Contrastive Predictive Coding (3D-CPC)

Contrastive learning approaches have found recent success in unsupervised representation learning tasks, as described earlier in Sec. 2.3.2. Contrastive methods are a family of algorithms, which at their core all use negative sampling techniques to train the model. Negative sampling was first employed in the Noise Contrastive Estimation (NCE) [GH10] algorithm, and it made learning word representations from the context of the sentence and given a large vocabulary list more tractable [Mik+13]. Contrastive Predictive Coding (CPC) [Hen20; OLV18] is an example of the contrastive methods family. This universal technique learns by predicting the latent representations for future, e.g. next or adjacent, samples. CPC has been applied to 1D Audio Signals [OLV18], and to 2D Natural Images [Hen20; OLV18]. Our proposed CPC algorithm, illustrated in Fig. 4.2, generalizes this technique to 3D image inputs, hence the naming 3D-CPC.

The first step of the algorithm is cropping each input 3D scan to equally-sized and overlapping 3D patches. Then, the encoder model being trained g_{enc} maps each patch $x_{i,j,k}$ to its latent representation $z_{i,j,k} = g_{enc}(x_{i,j,k})$. Next, the latent vectors of the patches are processed by a subsequent model called the context network g_{cxt}.

The latter network aims to summarize the contents of the patches in the context of $x_{i,j,k}$ and produce the context vector $c_{i,j,k} = g_{cxt}(\{z_{u,v,w}\}_{u \leq i,v,w})$, where $\{z\}$ denotes the set of latent vectors. Finally, since $c_{i,j,k}$ is assumed to capture the high level content of the context of $x_{i,j,k}$, it can be used for predicting the latent representations $z_{i+l,j,k}$ of next (adjacent) patches in the same context, where $l \geq 0$. However, in order to realize this prediction task, it is cast as an N-way classification problem by utilizing the InfoNCE loss [OLV18], which derives its name from its ability to maximize the Mutual Information between $c_{i,j,k}$ and $z_{i+l,j,k}$ and the fact that it uses NCE. In this formulation, the N classes are the patches latent representations $\{z\}$, among which is one target *positive* representation and the other $N-1$ classes are *negative*. Formally, the CPC loss can be written as follows:

$$\begin{aligned} \mathcal{L}_{CPC} &= -\sum_{i,j,k,l} \log p(z_{i+l,j,k} \mid \hat{z}_{i+l,j,k}, \{z_n\}) \\ &= -\sum_{i,j,k,l} \log \frac{\exp(\hat{z}_{i+l,j,k} z_{i+l,j,k})}{\exp(\hat{z}_{i+l,j,k} z_{i+l,j,k}) + \exp(\sum_n \hat{z}_{i+l,j,k} z_n)} \end{aligned} \tag{4.1}$$

This loss corresponds to the categorical cross-entropy loss. The target positive representation is $z_{i+l,j,k}$, and $\{z_n\}$ is the list of negative representations. These negative 3D patches, from which the negative representations are extracted, are chosen from other random locations in the input scan. In practice, similar to the original NCE [GH10] formulation, the classification task is solved as a *binary* pairwise classification task, in order to make the task more tractable.

The 3D context of each patch $x_{i,j,k}$ resembles an inverted pyramid neighborhood, which is inspired from [Sto+15; VKK16]. This particular context is chosen based on a tradeoff between computational cost and performance. Large contexts (e.g. full surrounding of a patch) incur prohibitive computations and memory use. The inverted-pyramid context was an optimal tradeoff. It is noteworthy that the proposed 3D-CPC task may employ any network architecture in the encoder g_{enc} and the context g_{cxt} networks. More architecture details in Appendix B.

4.3.2 3D Simple Contrastive Learning of Representations (3D-SimCLR)

Another more recent successful variant from the contrastive learning family is the *Sim*ple framework for *C*ontrastive *L*earning of visual *R*epresentations (SimCLR) algorithm. First proposed by Chen *et al.* [Che+20a], this method was able to advance the results of self-supervised learning methods to make their performance

comparable with supervised baselines on the ImageNet [Den+09] benchmark. SimCLR leverages the normalized temperature-scaled cross-entropy loss (NT-Xent) loss, which at its core also relies on negative sampling [GH10]. However, there are two main differences to CPC approaches ([Hen20; OLV18] and our 3D variant proposed in Sec. 4.3.1). In CPC, the processed samples by the encoder are usually image patches or audio parts derived from the same signals (samples), but in SimCLR usually full images are processed, thus allowing for capturing larger context inherently. In CPC, this is solved by using a context aggregation network, which is required to perform the autoregressive prediction task. In SimCLR, on the other hand, no autoregressive context network is required, therefore decoupling (or simplifying) the task to classify positive versus negative samples. Here, the positive samples are created by employing data augmentation techniques on each sample in the data batch. Hence, the NT-Xent loss used in SimCLR aims to maximize the similarity between latent representations of various augmented versions of the same sample, i.e. the sample and its associated positives. And at the same time, maximize the dissimilarity to the negative samples, which are augmented versions of the other samples in the same data batch. In essence, employing various augmentations improves the learned representations in contrastive methods by making them invariant to these augmentations.

In this section, we introduce a 3D-SimCLR algorithm, which operates on volumetric 3D input images, aiming to capture the full 3D spatial context of the scans. The first step is sampling a random batch of M 3D scans. Then, each 3D scan is split into P equally-sized non-overlapping 3D patches resulting in $N = M * P$ input samples. Before processing the input by the model, two composite augmentations are randomly chosen from the set of augmentations T, and then applied to each 3D patch leading to a dataset size of $2N$ samples. Hence, there exists one positive pair for every input sample, i.e. one pair originating from the same original 3D patch, and $2(N-1)$ negative pairs, i.e. originating from different 3D patches. It should be noted here that splitting each 3D scan into a set of 3D patches, in our 3D-SimCLR variant, is motivated by the different nature of 3D images in comparison to 2D images. Naturally, 3D images exhibit larger resolutions and hence are more computationally expensive when considering the whole scan as a sample. In terms of pretraining with 3D-SimCLR, this entails creating $2N$ positive and negative large samples, deeming the task prohibitively expensive memory and compute wise. Therefore, splitting each full scan into smaller 3D patches, similar to how our 3D-CPC task is formulated in Sec. 4.3.1, offers an optimal trade-off. It should be noted however, that in 3D-SimCLR, the positive and negative samples are created from the 3D patches in the data batch, which can stem from other volumes. As opposed to 3D-CPC, where these are only derived from each volume separately.

The training process with 3D-SimCLR, which is illustrated in Fig. 4.3, is as follows. First, the $2N$ samples in the data batch are processed by the encoder network g_{enc} to produce their hidden representations $\{h_n\}$. Then, a projection head, which is a non-linear fully-connected (dense) layer $p(.)$ is applied to produce the latent representations of each patch $\{z_n\}$. This non-linear projection head has been found useful in obtaining the representations of the samples in [Che+20a]. Afterwards, the loss function is evaluated. Here, we use the normalized temperature-scaled cross entropy loss [Che+20a], which takes the set of latent representations $\{z_n\}$ as inputs. Assuming the representations of a positive pair of 3D patches z_i and z_j and of a set of negative samples $\{z_k\}$:

$$l_{i,j} = -log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(sim(z_i, z_k)/\tau)} \tag{4.2}$$

where sim is the cosine similarity and τ is the temperature parameter. Finally, the overall NT-Xent loss, which we formalize by $\mathcal{L}_{Sim}$, is obtained by aggregating across all the pairs.

Image augmentations or transformations play a crucial role in contrastive learning methods, including 3D-SimCLR. As mentioned earlier, varying augmentation types inflicts invariance in the learned representations to these augmentations. In other words, this helps improve the model generalizability by exposing complicated patterns in the images. Nevertheless, the choice of augmentation types is rather an important aspect to avoid introducing structural changes on the image which might cause deterioration in the model performance. In addition, we find that varying the types of augmentation allows for accommodating domain knowledge, as we explore in a subsequent Chapter 6. In this 3D-SimCLR version, we employ the following augmentations set T:

- **3D Rotation**: The 3D patch is rotated on a randomly selected axis (x, y, z) with a randomly selected degree $(90°, -90°, 180°)$. Hence, effectively one random rotation is selected out of 9 possible ones.
- **3D Crop and Resize**: A cube is cropped out of the 3D patch at random a position, then it is resized to the original 3D image size. The size of the cube is a hyper-parameter. Additionally, the resized image is flipped on the z axis with 0.5 probability.
- **3D Cut Out**: A cube is cropped out of the 3D patch at a random position, then empty pixels are padded to fill the cropped out area. The size of the cube is a hyper-parameter.

- **Gaussian Noise**: Randomly generated Gaussian noise is added to the image pixel (here Voxel) intensities.
- **Gaussian Blur**: The 3D image is blurred using a Gaussian filter.
- **Sobel Filter**: The edges within the 3D images are emphasized using a Sobel filter.
- **Color Distortion**: The 3D image colors are distorted by consecutively applying a random brightness adjustment and a random contrast adjustment in a random order.
- **Identity**: The 3D image remains unchanged.

4.3.3 Relative 3D patch location (3D-RPL)

As one of the earliest self-supervised proxy tasks proposed, relative patch location (RPL) prediction [DGE15] utilizes the spatial context in images as a source of supervision. This task, in fact, is inspired from the skip-gram word2vec algorithm [Mik+13], where the anchor (center) word in a sentence is used to predict its surrounding words in the context window. In our proposed 3D-RPL algorithm version, depicted in Fig. 4.4, we leverage the full 3D spatial context in the design of this task. In each input 3D image, a grid of N non-overlapping 3D patches $\{x_i\}_{i\in\{1,..,N\}}$ is sampled at random locations. Then, the patch x_c in the center of the grid is used as a reference, and a query patch x_q is selected randomly from the surrounding $N-1$ patches. Next, the location of x_q relative to x_c is used as the positive label y_q. This casts the task as an $N-1$-way classification problem, in which the locations of the remaining grid patches are used as the negative samples $\{y_n\}$. The process is repeated to sample pairs of multiple positive examples. Formally, the cross-entropy loss in this task is written as:

$$\mathcal{L}_{RPL} = -\sum_{k=1}^{K} \log p(y_q \mid \hat{y}_q, \{y_n\}) \tag{4.3}$$

where K is the number of queries extracted from all samples. In order to prevent the model from solving this task quickly by finding shortcut solutions, e.g. edge continuity, we follow [DGE15] in leaving random gaps (jitter) between neighboring 3D patches. More details in Appendix B.

4.3.4 3D Jigsaw puzzle Solving (3D-Jig)

The relative patch location prediction task has inspired solving Jigsaw puzzles as a natural extension [NF16] for a self-supervised proxy task. In our proposed 3D Jigsaw puzzle task, which is illustrated in Fig. 4.5, the puzzles are formed by sampling an $n \times n \times n$ grid of 3D patches. Then, these 3D patches are shuffled according to an arbitrary permutation, selected randomly from a set of predefined permutations. This set of permutations with size P is chosen out of the $n^3!$ possible permutations, by following the Hamming distance based algorithm in [NF16] (details in Appendix B), and each permutation is assigned an index $y_p \in \{1, .., P\}$. Therefore, the task is reformulated as a P-way classification problem, i.e. the model is trained to simply recognize the applied permutation index p. This allows solving the 3D Jigsaw puzzles in a computationally efficient manner. Formally, we minimize the cross-entropy loss of $\mathcal{L}_{Jig}(y_p^k, \hat{y}_p^k)$, where $k \in \{1, .., K\}$ is an arbitrary 3D puzzle from the list of extracted K puzzles. Similar to 3D-RPL, we use the trick of adding random jitter in 3D-Jig.

4.3.5 3D Rotation prediction (3D-Rot)

Rotation prediction is one of the intuitive proxy tasks, for which the supervision labels for the self-supervision pretraining stage are obtained simply by predicting the angle of rotation applied artificially on input images. Originally proposed by Gidaris *et al.* [GSK18], the rotation prediction task encourages the model to learn visual representations that are invariant to the rotation transformation. In our proposed 3D Rotation prediction task, input 3D images are rotated randomly by a random degree $r \in \{1, .., R\}$ out of the R considered degrees. In this task, for simplicity, we consider the multiples of 90 degrees (0°, 90°, 180°, 270°, along each axis of the 3D coordinate system (x, y, z). There are 4 possible rotations *per axis*, amounting to 12 possible rotations. However, rotating input scans by 0° along the 3 axes will produce 3 identical versions of the original scan, hence, we simply consider 10 rotation degrees. Therefore, in this setting, the 3D rotation prediction task is solved as a 10-way classification problem, as shown in Fig. 4.6. Formally, we minimize the cross-entropy loss $\mathcal{L}_{Rot}(r^k, \hat{r}^k)$, where $k \in \{1, .., K\}$ is an arbitrary rotated 3D image from the list of K rotated images. It is noteworthy that we create multiple rotated versions for each 3D image.

4.3.6 3D Exemplar networks (3D-Exe)

The task of Exemplar networks, proposed by Dosovitskiy *et al.* [Dos+14] derives supervision labels using image augmentation techniques, i.e. transformations. This task, in particular, is viewed in literature to have inspired the family of contrastive algorithms, such as [Che+20a; Gri+20; He+20]. The core idea here is that this line of algorithms employs image transformations (augmentations) as a source of supervision, aiming to learn transformation-invariant data representations. In the original formulation of exemplar networks, the task is cast as a classification task. Assuming a training set of $X = \{x_1, ...x_N\}$, and a set of K image transformations $\mathcal{T} = \{T_1, ..T_K\}$, a new surrogate class S_{x_i} is created by transforming each training sample $x_i \in X$, where $S_{x_i} = \mathcal{T}x_i = \{Tx_i \mid T \in \mathcal{T}\}$. However, this classification task becomes prohibitively expensive as the dataset size grows larger, as the number of classes grows accordingly, since each class represents a sample and its augmented versions. Thus, in our proposed 3D version of Exemplar networks, shown in Fig. 4.7, we employ a different mechanism that relies on the triplet loss instead [WG15b]. Formally, assuming x_i is a random training sample and z_i is its corresponding embedding vector, x_i^+ is a transformed version of x_i (seen as a positive example) with an embedding z_i^+, and x_i^- is a different sample from the dataset (seen as negative) with an embedding z_i^-. The triplet loss is written as follows:

$$\mathcal{L}_{Exe} = \frac{1}{N_T} \sum_{i=1}^{N_T} \max\{0, D(z_i, z_i^+) - D(z_i, z_i^-) + \alpha\} \tag{4.4}$$

where $D(.)$ is a pairwise distance function, for which we use the L_2 distance, following [SKP15]. α is a margin (gap) that is enforced between positive and negative pairs, which we set to 1. The triplet loss enforces $D(z_i, z_i^-) > D(z_i, z_i^+)$, i.e. the transformed versions of the same sample (positive samples) to come closer to each other in the learned embedding space, and farther away from other (negative) samples.

In terms of 3D transformations used to create the positive and negative samples, we apply: random flipping along an arbitrary axis, random rotation along an arbitrary axis, random brightness and contrast, and random zooming.

It is noteworthy that by replacing the triplet loss with a contrastive loss [GH10] converts the Exemplar networks algorithm to SimCLR [Che+20a]. In other words, the triplet loss can be seen as a special case of the contrastive loss, where the number of positive and negative samples for each anchor sample is one each.

4.4 Experimental Results

We present the evaluation results for our developed self-supervised methods in this section. In order to assess the quality of the learned representations by our methods, we fine-tune them on two different downstream tasks. To perform the evaluation in each downstream task, we analyze the obtained gains in data-efficiency, performance, and speed of convergence. Additionally, each downstream task demonstrates a certain use-case for our methods. We follow the commonly used evaluation protocols for self-supervised methods in each of these downstream tasks. Namely:

- Brain Tumor Segmentation on 3D MRI (Sec. 4.4.1): in which we study the possibility for transfer learning from a different unlabeled 3D corpus, following [Goy+19].
- Pancreas Tumor Segmentation on 3D CT (Sec. 4.4.2): to demonstrate how to use the same unlabeled dataset, following the data-efficient evaluation protocol in [Hen20].

We provide additional details about architectures, training procedures, augmentation details, and decoders initialization for segmentation tasks in the Appendix B.

4.4.1 Brain Tumor Segmentation Results

In this downstream task, we evaluate the representations learned by our 3D self-supervised methods by fine-tuning them on the Multimodal Brain Tumor Segmentation (BraTS) 2018 [Bak+17; Men+15] dataset. However, the model semantic representations are all obtained by pretraining with our methods on the Brain MRI data from the UK Biobank [Sud+15] (UKB) corpus. We use roughly $22K$ 3D Brain MRI scans from the UK Biobank. Due to this large number of unlabeled scans, UKB is suitable for unsupervised pretraining. The BraTS dataset contains labelled MRI scans for 285 training and 66 validation cases. We fine-tune on the BraTS' training set, and we evaluate on the validation set. Following the official BraTS challenge, we report Dice scores for the Whole Tumor (WT), Tumor Core (TC), and Enhanced Tumor (ET) tasks. The Dice score (F1-Score) is twice the area of overlap between two segmentation masks divided by the total number of pixels in both.

In order to assess the quality of the learned representations by our 3D self-supervised methods, we compare to the following baselines:

- Training from scratch: the first sensible baseline for any self-supervised method, in general, is the same model architecture trained on the downstream

task when initialized from random weights. Comparing to this baseline provides insights about the benefits of self-supervised pretraining.

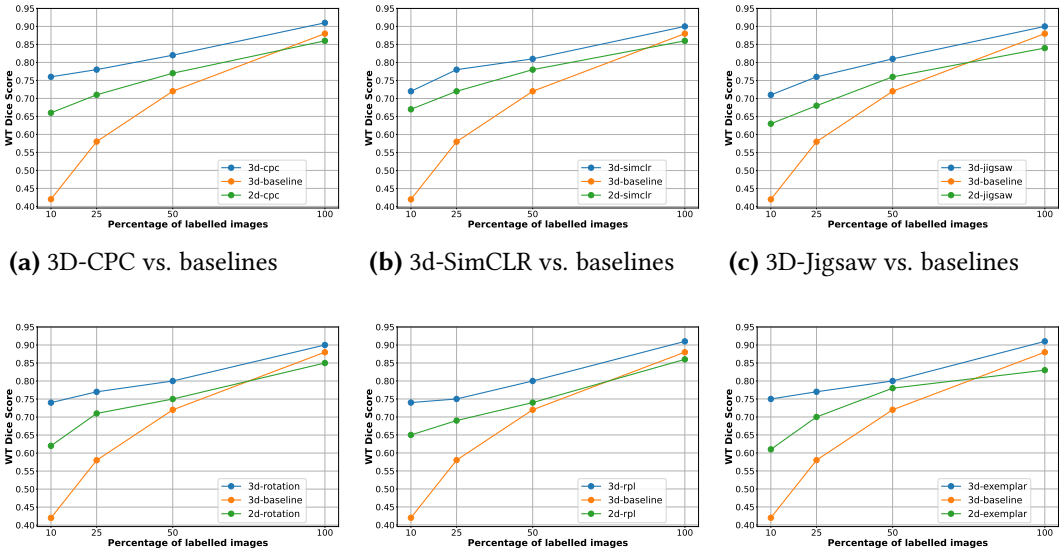
- Training on 2D slices: this baseline aims to quantitatively demonstrate how operating on the 3D spatial context benefits the learned representations, compared to the 2D spatial context.

- Supervised pretraining: this baseline was trained with automatically generated segmentation labels by FSL-FAST [Woo+09] for UK Biobank scans. The labels include masks for three brain tissues.

- Baselines from the BraTS challenge: we compare to the methods [Bai+18; Cha+18a; Ise+18; PAP18], which all use a single model with an architecture similar to ours, i.e. 3D U-Net [RFB15].

Discussion. We first assess the gains in data-efficiency in this task. To quantify these gains, we measure the segmentation performance at different sample sizes. We randomly select subsets of patients samples at rates 10%, 25%, 50%, and 100% of the full dataset size, and we fine-tune the pretrained models on these subsets. As shown in Fig. 4.8, the models pretrained with our 3D methods outperform the baseline model trained from scratch by a large margin when using few training samples, and behaves similarly as the number of labeled samples increases. The low-data regime case at 5% suggests the potential for generic unsupervised features, and highlights the huge gains in data-efficiency. In addition, the models pretrained with our proposed 3D versions considerably outperform their 2D counterparts, which are trained on slices extracted from the 3D images. The latter result confirms that learning from the 3D spatial context considerably improves the quality of the learned representations.

Pretraining with our self-supervised methods is also expected to improve the overall segmentation performance of the models, an aspect we assess in this downstream task by measuring Dice scores on the full BraTS dataset. As Tab. 4.1 shows, all models pretrained with our 3D methods outperform the baseline trained from scratch as well as their 2D counterparts. Additionally, our models achieve comparable results to the baselines from the BraTS challenge, and, in some cases, our models outperform these baselines, e.g. our 3D-RPL method outperforms all baselines in terms of ET and TC dice scores. Moreover, the model pretrained with 3D-Exemplar matches the result of Isensee *et al.* [Ise+18] in terms of WT dice score, which is one of the top results on the BraTS 2018 challenge, even though 3D-Exemplar uses fewer downstream training epochs. In comparison to the supervised baseline



(a) 3D-CPC vs. baselines **(b)** 3d-SimCLR vs. baselines **(c)** 3D-Jigsaw vs. baselines

(d) 3D-Rotation vs. baselines **(e)** 3D-RPL vs. baselines **(f)** 3D-Exemplar vs. baselines

Figure 4.8: Data-efficient segmentation results in BraTS. With less labeled data, the **supervised baseline** fails to generalize, as opposed to **our methods**. Also, the proposed 3D methods outperform all **2D counterparts.**

pretrained with automatic FAST labels (3D Supervised), we find that our results are comparable, outperforming this baseline in some cases.

Our results in this downstream task also demonstrate a generalization ability of our 3D tasks across different domains, i.e. our models are pretrained on UK Biobank and fine-tuned on BraTS. This result is significant, because medical datasets are supervision-starved, e.g. images may be collected as part of clinical routine, but much fewer high-quality labels are produced, due to annotation costs.

4.4.2 Pancreas Tumor Segmentation Results

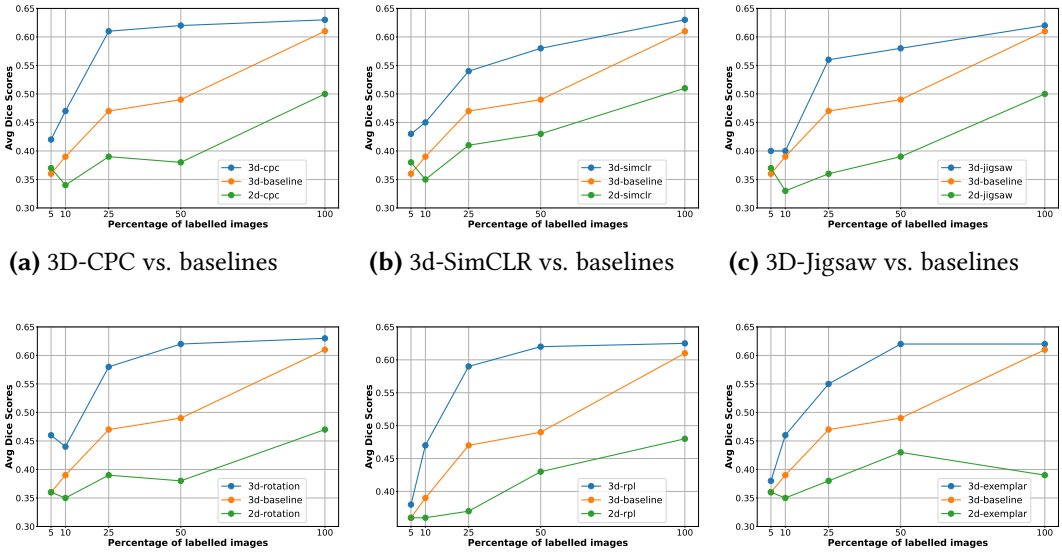
In this downstream task, we evaluate models pretrained with our methods on 3D CT scans of Pancreas tumor from the medical decathlon benchmarks [Sim+19b]. The Pancreas dataset contains annotated CT scans for 420 cases, where each scan contains 3 different classes: pancreas (class 1), tumor (class 2), and background (class 0). To measure the performance on this benchmark, two dice scores are computed for foreground classes 1 (pancreas) and 2 (tumor). The nature of the pancreas dataset in this downstream task is challenging, because the distribution of the

Table 4.1: Segmentation results of the proposed 3D SSL methods on BraTS data

Model	ET	WT	TC
3D-From scratch	76.38	87.82	83.11
3D Supervised	78.88	90.11	84.92
2D-CPC	76.60	86.27	82.41
2D-SimCLR	77.36	86.33	82.77
2D-RPL	77.53	87.91	82.56
2D-Jigsaw	76.12	86.28	83.26
2D-Rotation	76.60	88.78	82.41
2D-Exemplar	75.22	84.82	81.87
Popli *et al.* [PAP18]	74.39	89.41	82.48
Baid *et al.* [Bai+18]	74.80	87.80	82.66
Chandra *et al.* [Cha+18a]	74.06	87.19	79.89
Isensee *et al.* [Ise+18]	80.36	**90.80**	84.32
3D-CPC	80.83	89.88	85.11
3D-SimCLR	79.76	90.02	84.98
3D-RPL	**81.28**	90.71	**86.12**
3D-Jigsaw	79.66	89.20	82.52
3D-Rotation	80.21	89.63	84.75
3D-Exemplar	79.46	**90.80**	83.87

classes exhibits imbalance, i.e. the tumor class is rare compared to the background class.

In the self-supervised stage, the models are pretrained with our proposed 3D self-supervised methods on pancreas scans by discarding their associated masks (labels). Afterwards, the models are finetuned on subsets of annotated data to assess the gains in both data-efficiency and performance. In terms of baselines, similarly to the previous downstream task, we establish the baseline model trained from scratch as well as models trained on 2D slices extracted from the pancreas data samples. The obtained gains in data-efficiency are illustrated in Fig. 4.9. To quantify these gains, we fine-tune the pretrained models on 5%, 10%, 25%, 50%, and 100% of the full dataset size. Then, we evaluate the fine-tuned models on a held-out labeled test set from the Pancreas dataset that was not used for pretraining nor fine-tuning. The results obtained by our 3D methods outperform the baselines in this task with a margin when using only few training samples, e.g. 5% and 10% cases. This behavior is also consistent when the data size increases to 50% and 100% of the full size, confirming the benefits on downstream performance in this task too. We provide additional experimental details in Appendix B.



(a) 3D-CPC vs. baselines **(b)** 3d-SimCLR vs. baselines **(c)** 3D-Jigsaw vs. baselines

(d) 3D-Rotation vs. baselines **(e)** 3D-RPL vs. baselines **(f)** 3D-Exemplar vs. baselines

Figure 4.9: Data-efficient segmentation results in Pancreas. With less labeled data, the **supervised baseline** fails to generalize, as opposed to **our methods**. Also, the proposed 3D methods outperform all **2D counterparts.**

4.5 Discussion

In this Chapter, we asked whether designing self-supervised methods to operate on the full 3D spatial context could benefit the learned representations from unlabeled 3D images, and found that it is indeed the case. Employing the 3D context in self-supervision improves downstream performance, an effect that appears larger when fine-tuned on only few samples of labeled 3D data. In other words, learning data representations from 3D data provides considerable gains in data (or label) efficiency, in a way reducing the efforts of required manual annotation. This result is significant for the medical imaging domain, where data and annotation scarcity is an obstacle.

We showcase the obtained gains in downstream data-efficiency, performance, and even speed of convergence (see Appendix B) on two semantic segmentation tasks. To highlight these gains, we compare models pretrained with our 3D methods to the baselines of training from scratch and of pretraining on 2D inputs. The experimental results demonstrate how utilizing the 3D context improves the representation quality considerably. Furthermore, we observe performance gains when pretraining models on a large unlabeled corpus with our proposed methods, and subsequently

fine-tuning them on a different smaller downstream-specific dataset. This result suggests alternatives for transfer learning from ImageNet features, which can be substantially different from the medical domain.

5 Self-supervision from Medical Images with other Modalities

This chapter extends own work in [Tal+22a], published in the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022). Therefore, many figures, tables, and statements have been quoted verbatim or reproduced with permission.

5.1 Introduction

Disease patterns in medical images are numerous and their incidence exhibits a long tail distribution in nature [Zho+21]. In other words, most diseases are infrequent in clinics, and only a small number of common diseases have sufficiently observed cases to allow large-scale analysis [BLK14]. Therefore, as a natural next step for improving our knowledge of disease traits is to fuse knowledge from additional data modalities, e.g. genomics or clinical data. In this Chapter, we investigate integrating imaging with genetic modalities, which, as we elaborate below, have significant causal relations to diseases.

Biobanks are organized collections of biological materials and associated information stored for research purposes [Hew11]. While biobanks may include plant or animal material, the term is mostly used to refer to datasets of human specimens. Recently, large-scale biobank studies have begun to aggregate unprecedented quantities of multimodal data on human health. For example, the UK Biobank (UKB) [Sud+15] contains data for 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) [Bioc], BioMe [Bioa], FinnGen [Fin], Estonia Biobank [Biob], and others. While some of these biobanks 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 [Oma+05], deeming learning from them challenging and error-prone.

On the other hand, genetic data is increasingly abundant in these studies. Studying common genetic variation at scale have been made possible by chip-based genotyping technologies [Ver+21]. In addition, the exponentially decreasing costs of genomic sequencing is driving progress for rare genetic variation [PK16]. Due to

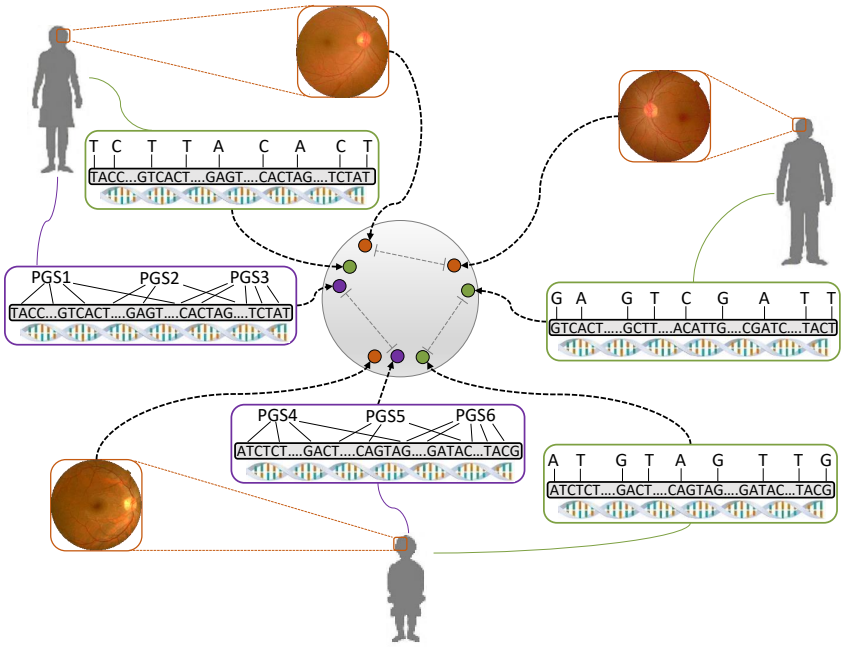


Figure 5.1: Overview of our contrastive learning method from imaging and genomic data. It learns representations by bringing the modalities of each individual closer in the embedding space, and apart from different individuals'. In this example, the modalities are retinal fundus images (in **brown**), 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). Figure source [Tal+22a], reprinted with permission.

these advances, the UKB and other biobanks often contain a rich array of genetic 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 [WVW14]. Similarly, the majority of other traits that are not directly related to diseases, *e.g.* height and human personality, are also strongly influenced by genetics [Lip+17; Zwi+20]. Similarly, complementary imaging-genetics datasets are increasingly also available in other application settings, *e.g.* plant breeding [Yan+20].

Unlabelled medical images carry valuable information about organ structures, and an organism's genome is the blueprint for biological functions in the body. Clearly, integrating these distinct yet complementary data modalities can help create a more holistic picture of physical and disease traits. Such integration step, however, is non-trivial and challenging. The human genome consists of three billion base pairs, yet most genetic differences between individuals have little ef-

fect. 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) [Insb; Man10] 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 [Insc]. 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" [Man+09]. Therefore, seeking better methods or solutions may help in finding and explaining the relationships between genetic and imaging modalities.

The growing number of biobanks of imaging-genetics data, which are unlabeled and multimodal in nature, calls for solutions that can: *(i)* learn semantic data representations without requiring costly expert annotations, *(ii)* integrate these data modalities end-to-end in an efficient manner, and *(iii)* explain discovered cross-modal correspondences (associations). Hence, as explained in earlier Chapters, self-supervised representation learning offers a pertinent solution when unlabeled data is abundant and labels are scarce. Furthermore, these algorithms also 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, 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 labelling challenges inherent to the medical domain.

As a result, we propose a novel self-supervised method in this Chapter, called ContIG, that can learn from datasets of unlabeled medical images and genetic data modalities. ContIG is short for Contrastive Learning for Medical Imaging with Genetics. This algorithm aligns imaging and genetic 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. Nevertheless, a main requirement we proposed above is explainability of discovered imaging-genetic associations, for which we adapt gradient-based explainability algorithms. Our method discovers interesting associations across these modalities, and we confirm their relevance by cross-referencing biomedical literature. We evaluate the representations learned by ContIG via transfer learning on several downstream tasks, and the results outperform state-of-the-art self-supervised methods on all benchmarks. We also perform genome-wide

association studies on the learned features, and we find they uncover interesting relationships between images and genetic data.

5.2 Related Work

Deep learning from multiple data modalities. Learning from multimodal data presents several inherent challenges, such as: multimodal fusion, alignment, and representation [BAM19; Ngi+11]. Prior works, some of which are self-supervised, learn from diverse sets of modalities, such as: image with text (vision and language) [Ayt+18; Joh+16; Li+19; Lu+19; Sun+19a; Sun+19b; TB19; Xu+15], image with audio [Alw+19; Asa+20; AVT16; AZ17; OE18; Owe+18], audio with text [AGG18; YBJ18], and multi-view (multimodal) images [PG16a; SBO18; SZ14; TKI20]. More recent self-supervised works employed contrastive learning for multimodal inputs, *e.g.* images with text captions [Ala+20; Pat+21; Rad+21; Yua+21; Zha+20a; Zha+20b]. We follow this line of work, and we extend contrastive pretraining to novel modalities, *i.e.* images and genetics, for the first time.

Deep learning from imaging with genetics. Not only have deep learning methods been successfully applied to medical imaging [Ma+20], they also found success in application domains with genomics data [Era+19; Kou20; Wu+21; Zit+19; Zou+19]. Few recent works have utilized deep learning to learn jointly from both data modalities, such as [Ash+21; BS20; Cha+18c; Dai+21; Fuj+21; Gun+20; Kir+21; Nin+18; Ven+21; Zho+19a]. 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 combined imaging and genetic data.

5.3 Methods

In Sec. 5.3.1 we first review some biomedical foundations and motivate the genetic modalities chosen in this Chapter, which exhibit complementary biological properties. Then, we describe our contrastive method in Sec. 5.3.2, and different modality aggregation types. Finally, we detail the explanation methods for genetic features in Sec. 5.3.3. Some passages in these subsections have been quoted verbatim from own work in [Tal+22a], and are only insignificantly changed.

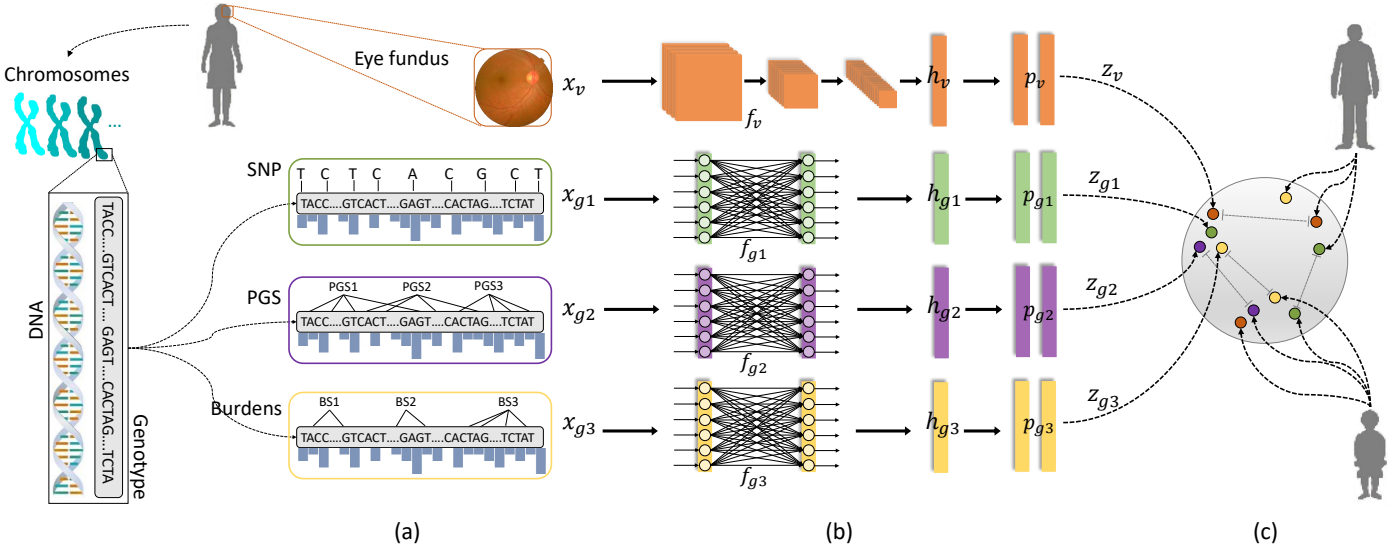


Figure 5.2: Schematic illustration for the steps of our proposed method. **(a)** Assuming one imaging modality (retinal fundus shown in **brown**), and three genetic modalities (Single-nucleotide polymorphisms (SNP) in **green**, polygenic risk scores (PGS) in **purple**, burden scores in **yellow**). 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. Figure source [Tal+22a], reprinted with permission.

5.3.1 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 [Insa]. A relatively small fraction of the genome codes for proteins, while the remaining parts have regulatory or structural functions. Proteins make up much of the structural components of the cell, and enable the thousands of biochemical reactions needed for survival. 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 diseases (*e.g.* Alzheimer's). "Genotyping" is the process of measuring these genetic changes [RH05]. 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, we consider *three* different ways to encode the genetic modalities 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 [Yan+10]. Many common diseases and impairments are complex traits, which makes them especially relevant to human health applications [Fra+09]. To best encode genetic architectures associated with complex traits, we utilize **polygenic risk scores (PGS)** [Dud13]. PGS aggregate many, mostly common, SNPs into a single score that reflects a person's inherited susceptibility to a specific disease [Khe+18]. 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 [Wai+21]. Rare variants occur at low frequencies (*e.g.* MAF[5] $< 1\%$ 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 Chapter, we use **burden scores** [Lee+12], 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 [Lip+11], but also tags highly diverse functional information.

The three chosen genetic modalities – polygenic risk scores, burden scores, and raw SNPs – capture complementary aspects and together paint a broad description of an individual's genetic predisposition. We employ them both individually and jointly as contrastive views to medical images.

5.3.2 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. 5.2, 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

5 minor allele frequency

$z_v, z_{gm} \in \mathbb{R}^d$. Following [Che+20a], each projection head is a non-linear MLP with one hidden-layer.

Formulation of 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:

$$
\begin{aligned}
L(v, g) &= -\sum_{j=1}^{b} \log \frac{\exp(\cos(z_v^{i_j}, z_g^{i_j})/\tau)}{\sum_{k=1, k\neq j}^{b} \exp(\cos(z_v^{i_j}, z_g^{i_k})/\tau)} \\
\mathcal{L}_{cont}(v, g) &= \lambda L(v, g) + (1 - \lambda) L(g, v),
\end{aligned}
\tag{5.1}
$$

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 exists 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:

$$
\mathcal{L}(v, g_1, \ldots, g_M) = \sum_{m=1}^{M} \mathcal{L}_{cont}(v, g_m) \tag{5.2}
$$

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. (5.2), 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.

5.3.3 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. (5.2) 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. (5.2), but $v_r, g_{1,r}, \ldots, g_{M,r}$ fixed. We can then use standard feature attribution methods such as Integrated Gradients [STY17] or DeepLift [SGK17] 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.

The setting with missing values can be handled analogously to the "outer" aggregation scheme in Sec. 5.3.2, by just omitting the respective modalities.

5.4 Experimental Results

We present the evaluation results of our method in this section. First, we detail the datasets used for both pretraining and evaluation purposes in Sec. 5.4.1. Then, we assess the quality of the learned representations, by: i) fine-tuning (transfer learning) on four downstream tasks in Sec. 5.4.2, ii) linear evaluation of the learned representations on the same downstream tasks in Sec. 5.4.3, iii) quantifying the gains in labelled data efficiency in Sec. 5.4.4, and iv) performing a genome-wide association study (GWAS) on the model features in Sec. 5.4.5. Finally, we present the genetic feature explanation results in Sec. 5.4.6, and we analyze the findings to check their relevance with medical literature resources. Similarly, few passages in this section have been quoted verbatim from own work in [Tal+22a], and are only insignificantly changed.

5.4.1 Datasets

We pretrain our models (and the unsupervised baselines) on data obtained from the UK Biobank (UKB) dataset [Sud+15]. This dataset contains multimodal data for almost 500k individuals, although imaging data is only available for a subset of those. The UKB contains an overwhelming majority of individuals of European descent; we therefore restrict our pretraining dataset to European descent individuals, as including individuals from other populations would likely introduce very large confounding effects [Lip+11]. 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 (see Sec. 5.3.1), amount to 155, 238 Raw-SNP samples, 145, 206 PGS samples, and 93, 216 burden scores. In terms of feature dimensions, for the raw-SNPs, we uniformly sample every 100^{th} SNP from 22 Chromosomes (excluding the X and Y chromosomes), resulting in 7, 854 SNPs per sample. For PGS, we used 481 scores for a wide variety of different traits downloaded from the PGS Catalog [Lam+21]. We created burden scores for 18, 574 protein-coding genes [Mon+21]. These binary scores indicate whether a participant has at least one potentially damaging rare (MAF $< 1\%$) variant within a given gene. 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 [19] dataset for Diabetic Retinopathy detection in Sec. 5.4.2, which has 3, 662 retinal fundus training samples. ii) Retinal Fundus Multi-disease Image Dataset (RFMiD) [Pac+21] for disease classification (Sec. 5.4.2), which has 3, 200 training images. iii) 102, 219 images from the UKB [Sud+15] training split, but now we predict cardiovascu-

lar risk factors (Sec. 5.4.2). iv) Pathologic Myopia challenge dataset [Fu+19] for Pathological Myopia Segmentation (Sec. 5.4.2), which has 400 image samples with segmentation masks. More datasets details in the Appendix C.

5.4.2 Transfer Learning (Fine-Tuning) Results

In this section, we evaluate the quality of representations by fine-tuning to downstream tasks. In other words, we unfreeze the encoder layers weights in this set of experiments, as opposed to the results in Sec. 5.4.3.

Models & architectures. Across the following experiments, we train neural network models with our proposed method ContIG. For the image encoder part of the model (f_v in Fig. 5.2), we employ a Resnet50 [He+16]. For the genetic encoders (f_{gm}), we vary the number of fully connected layers: "None" hidden layers, one hidden layer "H1", and two hidden layers "H12". We also vary the combination of genetic modalities (detailed in Sec. 5.3.1) used in pretraining, along with modality aggregation schemes (explained in Sec. 5.3.2).

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. The comparison with this baseline provides insights about the benefits of pretraining.
- Contrastive methods from state-of-the-art: 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** [Che+20a], **BYOL** [Gri+20], **Barlow Twins** [Zbo+21], **SimSiam** [CH21], and **NNCLR** [Dwi+21].

Diabetic Retinopathy Detection (APTOS)

Millions of people suffer from Diabetic Retinopathy, the leading cause of blindness among working aged adults. The APTOS dataset [19] 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 [Coh68]), which measures the agreement between two rating sets. Its values vary from random (0) to complete agreement (1), and if

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 [Che+20a]	-	81.83	91.88	70.41	3.451	59.38
SimSiam [CH21]	-	75.44	91.28	72.26	3.442	57.37
BYOL [Gri+20]	-	71.09	89.88	66.32	3.414	59.73
Barlow Twins [Zbo+21]	-	72.28	92.03	70.53	3.430	59.05
NNCLR [Dwi+21]	-	77.93	91.89	72.06	3.426	61.95
ContIG (Raw-SNP)	None	81.99	92.27	74.96	3.366	64.71
ContIG (Raw-SNP)	H1	84.01	93.22	76.98	3.254	70.10
ContIG (Raw-SNP)	H12	82.56	93.09	77.02	3.201	69.58
ContIG (PGS)	None	83.84	91.63	76.86	3.257	69.81
ContIG (PGS)	H1	<u>85.93</u>	93.31	**78.47**	<u>3.176</u>	**72.72**
ContIG (PGS)	H12	**86.44**	93.04	77.04	3.216	70.69
ContIG (Burden)	None	82.92	**93.68**	76.89	3.273	71.91
ContIG (Burden)	H1	83.22	93.03	76.49	**3.160**	<u>72.37</u>
ContIG (Burden)	H12	83.61	93.14	76.72	3.236	71.50
ContIG (Inner RPB)	None	83.49	93.31	77.11	3.195	71.68
ContIG (Inner RPB)	H1	81.52	92.95	77.34	3.202	70.80
ContIG (Inner RPB)	H12	80.24	92.94	75.37	3.235	68.89
ContIG (Outer RPB)	None	82.93	93.01	76.31	3.260	69.16
ContIG (Outer RPB)	H1	84.22	<u>93.62</u>	76.97	3.187	71.80
ContIG (Outer RPB)	H12	84.21	93.41	<u>77.51</u>	3.233	71.13

Table 5.1: Evaluation results by fine-tuning on downstream tasks. **Bold** indicates the best result, <u>underlined</u> is second best. RPB in our method stand for the genetic modalities used: *R*aw-SNPs, *P*GS-scores, and *B*urden-scores. ↑ means higher is better, and ↓ lower is better.

there is less agreement than chance it may become negative. The evaluation results in Tab. 5.1 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.

Retinal Fundus Disease Classification (RFMiD)

The Retinal Fundus Multi-disease Image Dataset (RFMiD) [Pac+21] 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 [Lea]. The results for this task in Tab. 5.1 also demonstrate the gains in performance obtained by pretraining with ContIG. Our models also outperform the self-supervised baselines in this task.

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 [Fu+19] 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 [RFB15] on top of the ResNet50 encoder. In terms of evaluation metrics, we use the dice score [Sor]. The results of this task in Tab. 5.1 showcase the quality of the learned representations by ContIG on semantic segmentation.

Cardiovascular Risk Prediction

Previous work has shown that retinal fundus images can predict a range of risk factors for cardiovascular diseases [Pop+18]. 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 [Sud+15] 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. 5.1 shows, models pretrained with ContIG outperform the baseline models in both classification and prediction (regression) tasks.

5.4.3 Linear Evaluation Results

In this section, we follow a linear evaluation protocol [Che+20a; OLV18; ZIE16], meaning that the encoder weights are kept frozen and only a linear classifier / regressor is trained on top. Similarly to the fine-tuning protocol, linear evaluation aims to provide an idea about the quality of semantic representations stored in

Model & Genetics Encoder		APTOS	RFMiD	PALM	Cardio. Risk Pred.	
		QwKappa ↑	ROC-AUC ↑	Dice-Score ↑	MSE ↓	ROC-AUC ↑
SimCLR [Che+20a]	-	35.02	86.53	59.77	3.998	52.26
SimSiam [CH21]	-	21.25	87.91	56.58	3.998	53.13
BYOL [Gri+20]	-	17.39	87.84	54.04	4.009	52.29
Barlow Twins [Zbo+21]	-	44.75	87.65	59.52	3.952	54.28
NNCLR [Dwi+21]	-	24.76	85.80	66.25	3.870	54.17
ContIG (Raw-SNP)	None	59.14	89.24	72.82	3.683	59.07
ContIG (Raw-SNP)	H1	69.85	89.99	75.25	3.443	64.36
ContIG (Raw-SNP)	H12	68.72	90.47	74.39	3.439	69.58
ContIG (PGS)	None	66.34	88.16	75.03	3.488	62.64
ContIG (PGS)	H1	**72.38**	90.43	76.35	3.426	63.98
ContIG (PGS)	H12	70.20	90.01	**77.13**	3.481	63.27
ContIG (Burden)	None	70.29	<u>91.08</u>	75.31	3.453	64.72
ContIG (Burden)	H1	70.67	90.62	75.42	3.421	64.70
ContIG (Burden)	H12	<u>71.22</u>	**91.10**	76.09	3.434	<u>64.84</u>
ContIG (Inner RPB)	None	70.26	89.94	75.27	3.439	63.84
ContIG (Inner RPB)	H1	66.94	88.65	75.00	3.404	64.73
ContIG (Inner RPB)	H12	68.41	90.56	73.08	3.457	63.45
ContIG (Outer RPB)	None	66.94	90.38	75.29	3.448	**65.20**
ContIG (Outer RPB)	H1	66.60	89.46	<u>77.04</u>	<u>3.398</u>	64.59
ContIG (Outer RPB)	H12	68.57	90.51	76.50	**3.388**	**65.20**

Table 5.2: Downstream evaluation results by linear evaluation on each task. Similarly, the results obtained by ContIG outperform all baselines. **Bold** indicates the best result, <u>underlined</u> is second best. RPB in our method stand for the genetic modalities used: *R*aw-SNPs, *P*GS-scores, and *B*urden-scores. ↑ means higher is better, and ↓ lower is better.

the model encoder. However, linear evaluation attempts to minimize the influence of encoder weight changes due to downstream task loss gradients. This in a way evaluates the relevance of learned generic features by the unsupervised task, here by ContIG. We compare to the same baselines presented in Sec. 5.4.2, and we use the same model architectures and evaluation metrics. As shown in Tab. 5.2, the models trained with our method "ContIG" consistently outperform the baselines, confirming the high quality of features learned by "ContIG".

5.4.4 Data-Efficiency Results

In this section, we assess the quality of semantic representations by measuring the gains in labelled data-efficiency. To quantify these gains, we follow a semi-

Model	Label Fraction			
	1%		10%	
	MSE ↓	ROC ↑	MSE ↓	ROC ↑
SimCLR [Che+20a]	4.029	51.43	3.762	54.29
SimSiam [CH21]	3.861	53.35	3.564	57.45
BYOL [Gri+20]	3.894	51.68	3.505	56.71
Barlow Twins [Zbo+21]	3.788	51.89	3.558	56.86
NNCLR [Dwi+21]	3.913	52.20	3.643	55.99
ContIG (Raw-SNP)	3.541	<u>60.11</u>	3.414	64.81
ContIG (PGS)	3.521	59.23	<u>3.391</u>	<u>65.86</u>
ContIG (Burden)	3.540	59.74	3.393	65.41
ContIG (Inner RPB)	<u>3.511</u>	59.95	3.397	65.71
ContIG (Outer RPB)	**3.490**	**60.39**	**3.378**	**65.99**

Table 5.3: Data-efficient evaluation results by fine-tuning on subsets of UKB samples. All our ContIG models use the "H1" genetic encoder variant. **Bold** indicates the best result, <u>underlined</u> is second best. ↑ means higher is better, and ↓ lower is better.

supervised experimental scheme, in which we choose randomly 1% and 10% of the labels provided by UK Biobank (UKB) [Sud+15], and perform the downstream tasks of Cardiovascular Risk Factors prediction. Then, we evaluate using the same fixed test split of 20% of UKB dataset size. We choose this particular downstream task as UKB's dataset size is large enough to allow a simulation for expert annotation collection process, *i.e.* 1% of the overall labels is approximately 1000 samples, and such number may simulate an annotation process. The other benchmark datasets (APTOS [19], RFMiD [Pac+21], and PALM [Fu+19]) are relatively small in size. The evaluation results shown in Tab. 5.3 compare models trained with ContIG to models trained with the self-supervised baselines. ContIG outperforms the baselines in this evaluation scheme too. Note that all models are trained on the same exact subset of individuals and also evaluated on the same test set. The results for this data-efficient evaluation scheme especially confirm the advantages of pretraining with multiple genetic modalities using the "Outer" aggregation scheme. Notably, semi-supervised pretraining of ContIG with only 1% labeled data still outperforms the self-supervised baselines when they have 10× as much labeled data available.

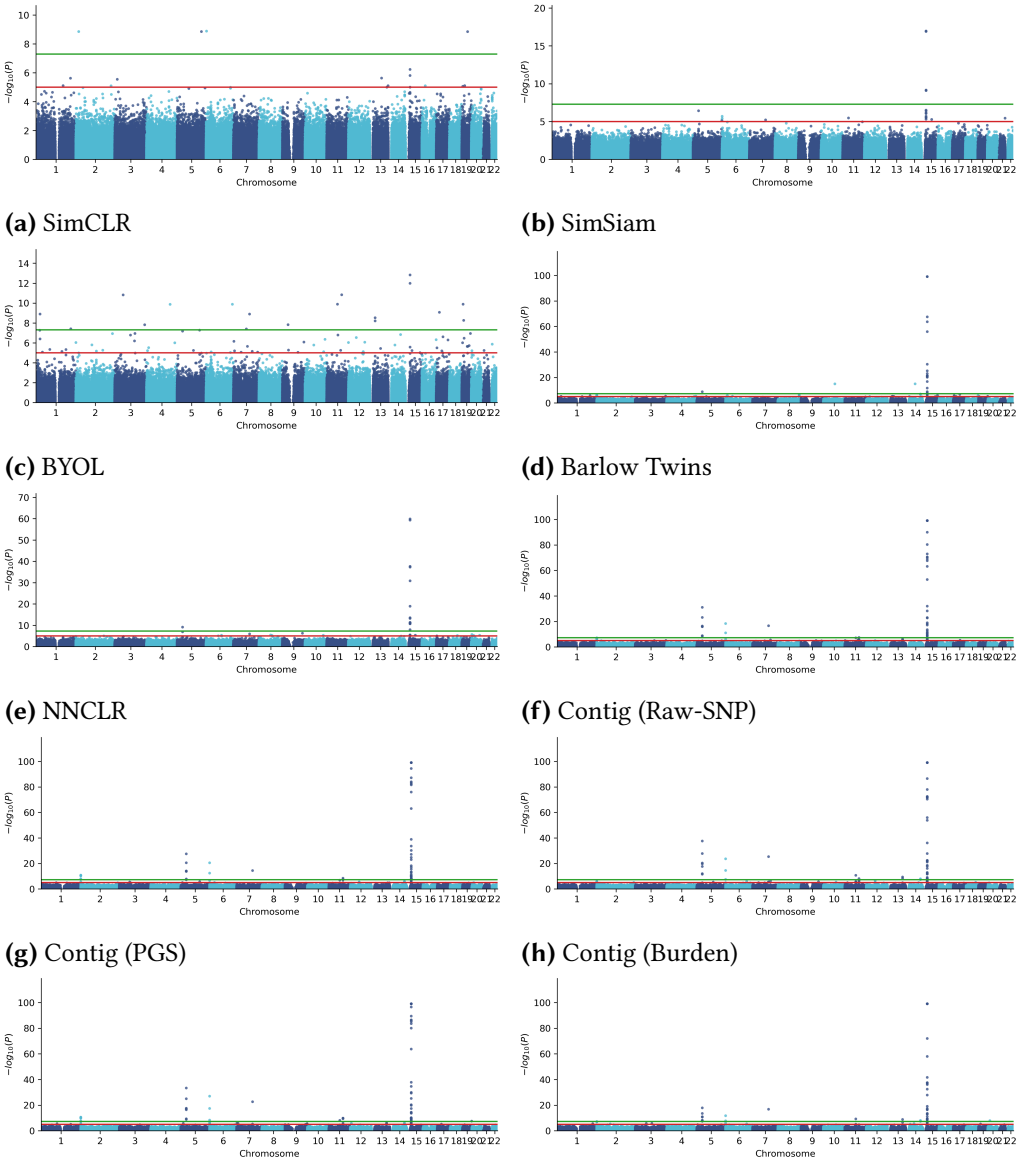
5.4.5 Genome-wide Association Study Results

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. GWAS

studies 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 [Ash+21; Kir+21], or high-dimensional representations of such data. Here, we follow the *transferGWAS* [Kir+21] 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 [Cha+15; Pur+07]. 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 complete analysis details to Appendix C.

We present the results of the performed GWAS study in Tab. 5.4 and Fig. 5.3. In Tab. 5.4, we report the number of found independent regions for each pretraining method. Genetic pretraining increases the statistical power of the genetic association study considerably. Only BYOL [Gri+20] achieves near-competitive results and all other self-supervised methods are outperformed by a large margin. We also looked up the found regions in the GWAS catalog [Insc] 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 [Kir+21], albeit with a larger sample size.

Fig. 5.3 shows the manhattan plot of genome-wide associations from the GWAS with ContIG and other pretraining methods. A number of very strong signals, *e.g.* on chromosomes 15 and 5, are known to be associated with skin pigmentation and cardiovascular traits. Manhattan plots for the other pretrained models look similar but with less signal. Almost all models found strong signals on chromosome 15. Interestingly, the manhattan plots for both SimCLR and BYOL (Fig. 5.3 (a) & Fig. 5.3 (c)) show clear signs of an ill-fitted association model. BYOL exhibits most likely spurious associations distributed over the whole genome but no signal in the chromosome-15 pigmentation region. This happens even after applying the inverse-normal transformation to counteract outliers and is likely due to different forms of confounding. This finding also explains the surprisingly large number of hits for BYOL – they are most likely false-positives. A more careful analysis with mixed effect models [Lip+11] and in-depth inspection of the image features is beyond the scope of this thesis.



(a) SimCLR **(b)** SimSiam

(c) BYOL **(d)** Barlow Twins

(e) NNCLR **(f)** Contig (Raw-SNP)

(g) Contig (PGS) **(h)** Contig (Burden)

(i) Contig (Inner RPB) **(j)** Contig (Outer RPB)

Figure 5.3: Manhattan plot for the GWAS with different methods. The x-axis shows the position of each SNP on the genome, the y-axis is the negative base-10 logarithm of the p-value for each SNP. Higher values correspond to lower p-values, correspond to stronger signal. The red line corresponds to a significance threshold of 0.05 Bonferroni-adjusted for the number of SNPs; the green line corresponds to "genome-wide significance" ($5 \cdot 10^{-8}$). P-values are clamped at 10^{-99} for clearer visualization (only relevant for the loci on chromosome 15 with a minimum p-value of 10^{-320}). Figure source [Tal+22a].

Model	Found Regions ↑
SimCLR [Che+20a]	4
SimSiam [CH21]	2
BYOL [Gri+20]	17
Barlow Twins [Zbo+21]	8
NNCLR [Dwi+21]	3
ContIG (Raw-SNP)	16
ContIG (PGS)	<u>20</u>
ContIG (Burden)	19
ContIG (Inner RPB)	**22**
ContIG (Outer RPB)	18

Table 5.4: GWAS results. Indicated is the number of independent regions associated with the image embeddings for each model

5.4.6 Genetic Feature Explanation Results

In this section, we inspect the representations learned by ContIG using the explanation methods developed in Sec. 5.3.3[6]. First, we analyze the models trained with a single genetic modality. Fig. 5.4 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 [Pop+18]. 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 [Gia+14; Sat+14].

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).

6 We validate that our explanation approach can in fact distinguish meaningful features from noise features in Appendix C

In addition to the global attributions, Fig. 5.5 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 [Che+11]. Analogously, for this instance there is a large number of PGSs very strongly related to cardiovascular 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.

5.4.7 Ablation Study

For the set of experiments reported in this section, we conduct ablations for the hyper-parameters of training batch size (b) and lambda (λ) from Eq. (5.1), used in the pretraining phase using our method ContIG. For the batch size, due to memory limits of available GPUs, 64 multimodal samples is the maximum we could fit, *i.e.* each sample in the batch contains an image with its corresponding genetic modalities. Nevertheless, ContIG outperforms state-of-the-art contrastive methods in the evaluated downstream tasks as shown in Tab. 5.1 and Tab. 5.2. In fact, this can be viewed as an advantage of training with ContIG, as it does not strictly require large batch sizes as opposed to SimCLR [Che+20a]. Therefore, for the purposes of this ablation study, we try smaller batch sizes of 16 and 32. As anticipated, we observe a slight drop in downstream performance ($\leq$ 2 p.p.) as Sec. 5.4.7 shows. For varying the values of lambda, for which we use the value 0.75 by default. In Sec. 5.4.7 we also evaluate the values of 0.25 and 0.5, and we find that the results are comparable to those obtained with original value of lambda ($\leq$ 1 p.p.). The comparable nature of the results when varying these hyperparameters – both affect the results with $\leq$ 2 p.p. –, somehow show that our method exhibits an improved robustness to smaller batch sizes and lambda values.

5.5 Discussion

In this Chapter, 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. While we believe this may be the reason why ContIG outperforms image-only self-supervised baseline methods –in some tasks by a margin–, but we

Batch (b)	Lambda (λ)	APTOS	RFMiD	PALM	Cardio. Risk Pred.	
		QwKappa ↑	ROC-AUC ↑	Dice ↑	MSE ↓	ROC-AUC ↑
64	0.75	**86.33**	**93.92**	**77.56**	3.180	72.65
64	0.5	84.13	93.52	77.32	**3.167**	**73.08**
64	0.25	84.91	93.77	76.64	3.174	72.37
32	0.75	84.01	93.41	76.59	3.182	72.11
16	0.75	84.09	92.77	76.40	3.296	67.41

Table 5.5: Ablation results for batch-size (b) and lambda (λ). Note that we vary one hyperparameters while fixing the other.

also conjecture that the reliance of self-supervised baselines 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 investigate how to improve the nature and types of employed augmentations in these algorithms in next Chapter.

We also attempt to assess the explainability of the learned representations by ContIG through performing a GWAS study and adapting attribution-based interpretability methods. Both aim to understand the relationship between imaging and genetic modalities in more detail and find interesting associations. This form of explainability by identifying imaging-genetic associations (or correspondences) was deemed a target (or a motivation) for our method. By cross-referencing the associations uncovered by ContIG with resources from literature, we find they are relevant. The significance of this result lies in that methods similar to ContIG may help discover novel genetic associations for traits visible in medical scans.

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, and only few studies are open to the public. 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.

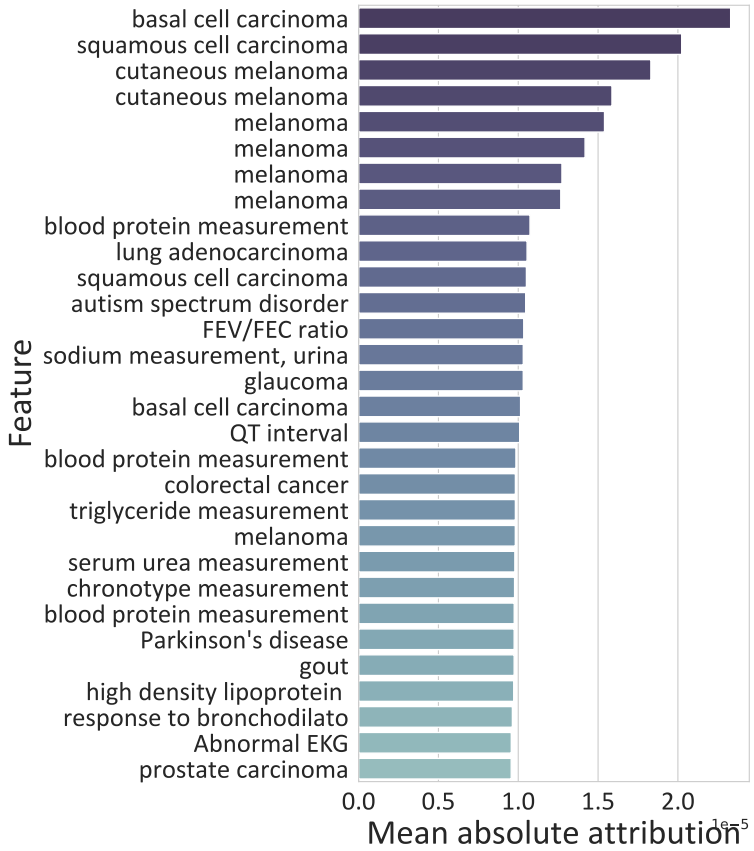


Figure 5.4: Global explanations for genetic features in ContIG (PGS only). Recorded is the mean absolute attribution per feature, aggregated over 1000 individuals, and the 30 PGS with highest associations are shown. Repeated traits (*e.g.* Melanoma) are due to multiple different risk scores published in the PGS catalog. Figure source [Tal+22a], reprinted with permission.

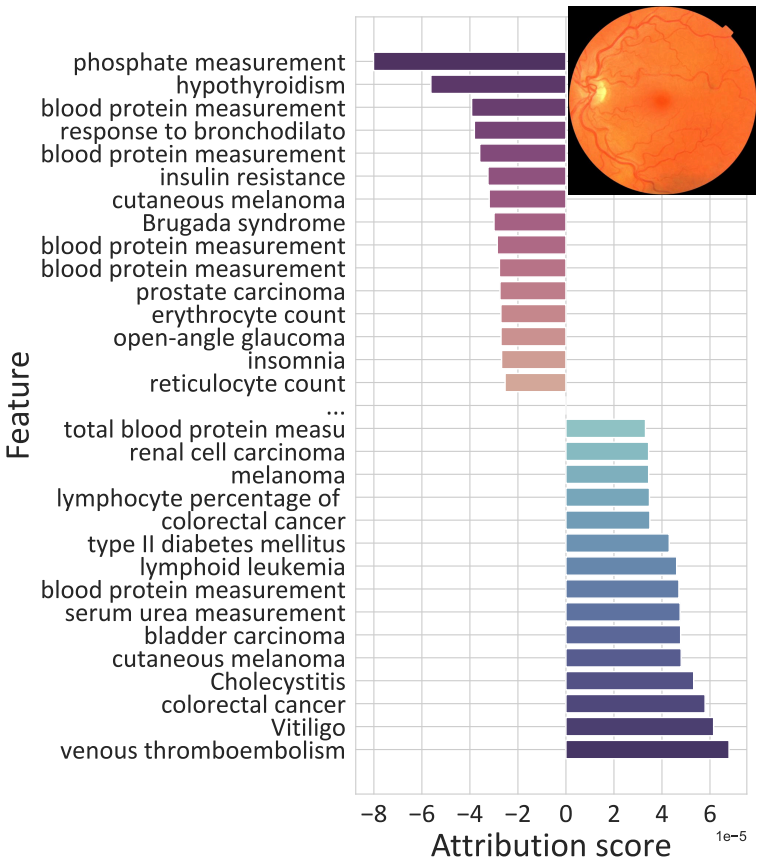


Figure 5.5: Local explanation attributions (signed) of genetic features for one image-PGS pair. Only the risk scores with highest values in either positive direction are shown. Retinal fundus image reproduced by kind permission of UK Biobank ©. Figure source [Tal+22a], reprinted with permission.

6 Self-supervision from Homogeneous Medical Scans

This chapter extends own work in [Tal+22b], published in the 12th volume of the Diagnostics Journal in 2022. Therefore, some figures, tables, and statements have been quoted verbatim or reproduced with permission.

6.1 Introduction

Medical scans normally exhibit more uniform appearance characteristics than natural images, such as more coherent color density distributions. This is consistent within each medical imaging modality (see Fig. 6.1) or even across different imaging modalities (see Fig. 2.7). Most medical scans use a single channel to capture colors, and therefore appear in levels of gray-scale. This homogeneous nature of medical scans may become an obstacle for several state-of-the-art contrastive learning methods, which rely on image transformations (or augmentations) to create image views, which may be for positive or negative samples, depending on the chosen loss. Ideally, these contrastive methods aim to learn data representations that are invariant to the chosen augmentations. However, most employed image augmentations assume natural imaging inputs, and are designed to operate on such domains – deeming many of these augmentations incapable to learn rich representations from medical images. As a result, to better adapt such self-supervised contrastive methods to such homogeneous medical scans, we propose several domain-inspired changes to the employed image augmentations in training in this Chapter. As a testbed for these proposed changes, we evaluate on dental X-Rays, namely Bitewing Radiograph (BWR) images, for the reasons described below.

Dental caries is the most prevalent health disease, affecting more than three billion people worldwide [Kas+17]. The estimated global annual total of conducted dental x-ray examinations is in the scale of 520 million [Cha01]. Nevertheless, almost all machine learning applications in the dentistry domain followed the supervised learning paradigm, e.g. [Kha+21; Kim+19a; Set+20]. Notably, these studies report diagnostic performances that are considered to be clinically useful. However, the referenced datasets used in these studies are only scratching the surface, with respect of being comprehensive or representative. For instance, [Kim+19a] used 12,179 labeled panoramic radiographs, whereas [Set+20] used only

Figure 6.1: Bitewing Radiograph (BWR) Examples. Collected in own study [Tal+22b].

20 CBCT volumes. This in turn hinders transferabilty and generalizability, and finally the dissemination of machine learning applications into clinical settings.

For diagnosing dental caries, the clinicians commonly analyze Bitewing Radiographs (BWRs), a specific medical imaging modality from the X-Ray family that typically capture the teeth from the jaw sides (see Fig. 6.1). Notably, the assessment of BWRs by dentists is associated with low sensitivity and shows considerable inter-examiner variation [STP15; Wal+21]. The growing quantities of dental data and the challenging nature of dental caries detection motivated employing deep learning techniques for this task [BA21; Can+20; Kim+19a; MK20]. Additionally, the high costs associated with labeling caries in BWRs, make this domain a pertinent test-bed for self-supervised representation learning algorithms. For instance, annotating the curated test set used in this Chapter (see Sec. 6.3.1) required 71 man-hours approximately. At the same rate, annotating the full training dataset would have required more than 7600 man-hours (approx. 950 work-days).

To evaluate our proposed augmentation changes to the chosen self-supervised learning methods, we assess both the diagnostic performance of the model when fine-tuned into dental caries classification tasks and the effects on label-efficiency. Our experimental results demonstrate the obtained gains, including improved caries classification performance (6 p.p. increase in sensitivity) and improved label-efficiency. In other words, the resulting models can be fine-tuned using fewer labels; using as few as 18 annotations can produce $\geq$45% sensitivity, which is comparable to human-level diagnostic performance.

6.2 Methods

6.2.1 Self-Supervised Learning Algorithms

In this section, we present the employed self-supervised algorithms in this Chapter used for pretraining on raw BWR images. Each algorithm results in an encoder model that can be fine-tuned on subsequent downstream tasks, here dental caries classification. Three algorithms were employed, which all aim to learn semantic representations that are invariant to the augmentations applied to input samples [Che+20a; Gri+20; Zbo+21]. While we mentioned these methods briefly in Sec. 2.3.2, we elaborate more on them in this Chapter. These approaches build upon the cross-view prediction framework introduced in [BH92], e.g. predicting random crops of the same image from each other. Such approaches solve the problem in the feature space, i.e. the representation of an image view should be predictive of another view. However, predicting in feature space directly can lead to collapsed representations, i.e. a trivial constant solution across views. Therefore, these algorithms differ in the techniques used to avoid such collapsed representations.

SimCLR

First proposed by [Che+20a], this method follows the Contrastive family of algorithms [Hen20; OLV18]. At the core of these algorithms is the Noise Contrastive Estimation (NCE) loss [GH10], which aims to maximize the mutual information between related signals, in contrast to other signals, in the embedding space. In order to circumvent the aforementioned collapsed representations problem, SimCLR reformulates the embedding prediction task into one of classification. To achieve that, it discriminates (classifies) artificially created "positives" and "negatives" from unlabeled samples, as illustrated in Fig. 6.2a. In this context, the terms "positive" and "negative" have no relation to manually acquired human labels whatsoever; here, they indicate views of the same image (positives) and of other images (negatives).

SimCLR learns semantic representations from unlabeled data as follows. The image dataset is processed in batches, where positive and negative samples are created from each batch. For each input image, a pair of positive images is created using image augmentations. The negatives are then the remaining images in the batch. All images are then processed by the encoder network, to produce a vector representation, i.e. embedding, for each image. We employ a CNN for the encoder architecture, but other architectures such as image-transformers are possible. During training, the encoder is replicated to process *pairs* of samples, constituting a Siamese architecture [Bro+93]. Each representation is then processed

by a small projection head, which is a non-linear multi-layer perceptron (MLP) with one hidden layer. Finally, the NCE loss computes the cosine similarity across all samples. This loss encourages the similarity between positive samples to grow larger (attracts their embeddings in the feature space), and the similarity to negative samples to become smaller (repels their embeddings in the feature space).

BYOL

BYOL [Gri+20], which is short for Bootstrap Your Own Latent, attempts to avoid the mechanism of negative mining (or sampling) used in SimCLR. The motivations for such avoidance are two folds. First, it may be computationally expensive, as the NCE loss may require a large number of negative samples to learn rich representations. SimCLR [Che+20a] addresses this by using larger batch sizes ($\geq$512). Second, the semantics of negative samples may require special treatment [Wu+17] to ensure they encourage learning rich representations. Therefore, BYOL introduces asymmetric parameter updates to the encoder architecture as an alternative for negative sampling. In other words, the two encoder models in the Siamese architecture, illustrated in Fig. 6.2b, do not have identical weights.

The Siamese architecture processes a pair of augmented views of each image, similarly to SimCLR. However, the architecture in BYOL is modified to be asymmetric as follows. The first *online* network is trained to predict the representations of the other *target* network. Here, the weights of the target network are an exponential moving average of the online network. This means that the actual parameter updates, i.e. gradients of the loss, are applied on the *online* network only. This is ensured by a "stop gradient" technique on the target network, which has been found, empirically, to be essential [CH21] to avoid collapsed representations. The overall training loss is the Mean Squared Error (MSE) between the predictions of online and target networks. Note that both networks use a projection head similar to SimCLR's. After training, only the encoder of the online network is kept, and everything else is discarded.

Barlow Twins

This method, which was first proposed in [Zbo+21], and illustrated in Fig. 6.2c, avoids both negative sampling of SimCLR and the asymmetric updates of BYOL. It rather relies on a statistical principle called *redundancy reduction* to learn representations. The algorithm steps are the following. Assuming a batch of images, from which two sets of augmented versions are created by applying different augmentations. These versions are processed concurrently with a Siamese encoder,

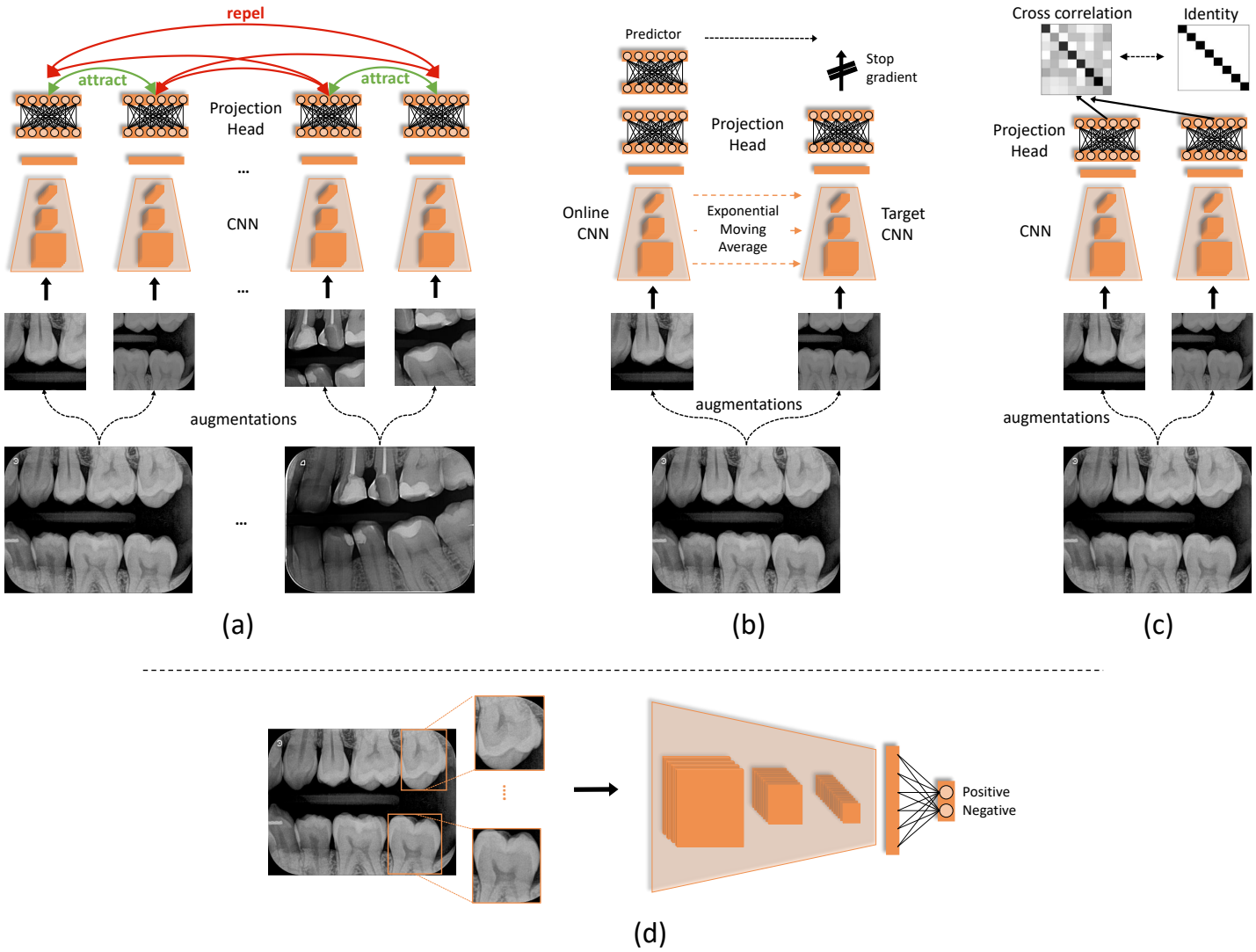


Figure 6.2: Illustration scheme of the three self-supervised algorithms and how to fine-tune the resulting encoder CNN. (**a**) SimCLR relies on **attracting** the views of each image (positives) together and **repelling** them from the views of other images (negatives). (**b**) In BYOL the target network calculates moving averages of the online network, which is updated with loss gradients. (**c**) Barlow Twins computes the cross-correlation matrix of two batches of image views and minimizes its difference to the identity matrix. (**d**) The obtained CNN encoder is fine-tuned on input tooth images for caries classification. Figure source [Tal+22b], reprinted with permission.

one set per encoder replica. Here, similar to SimCLR, the encoder weights are replicated, and the representations are projected with a projection head. Then, the cross-correlation matrix of the two sets of the resulting representations is computed. Each entry of this matrix encodes the correlation between the corresponding representation entries. Finally, the overall loss is defined as the difference between the computed cross-correlation matrix and the identity matrix. The intuition behind this is that it encourages the representations of positive image views to be similar, while minimizing the redundancy between their components. Explaining the name of the redundancy reduction principle. After training, one encoder replica is kept and utilized in subsequent downstream tasks, similar to SimCLR.

6.2.2 Modified Image Augmentations for Self-Supervision from Medical Scans

As explained earlier, many state-of-the-art self-supervised algorithms employ image augmentations in learning data representations that are invariant to the chosen augmentations. In particular, these augmentations are used in creating positive samples and also negative samples in contrastive methods. As shown previously in [Che+20a; Gri+20; Zbo+21], the choice of image augmentations considerably influences downstream performance. In our experiments, we find that the default augmentations used by these methods fail to learn semantically rich data representations from medical scans. In other words, the resulting representations do not improve downstream task performance, here in caries classification results. As mentioned earlier, we believe the nature of the data has a role in this, i.e. medical images exhibit a more uniform nature than natural images, e.g. color distributions. Hence, we employ a modified set of image augmentations, which better fits the more homogeneous medical imaging domain.

As Tab. 6.1 shows, the employed image augmentations by [Che+20a; Gri+20; Zbo+21] are stronger than the ones we use. As confirmed by [Che+20a], such strong augmentation regimes were essential to learn representations from natural images. On the contrary, we find that less aggressive image transformations can learn better data representations from medical images. In particular, the reduced probabilities of color adjustments benefit the learned representations the most in our evaluations. In fact, we find that applying the Gaussian Blur augmentation is detrimental to the learned representations in our case. Note that some of the augmentations in the original set are not applicable in our case, since they assume colored image inputs, which is not our case. Finally, different from the original

Table 6.1: Comparing the employed set of image augmentations with the original one used by [Che+20a; Gri+20; Zbo+21]

Image Augmentation	Modified Set	Original Set
Random resized cropping (% of input size)	50–100%	8–100%
Random horizontal flip (probability)	50%	50%
Random rotation (angle)	-20° – 20°	–
Image Brightness (probability)	20%	80%
Image Contrast (probability)	10%	80%
Image Saturation (probability)	10%	80%
Image Hue (probability)	–	20%
Image Color to Grayscale (probability)	–	20%
Gaussian Blur	–	50%

set of augmentations, we apply image rotation with small angles, which we find beneficial in our case.

6.3 Experimental Results

In this section, we report the evaluation results of the supervised caries classification on tooth segments. Notably, for training, i.e. fine-tuning pretrained models, we use EHR labels but all of the reported metrics are computed on the curated test set of 343 BWRs. We evaluate whether fine-tuning pretrained models via self-supervision improves the diagnostic performance of the classifier compared to a baseline model that was initialized with random model weights. In addition, we assess if self-supervised pretraining improves the label-efficiency by successively increasing the size of labelled data for training the model. Both Secs. 6.3.1 and 6.3.2 have been quoted verbatim from own work [Tal+22b], and only slightly modified.

6.3.1 Dataset

The dataset was collected by three dental clinics in Brazil, which are specialized in radiographic and tomographic examinations. The dataset consists of 38,094 BWRs taken between 2018 and 2021. In total, 9779 patients with an average [min–max, sd] age of 34 [3–88, 14] years constitute the sample. The average [min–max, sd] number of scans per patient is 4 [1–11, 1]. We preprocess the radiographs by extracting individual tooth images using a helper model, a deep-learning based tooth instance-segmentation model (unpublished). Each detected tooth is then cropped from the

BWR using a bounding box that fully contains the tooth. The procedure resulted in a dataset of 315,786 cropped tooth images. Out of those, we observe 49.9% of molars, 40.5% of premolars and 9.6% of canines and incisors, respectively. It is noteworthy that the tooth classification with the helper model may not be perfect, but as we are interested in the tooth as an object, we ignore these imperfections in automated tooth labeling. Tooth-level caries labels were extracted from electronic health records (EHRs) that summarize the patient's dental status. The dataset has a caries prevalence of 19.8%. Although EHR-based ground truth labels are known to come with uncertainties and biases [Gia+18], we find that they provide sufficiently rich signals (semantically) when fine-tuning self-supervised models. For model evaluation purposes, a hold-out test set was curated by dental professionals. The test set consists of a random sample of 343 BWRs. The average [min-max, sd] age of the patients within the test set is 33 [5–80, 13]. The BWR samples were annotated for dental caries by four independent dentists. These annotations were reviewed by a senior dentist (+13 years of experience) to resolve conflicts and establish the ground truth in the test set. After extracting tooth-level images with the helper model, the test set contains 2846 tooth samples with 29.9% caries prevalence (850 positive and 1996 negative). We observe 49.2% molars, 40.5% premolars, and 10.3% canines and incisors, respectively. We ensure that the set of patients is independent in the training and the test datasets.

6.3.2 Implementation Details

All images were resized to the resolution of 384×384 pixels. We employ the Resnet-18 [He+16] architecture as the neural network encoder. During the self-supervision stage only, the used projection head has an output dimension of 128. For all training procedures, we employ the Adam optimizer [KB14a]. During the self-supervised pretraining stage we train with batch sizes of 224 images and set the initial learning rate to 0.001, while using cosine annealing [LH17]. After the self-supervised pretraining stage, the resulting encoder is employed in supervised dental caries classification, as illustrated in Fig. 6.2d. To that end, a fully-connected layer with output units equal to the number of classes is added on top. In this stage we train with a batch size of 92 images, set a fixed learning rate of 0.0001 and use the cross-entropy loss to learn from the EHRs labels. We do not tune the classification threshold and we use a confidence score of 0.5 to discriminate between the positive (has caries) and the negative prediction label. As evaluation metrics we compute ROC-AUC, sensitivity, and specificity. Our implementations are in `Python`, using the libraries `PyTorch v1.10.0`, `Pytorch-Lightning v1.5.4`,

and `Lightly` [lig]. We ensure reproducibility of results by setting a unified random seed of 42 for all scripts and workers.

6.3.3 Transfer Learning (Fine-Tuning) Results

In this set of results, we report on the performance of models that were fine-tuned on the full image dataset (315,786 tooth segments extracted from 38,094 BWRs) using the automatically-generated EHR labels. The models are initialized at the beginning of the fine-tuning phase, with model weights obtained by the self-supervised methods SimCLR, BYOL and Barlow Twins, described in Sec. 6.2.1. In terms of baselines, we compare to:

- A model with identical architecture (Resnet-18) trained from scratch, i.e. whose weights are initialized randomly.
- The set of models pretrained with the same three algorithms but with the original set of image augmentations from [Che+20a].

The evaluation results for this set of experiments are shown in Tab. 6.2.

The highest sensitivity, with 57.9% was observed with Barlow Twins, followed by SimCLR and BYOL, with 57.2% and 54.6%, respectively. All using our models pretrained with our modified augmentation types. These values are considerably higher than 51.8%, obtained by the baseline model trained from scratch. They also outperform the sensitivity results obtained by the models trained using the original set of augmentations proposed in [Che+20a]. For specificity all models perform similarly. With respect to the ROC-AUC values, all pretrained models with our modified augmentations are close to each other (73.3%, 73% and 73.4% for SimCLR, BYOL and Barlow Twins, respectively) but consistently higher than the baselines.

Method	Sensitivity	Specificity	ROC-AUC
Baseline (from scratch)	51.80	**91.30**	71.50
SimCLR (original augmentations)	52.29	89.14	72.09
BYOL (original augmentations)	52.72	90.52	71.89
Barlow Twins (original augmentations)	53.51	88.71	72.48
SimCLR (new augmentations)	57.20	89.30	73.30
BYOL (new augmentations)	54.60	**91.30**	73.00
Barlow Twins (new augmentations)	**57.90**	88.90	**73.40**

Table 6.2: Caries classification results when fine-tuning on the full caries classification training set. We highlight in **bold** the best models.

6.3.4 Data-Efficiency Results

The results in this section demonstrate the obtained gains in data efficiency. For that purpose, we report on the performance of caries classification for varying fine-tuning dataset sizes up to 10% of the total dataset, which is almost ~3.8K BWRs or 30K tooth segments. In detail, the subset sizes we considered are as follows in terms of No. Teeth/ No. BWRs: {152/18, 305/37, 1.5 K/190, 3 K/380, 15 K/1.9 K, 30 K/3.8 K}. Fine-tuning in all experiments is done for a fixed number of epochs (50 epochs each). For each subset, we compare the performance of the fine-tuned models to the following baselines:

- The model with an identical architecture (Resnet-18) trained from scratch, i.e. whose weights are initialized randomly.
- The set of models pretrained with the same three algorithms but with the original set of image augmentations from [Che+20a].

We repeat this process for each subset five iterations to account for random sampling effects, i.e. the samples at each iteration are chosen randomly, resulting in ~20% caries prevalence, which is close to the actual prevalence of the full dataset of 19.8%.

As shown in Tab. 6.3, the models pretrained with self-supervised algorithms accompanied by our modified augmentations outperform the baselines in terms of sensitivity. Interestingly, the model pretrained with the Barlow Twins and our proposed augmentations obtains a sensitivity value of 46.28% even when fine-tuned with only 18 BWRs.

Method | **#Teeth/#BWRs**	**152/18**	**305/37**	**1.5K/190**	**3K/380**	**15K/1.9K**	**30K/3.8K**
Baseline (from scratch)	32.87	41.74	42.45	46.61	44.78	50.19
SimCLR (original augmentations)	33.21	42.76	44.87	46.77	46.86	50.55
BYOL (original augmentations)	34.19	43.62	45.18	47.65	48.56	49.25
Barlow Twins (original augmentations)	34.68	44.42	45.41	47.01	47.44	51.22
SimCLR (new augmentations)	40.02	**50.05**	46.40	52.99	48.96	**54.80**
BYOL (new augmentations)	44.78	48.35	**60.92**	**53.88**	**55.32**	51.18
Barlow Twins (new augmentations)	**46.74**	45.01	51.21	51.28	53.42	52.85

Table 6.3: Caries classification results when fine-tuning on the full training set. We highlight in **bold** the best models. The results are reported in terms of Sensitivity.

6.4 Discussion

In this chapter, we assess the effect of self-supervised pretraining on a real-world supervised learning task, by training caries prediction models on EHR-based labelled data and evaluating them on a test set with manually acquired ground-truth labels. We show that for the downstream task of caries classification, pretraining with self-supervised algorithms provides a considerable performance boost, especially for the sensitivity of the model. All three presented methods outperformed the baselines, and added up to 6% in sensitivity (see Tab. 6.2). The effectiveness of leveraging pretrained models in a transfer learning scheme for boosting the performance of prediction tasks is a popular practice in caries classification tasks [Far+20; Zhu+20b]. In this work, however, we show that pretraining can be effectively done on domain-specific image data via self-supervision and does not have to stem from large open (non-medical) image datasets [Den+09; Kuz+20; Lin+14].

This chapter also highlights that self-supervised models outperform baseline models trained from scratch with a significant margin when using few training samples, particularly in terms of sensitivity. A gain that is usually called label-efficiency. In a low-data regime only as few as 18 BWR samples (=152 tooth images) yields $\geq$ 45% sensitivity, which is competitive compared to the diagnostic performance of domain experts, who reportedly show sensitivities of around 47% (95% confidence interval (CI) 40% to 53%) [Wal+21]. Hence, by using self-supervision techniques, the annotation process and data-efficiency may be improved as only a fraction of labeled data is required to achieve competitive results.

Lastly, in terms of methodological aspects, this chapter illustrates the need to choose appropriate image augmentation types when pretraining with self-supervised (contrastive) learning methods. Our empirical results in Tabs. 6.2 and 6.3 support choosing domain-inspired augmentation techniques, rather than relying simply on default augmentations used in these methods. In fact, the default augmentation types used by existing contrastive methods are mostly based on optimizing the results on natural imaging benchmarks, such as ImageNet [Den+09].

The results presented in this chapter come with a number of strengths and weaknesses. First, the training dataset contains more than 30 K BWRs and their labels (used for the fine-tuning stage) are based on EHR texts. EHR data is fairly abundant in dentistry. However, EHR-based labels are associated with uncertainties and biases [Gia+18]. In other words, the usage of labels extracted from EHRs may be problematic, as they stem from routine care and are affected by biases, incompleteness, inconsistencies, and limited accuracy. On the other hand, this is "real" data as it is stored in large amounts in data silos all over the world. Therefore, making use of this treasure trove of data is an advantage, despite the biases and

uncertainties associated with it. Second, the caries classification model is trained on the tooth level, even though the raw image data in the used dataset is in Bitewing (BWR) format, i.e. each image contains multiple teeth. As a downside, the step to preprocess these images to extract teeth samples requires an understanding of the tooth as an object on the BWR. To this end, we use a helper pretrained object detection model to crop tooth segments from BWRs.

7 Conclusions & Outlook

This thesis investigates how to exploit the ever growing amounts of unlabeled medical images in learning semantic data representations, a feat that aims to mitigate the costs of expert annotation required. To that end, we propose to use the self-supervised learning scheme to harness unlabeled samples for representation learning, followed by an annotation-efficient downstream task solving stage. Nevertheless, we identify multiple unique and inherent characteristics of medical images that deem representation learning from medical imaging with existing self-supervised solutions more challenging. In particular, we utilize the characteristics of multimodality (in chapter 3), multi-dimensionality (in chapter 4) and homogeneous density distributions (in chapter 6) in medical images. Also, we fuse knowledge about disease patterns from medical scans with genomic data modalities (in chapter 5) in an attempt to create a more holistic view on human disease. All of our proposed self-supervised methods in this thesis aim to address the challenges associated with learning from medical scans. In general, the experimental results obtained when pretraining with our proposed approaches exhibit both improved downstream task performance and annotation-efficiency. In other words, initializing deep neural network models with features learned with our methods significantly reduces the quantities of required annotations.

7.1 Findings and Limitations

We briefly summarise the findings and limitations of each chapter in this thesis.

chapter 3 proposed a multimodal Jigsaw puzzle-solving task, which exploits multiple medical imaging modalities, e.g. MRI and CT, to learn data representations in a self-supervised manner, even when these modalities are unregistered. The proposed multimodal puzzles outperform their single-modal counterparts, and are also more computationally efficient, thanks to the Sinkhorn operator. As part of this framework, a cross-modal conversion (generation) method is employed, which addresses real-world imaging modality imbalance issues. While the empirical results are consistent with the gains mentioned above, the main limitation lies in that it assumes multimodal imaging inputs. However, one can circumvent this limitation by utilizing image augmentation techniques to create other image views [TKI20],

and we show that using a cross-modal generation step with CycleGAN is able to mitigate this shortage of multimodal images.

In chapter 4, we show that designing self-supervised tasks that operate on the 3D spatial context proves more effective than the sub-optimal 2D context for learning representations from unlabeled 3D images. Furthermore, we observe performance gains when pretraining models on a large unlabeled medical imaging corpus different from smaller downstream datasets, suggesting alternatives for transfer learning from ImageNet features. However, a key limitation for 3D deep learning is the increased computational and memory requirements, prompting the search for computationally-efficient network architectures. We mention a few examples next section. The alternative being utilizing more advanced hardware (GPU) with larger memory capacities.

chapter 5 presented ContIG, a self-supervised contrastive representation learning algorithm for imaging-genetics datasets. We show the benefits of including genomic modalities in conjunction with medical images in the self-supervised pretraining stage, by evaluating on a variety of tasks relevant for clinical practice and genetic research. The latter would benefit the most from the proposed explainability mechanism for learned representations by ContIG through GWAS studies and attribution-based interpretability methods. This form of explainability by identifying imaging-genetic associations was in fact a target for ContIG. Here, the main limitation lies in the requirment of datasets with imaging and their corresponding genetic data for pretraining with ContIG. Nevertheless, an increased number of imaging-genetics studies on humans are being conducted, and also in live-stock and plant breeding. We should mention here that even though our evaluation experiments focused on combining medical images with genetics in chapter 5, we ensure ContIG is able to combine medical images with other types of modalities, such as markers of blood samples, or even phenotypic descriptions, *e.g.* a person's medical history. This is made possible in the method design by making minimal assumptions about input data modalities.

In chapter 6 we illustrate the effect of self-supervised methods on a downstream task from the field of dentistry, where medical imaging plays a vital role in clinical practice. Namely, we evaluate on caries prediction (classification) models. In this context, we find that image augmentation techniques employed by existing self-supervised algorithms do not learn rich representations from medical scans. Hence, we show how domain knowledge can be used to adapt the types of augmentation techniques. The results also highlight that fine-tuning self-supervised models with noisy and unreliable EHR-based labels is possible. This can be viewed as both a strength, as it further reduces manual annotation efforts, but also is a

limitation, as labels extracted from EHRs may be affected by biases, incompleteness, inconsistencies, and limited accuracies.

7.2 Applications and Future Work

This work has an array of practical applications.

First and foremost, the continuously growing numbers of medical scans create a natural workload on human radiologists, justifying the increasing adoption of computer-aided diagnosis (CAD) and detection systems [Doi07]. These systems, in particular, can benefit the most from the self-supervised methods proposed in this thesis. Mitigating the huge requirements on manual expert annotation is the key target of the methods developed in this work, and is a fundamental step to improve the adoption of deep learning methods in CAD systems and harness their human-level performances in many downstream tasks. With the continuous digitization of medical images, the hope that physicians and radiologists are able to instantly analyze them with machine learning algorithms is slowly shaping as a reality. Medical imaging allows instant insights into human body organs, thus the growing attention from both machine learning and medical communities.

Second, utilizing the aspect of multimodal images in self-supervised pretraining has a direct impact in the medical imaging domain, as shown in chapter 3, since medical imaging modalities capture complementary aspects of organs and tissues. Certain disease phenotypes are only visible in specific types of imaging modalities, such as Brain Tumor in T2 MRI, and thus training models to combine knowledge from several medical imaging modalities improves their downstream performance. Nevertheless, the aspect of multimodality in imaging and related sensory data has applications in robotics and autonomous driving. Most autonomous vehicles exploit a variety of sensors [RBZ22], including color with vision cameras, depth or thermal cameras, LiDARs, and RADARs.

Third, due to technological advancements in 3D sensing, and the growing number of its applications, e.g. in Robotics, CAD imaging, Geology, and Medical Imaging, the attention to 3D deep learning has been growing rapidly in the past few years [Ioa+17]. Our 3D self-supervised algorithms proposed in chapter 4 can improve annotation-efficiency of 3D deep learning across its application domains. Here, however, relying simply on existing architectures, e.g. convolutional neural networks, may incur large computational costs. Therefore, a theme we deem critical is the search for more computationally-efficient architecture alternatives, such as Shift blocks[Wu+18a], Squeeze-and-excitation blocks [HSS18], and Fire modules [Ian+16]. Another approach to reduce the model complexity is by uti-

lizing model compression techniques, such as by pruning [HMD15; Ian+16] and sparsification [FC18].

Finally, pretraining with our method ContIG presented in chapter 5 has clear applications in domains where imaging-genetic associations are important. This includes human healthcare, as the majority of human diseases have genetic roots. Understanding the role of genetics in how disease traits evolve in living organisms can help in early detection or even prevention. In addition, applications of live-stock and plant breeding can considerably benefit from the discovered associations by ContIG, when trained on relevant corpora. After all, breeding plants found to be economically or aesthetically desirable can only be done with a comprehensive understanding of their genetic predisposition [Yan+20]. Nevertheless, as mentioned above, we ensure ContIG is able to process input data modality types other than genetics in conjunction with images, such as biomarkers or medical history, a direction we deem essential in the future of machine learning in patient healthcare.

This thesis presents a framework, or a blueprint, for how to exploit raw unlabeled data samples to reduce human annotation required. By viewing all the methods as utilizing complementary aspects of input data, one can integrate knowledge from multiple angles. After all, patient data is in fact multi-modal; a single patient can have several medical imaging modalities, some of which are in 3D, and they may have genetic samples and clinical data as well. Transitioning to the patient-level by combining knowledge from all existing modalities is the cornerstone of personalized medicine, which may only become a reality by exploring methods to create such holistic views on patients.

A Experimental Details for SSL with Multimodal Jigsaw Puzzles

A.1 Model Training for all tasks

Input preprocessing. For all input scans, we perform the following pre-processing steps:

- First, we create 2-dimensional slices by navigating the scans from all datasets over the axial axis (z-axis).
- We resize each slice to a resolution of 128×128 for samples from BraTS, 256×256 for both Prostate and CHAOS.
- Then, each slice's intensity values are normalized by scaling them to the range $[0, 1]$.

Training details. For all tasks, we use Adam [KB14b] optimizer to train our models. The initial learning rate we use is 0.001 in puzzle solving tasks, 0.0002 in cross-modal generation tasks, and 0.00001 for segmentation and regression tasks. The network weights are initialized from a Gaussian distribution of $\mathcal{N}(\mu = 0.1, \sigma = 0.001)$ in puzzle solving and segmentation tasks, and from the distribution $\mathcal{N}(\mu = 0, \sigma = 0.02)$ in the cross-modal generation task. An L_2 regularizer with a regularization constant $\lambda = 0.1$ is imposed on the network weights in puzzle solving and downstream tasks. In terms of training epochs, we train all the puzzle solving tasks for 500 epochs, the cross-modal generators for 200 epochs, and all fine-tuning on downstream tasks for 50 epochs.

Network architectures. All of our network architectures are convolutional, and they vary in small details per task:

- For jigsaw puzzle solving tasks: we use 5 convolutional layers, followed by one fully-connected layer and one Sinkhorn layer.
- For downstream segmentation tasks: we use a U-Net [RFB15] based architecture, with 5 layers in the encoder, and 5 layers in the decoder. When fine-tuning, the weights of the encoder layers are copied from a pretrained model. The decoder layers, on the other hand, are randomly initialized. In terms of training losses in these tasks, we utilize a combination of two losses:

i) weighted cross-entropy, ii) dice loss. We use the same importance to both losses in the total loss formula.

- For cross-modal generation tasks: as mentioned earlier, we largely follow the architecture of the CycleGAN [Zhu+17] model. For the *generators*, we use the Johnson *et al.*'s [JAF16] architecture. We use 6 residual blocks for 128×128 training images, and 9 residual blocks for 256×256 or higher-resolution training images. With regards to the network *discriminators*, we utilize the PatchGAN [LW16] discriminator architecture, which processes 70×70 input patches.

Processing multi-modal inputs. In Brain tumor and Prostate segmentation tasks, the reported methods from literature use all available modalities when performing the segmentation, e.g. in table 3.1 in our paper. They typically stack these modalities in the form of image color channels, similar to RGB channels. However, our proposed puzzle-solving method expects a single channel input at test time, i.e. one slice with multi-modal patches. This difference only affects the input layer of the pretrained network, as fine-tuning on an incompatible number of input channels causes this process of fine-tuning to fail. We resolve this issue by duplicating (copying) the weights of *only* the pretrained input layer. This minor modification only adds a few additional parameters in the input layer of the fine-tuned model, but allows us to leverage its weights. The other alternative for this solution is to discard the weights of this input layer, and initialize the rest of the model layers from pretrained models normally. However, our solution for this issue takes advantage of any useful information encoded in these weights, allowing the model to fuse data from all the channels. The exact numbers of channels in each downstream task is as follows:

- BraTS Brain Tumor Segmentation: in each input slice, the MRI 4 modalities are stacked as channels.
- BraTS Number of Survival Days Prediction: for each input slice we also stack the 4 MRI modalities, on top of the predicted tumor segmentation mask; summing up to 5 channels for each input slice. The predicted masks are produced by our best segmentation model.
- Prostate segmentation: we stack the 2 available MRI modalities in each input slice.

In the Liver segmentation task, however, stacking the input modalities as channels is not possible. This is due to the fact that the modalities in this task (CT and

MR-T2) are non-registered. Hence, we process these modalities using the Joint Learning scheme used in [Val+18]. This means we process each modality as a single input slice. Pretrained models using our multimodal puzzles, learns to disregard the modality type during training and testing.

Training the multimodal puzzle solver It is noteworthy that after we sample patches from input slices, we add a random jitter of 5 pixels in each side before using them in constructing puzzles. This mechanism ensures the model does not use any shortcuts in solving the puzzles, thus enforcing it to work harder and learn better representations.

Algorithm 2 provides the detailed steps of the training process of our proposed multimodal puzzle solver. After obtaining the network parameters, the yielded representations capture different tissue structures across the given modalities as a consequence of the multimodal puzzle solving. Therefore, they can be employed in downstream tasks by simply fine-tuning them on target domains.

Algorithm 2: One epoch of training multimodal puzzle solver

Algorithm *Train Puzzle Solver*
  Input: list of multimodal *puzzles*
  Output: trained model G
  $G \leftarrow$ initialize model weights w
  foreach P *from puzzles* **do** `// each puzzle contains` N `patches`
    foreach *patch* x *in* P **do**
      $v \leftarrow G(x)$ `//` N`-dimensional feature vector`
    $V \leftarrow$ concat. vectors v `// form a matrix with size` $N \times N$
    $S \leftarrow Sinkhorn(V)$ `// permutation matrix`
    $P_{rec} \leftarrow S^T.P$ `// reconstructed version`
    $loss \leftarrow MSE(P^*, P_{rec})$

B Experimental Details for SSL from 3D Images

B.1 Implementation and training details for all tasks

It is noteworthy that our attached implementations are flexible enough to allow for evaluating several types of network architectures for encoders, decoders, and classifiers. We also provide implementations for multiple losses, augmentation techniques, and evaluation metrics. More information can be found in the `README.md` file in our attached code-base. We rely on `tensorflow v2.1` [ten20] with `Keras API` in our implementations. Below, we provide the training details we used in implementing our 3D self-supervised tasks (and their 2D counterparts), and when fine-tuning them in subsequent downstream tasks.

Architecture details. For all 3D encoders g_{enc}, which are pretrained with our 3D self-supervised tasks and later fine-tuned on 3D segmentation tasks, we use a 3D U-Net [RFB15]-based encoder (the downward path), which consists of five levels of residual convolutional blocks. The numbers of filters in these blocks are 32, 64, 128, 256, 512, respectively. The U-Net decoder (the upward path) is added in the downstream fine-tuning stage, and it includes five levels of deconvolutional blocks with skip connections from the U-Net encoder blocks. For the 2D encoders, we use a standard Densenet-121 [Hua+17] architecture, which is fine-tuned later on 2D classification tasks. When training our 3D self-supervised tasks, we follow [Che+20a] in adding nonlinear transformations (a hidden layer with ReLU activation) before the final classification layers. These classification layers are removed when fine-tuning the resulting encoders g_{enc} in downstream tasks.

Optimization details. In all self-supervised and downstream tasks, we use Adam [KB14b] optimizer to train the models. The initial learning rate we use is 0.001 in 3D self-supervised tasks, 0.00001 in 3D segmentation tasks, 0.0005 in 2D self-supervised tasks, and 0.00005 in 2D classification tasks. When we fine-tune our pretrained encoders in subsequent downstream tasks, we follow a warm-up procedure inspired from [KZB19] by keeping the encoder weights frozen for a number of initial warm-up epochs while the network decoders or classifiers are

trained. These warm-up epochs are 5 in 2D classification tasks, and 25 epochs in 3D segmentation tasks. The alternative options we evaluated were: 1) fine-tuning the encoder directly with a randomly initialized decoder, 2) keeping the encoder frozen throughout the training procedure. And the 3rd option we followed in the end was the hybrid approach of warm-up epochs described above, as it provided a performance boost over the other alternatives. For segmentation tasks, in particular, where a decoder is used in the architecture, these warm-up epochs prove indispensable. Otherwise, training the whole model with a randomly initialized decoder, while the encoder is not frozen, may harm the encoder representations.

Input preprocessing. For all input scans, we perform the following preprocessing steps:

- In self-supervised pretraining using 3D scans, we find the boundaries of the brain or the pancreas along each axis, and then we crop the remaining empty parts from the scan. This step reduces the amount of empty background voxels, as they might confuse patch-based self-supervised methods with no additional semantic information. This step is not performed when fine-tuning on 3D downstream tasks.

- Then, we resize each 3D image from BraTS or Pancreas to a unified resolution of $128 \times 128 \times 128$, and to the resolution 224×224 for 2D images from Diabetic Retinopathy.

- Then, each image's intensity values are normalized by scaling them to the range $[0, 1]$.

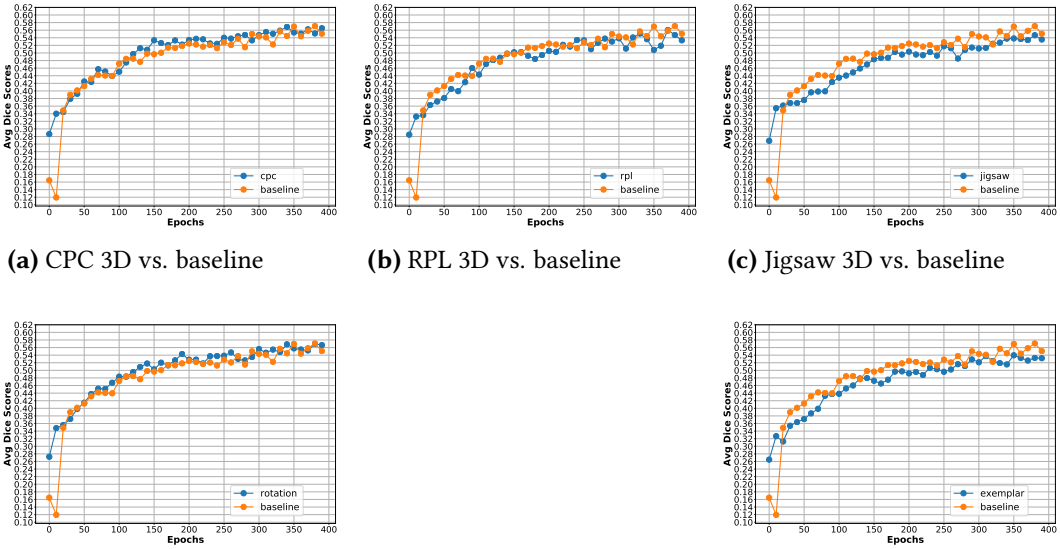
Processing multimodal inputs. In the first downstream task of brain tumor segmentation with 3D multimodal MRI, we pretrain using the UK Biobank [Sud+15] corpus, as mentioned earlier. Brain scans obtained from UKB contain 2 MRI modalities (T1 and T2-Flair), which are co-registered. This allows us to stack these 2 modalities as color channels in each input sample, similar to RGB channels. This form of early fusion [SWS05] of MRI modalities is common when they are registered, and is a practical solution for combining all information that exist in these modalities. However, as mentioned earlier, we use the BraTS [Bak+17; Men+15] dataset for fine-tuning, and each scan consists of 4 different MRI modalities, as opposed to only 2 in UKB that is used for pretraining. This difference only affects the input layer of the pretrained encoder, as fine-tuning on an incompatible number of input channels causes this process of fine-tuning to fail. We resolve this issue by

duplicating (copying) the weights of *only* the pretrained input layer. This minor modification only adds a few additional parameters to the input layer, but allows us to leverage its weights. The other alternative for this solution would have been to discard the weights of this input layer, and initialize the rest of the model layers from pretrained models normally. But we believe our solution for this issue takes advantage of any useful information encoded in these weights. This multimodal inputs problem does not occur in the other downstream tasks, as the inputs include only one modality/channel.

Task specific training details.

- **3D-CPC**: we follow [OLV18] in using an autoregressive network using GRUs [Cho+14] for the context network g_{cxt}, however, masked convolutions can be a valid alternative [Oor+16].
- **3D-CPC and 3D-Exe**: we use latent representation code size of 1024 in these tasks.
- **3D-Jig and 3D-RPL**: We split the input 3D images into $3 \times 3 \times 3$ patches in this task. We apply a random jitter of 3 pixels per side (axis).
- **Patch-based tasks (3D-CPC, 3D-RPL, 3D-Jig)**: each extracted patch is represented using an embedding vector of size 64.
- **3D-Exe**: the α value used for the triplet loss is 1.0.
- **3D-Jig**: the complexity of the Jigsaw puzzle solving task relies on the number of permutations used in generating the puzzles, i.e. the more permutations used, the harder the task is to solve. We follow the Hamming distance-based algorithm from [NF16] in sampling the permutations for this task. However, in our 3D puzzles task, we sample permutations that are more complex with 27 different entries. This algorithm works as follows: we sample a subset of 1000 permutations which are selected based on their Hamming distance, i.e., the number of different tile locations between 2 permutations. When generating permutations, we ensure that the average Hamming distance across permutations is kept as high as possible. This results in a set of permutations (classes) that are as far as possible from each other.

Augmentation in Exemplar. As mentioned earlier, we apply the following 3D transformations in Exemplar: random flipping along an arbitrary axis, random



(a) CPC 3D vs. baseline **(b)** RPL 3D vs. baseline **(c)** Jigsaw 3D vs. baseline

(d) Rotation 3D vs. baseline **(e)** Exemplar 3D vs. baseline

Figure B.1: Pancreas segmentation: Detailed speed of convergence results per method (blue) vs. the supervised baseline (orange). This benefit of our methods helps achieve high results using only few epochs. Figure Source [Tal+20], reprinted with permission.

rotation along an arbitrary axis, random brightness and contrast, and random zooming. These augmentations are utilized to produce the positive samples. We vary the percentages of applying these augmentations using these factors: $\alpha = 0.5$ for random rotations, $\beta = 0.5$ for color distortions (brightness and contrast), and $\gamma = 0.2$ for random zooming. When trying to omit a certain augmentation from the list above, we observe a drop in downstream performance. This is consistent with the findings of [Che+20a]. However, performing such transformations for high percentages is time-consuming, hence the reduced rates to 50%. Conducting a more thorough analysis of what *types* of augmentations are desirable is a future work.

B.2 Detailed experimental results

Fig. B.1 shows additional results for the proposed 3D self-supervised tasks in terms of speed of convergence.

C Experimental Details for SSL from Imaging and Genetics

C.1 Training & Implementation Details

C.1.1 Datasets Preprocessing

UK Biobank Genetic Modalities

During the pretraining phase using UK Biobank data, we choose the following feature dimensions. For the raw-SNPs, we uniformly sample every 100^{th} SNP from 22 Chromosomes (excluding the X and Y chromosomes), resulting in 7,854 SNPs per sample. For PGS, we used 481 scores for a wide variety of different traits downloaded from the PGS Catalog [Lam+21]. We created burden scores for 18,574 protein-coding genes [Mon+21]. These binary scores indicate whether a participant has at least one potentially damaging rare (MAF $< 1\%$) variant within a given gene.

Diabetic Retinopathy detection (APTOS)

In this task we use the APTOS 2019 Blindness Detection [19] dataset, which has 3,662 retinal fundus training samples. As explained in the main paper, the labels in this dataset have five levels of disease severity, defining five classes. However, these classes are not mutually exclusive, as a higher disease severity of *e.g.* four is also of level three and below. Hence, we employ a multi-hot encoding scheme for the labels. For instance, class three is encoded as `[1,1,1,0,0]` and two as `[1,1,0,0,0]`, and so on. We split the dataset into three different splits of training (60%), validation (20%), and test (20%). There is no overlap of patients across these splits.

Retinal Fundus Disease Classification (RFMiD)

For this task, we use the Retinal Fundus Multi-disease Image Dataset (RFMiD) [Pac+21], which has 3,200 images. The overall number of disease classes is 45. However, we found that two classes ("HR" and "ODPM") have no positive cases, so we exclude these two classes and only work with the remaining 43 classes. As mentioned before, we convert these classes to multi-hot labels and solve the task as multilabel classification. We use this dataset's official splits for training, validation, and test.

Pathological Myopia Segmentation (PALM)

We use the Pathologic Myopia challenge dataset [Fu+19] for this task, which has 400 image samples with segmentation masks. As for segmentation labels, this dataset has three annotated areas: i) peripapillary atrophy (available for 311 cases), ii) optic disc (available for all cases), and iii) detachment (available for 12 cases only). Given that detachment is rarely available, we omit it from this task and only predict the atrophy and disc classes. We stratify the patients using the atrophy labels, to ensure equal representation of classes in train (60% of dataset size) / val (20%) / test (20%) splits.

Cardiovascular Risk Prediction (UKB)

To predict the cardiovascular risk factors of (sex, age, BMI, SBP, DBP, smoking status) from retinal fundus scans, we use 102, 219 images from the UKB [Sud+15]. This corresponds to the training split (70% of UKB dataset size). We use the remaining scans for validation (10% of dataset size) and for the test split (20%). Each person only appears in one split. The training for this task is performed using two models: i) one model to classify the categorical labels (sex to binary labels {0,1}, smoking status to binary labels too), ii) a second model to predict – solved as a regression task – the remaining continuous variables (age, BMI, SBP, and DBP). We use two models because the loss values of these two tasks have different scales. We preprocess the values of the continuous factors by standardization (removing the mean and scaling to unit variance). Finally, we impute the missing values of these factors by using the "mean" for continuous factors and "median" for discrete factors.

C.1.2 Imaging Preprocessing

Image Quality Control

The UK Biobank contains a relatively large number of retinal fundus images with bad quality (*e.g.* completely black or extremely overexposed). To filter out extreme outliers, we performed two steps of quality control. First, we only included images where a simple circle-detection algorithm [IK87] could find a circle. In the second step, we filtered out the top and bottom 0.5% brightest and darkest remaining images.

Image transformations

We cropped images to the circles detected in Appendix C.1.2 and rescaled to 448×448 pixels. During training, we randomly transform images by a rotation of up to 20° and flip the image horizontally with a 50% probability. We also follow the common practice of normalizing (standardizing) all the image intensities using the mean and standard deviation from ImageNet [Den+09].

C.1.3 Genetics Preprocessing

In all our experiments we used the genetic data provided by the UK Biobank. The three different genetic modalities require different preprocessing steps, which we detail in this section.

Raw SNPs

The raw SNPs are a cross section of all SNPs collected on microarray chips, collecting approximately 800k genetic variants in total across all chromosomes. More information on data collection can be found at https://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=263.

The individual SNPs are coded in additive format, *i.e.* 0 stands for no deviation from the reference genome, 1 means that one of the two chromosome copies has a deviation and the other not, and 2 means that both chromosome copies show a deviation from the reference genome. We treated SNPs as continuous variables (opposed to, *e.g.* separating them into three classes each) and imputed missing values by mode imputation. Since 800k feature dimensions are challenging to handle, and SNPs are highly spatially correlated along the genome [Rei+01], we only sampled every 100-th SNP from the full microarray. We also only included SNPs on the 22 autosomal (=not sex-specific) chromosomes, as handling sex chromosomes requires special statistical care and leads to non-shared features between genetic males and females. Together, this means we include 7,854 SNPs in our models.

Polygenic Risk Scores

For computing polygenic risk scores, we downloaded all PGS weight files included in the PGS Catalog [Lam+21] (https://ftp.ebi.ac.uk/pub/databases/spot/pgs/, last accessed October 11, 2021; at the time of writing, a large batch of new scores has been added to the PGS catalog), a collection of published PGS. The PGS files provide weights for a linear model to compute risk scores from the raw genetic data. To have a large intersection of available SNPs for our UKB population and the

weights provided by the PGS catalog, instead of using the raw microarray data from Appendix C.1.3, we used *imputed* data. The imputed data uses prior knowledge about correlations between SNPs collected and not collected on the respective microarray ("linkage disequilibrium", LD) to infer the missing features with high accuracy. Imputed data was precomputed by the UKB, and more information can be found at https://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=100319. Using the imputed data, we computed 481 polygenic scores for our cohort using the PLINK software [Pur+07], ignoring scores that gave errors or that only recorded genome positions in a different reference genome build.

For some traits, there are multiple distinct risk scores in the PGS catalog, as multiple independent studies have been performed on the same trait. For example, the trait "melanoma" appears 9 times in our subset of selected PGS scores, while other traits, such as "insomnia" appear only once. The scores contain partially overlapping genetic markers, and the number of SNPs used for each individual score vary from only 1 to several millions.

Burden Scores

We ran the Functional Annotation and Association Testing Pipeline [Mon+21] to functionally annotate all the genetic variants present in the UK Biobank 200k exome sequencing release [Szu+21]. Protein loss of function and missense variants that were predicted to be damaging were used to construct burden scores across all protein coding genes. We considered only rare variants with minor allele frequencies below 1%. Of these variants 41% were "singletons", i.e. only observed once in our sample. Specifically, each participant was assigned a binary vector of length 18,574 corresponding to the number of protein coding genes. For every gene, the entry in this vector is 1 if the participant harbored at least one potentially damaging variant in that gene, or 0 if no potentially damaging variants were observed in that gene for that participant. This coding has been applied in rare-variant association studies in order to aggregate the effects of many rare variants within genes, where it can boost statistical power and reduce the burden of multiple testing[Lee+12; Mon+21].

C.1.4 Training Details

We provide the training details for all pretraining (self-supervised) and downstream tasks in this section.

- **Batch sizes**: we use a unified batch size of 64 across all pretraining and downstream tasks.

- **Optimizers**: we use Adam optimizer [KB14b] in all pretraining and downstream tasks.
- **Schedulers**: during self-supervised pretraining (with ContIG and the baselines), we decay the learning rate with the cosine decay schedule without restarts [LH17].
- **Learning rates**: we use an initial learning rate of 0.001 across all tasks. However, we reduce the learning rate during training in the PALM semantic segmentation task to 1×10^{-4} after 10 warum epochs.
- **Weight decay**: in pretraining tasks, we use a weight decay factor of 1×10^{-6}. In downstream tasks, we use a weight decay factor of 1×10^{-5}.
- **Number of epochs**: in pretraining tasks, we train all models for 100 epochs. In downstream tasks, we fine-tune for:
  - For the PALM, APTOS, and RFMiD tasks: we train all models for 50 epochs.
  - For Cardiovascular risk prediction tasks: we fine-tune all models for 5 epochs ($\approx$ 8000 steps).
- **Network architectures**: for the *image encoder*, as mentioned before, we use a Resnet50 [He+16] architecture across all pretraining and downstream tasks. For the *genetics encoders*, we vary between following choices:
  - None: here we do not have any hidden fully-connected layer for the genetics, and we feed them as inputs to the projection head directly.
  - H1: we process the genetic inputs with one hidden layer of size 2048. (followed by a `ReLU` activation and `Batchnorm1D` layers)
  - H12: we process the genetics with two hidden layers, both of size 2048. (Each layer is followed by a `ReLU` and `Batchnorm1D`)

  For the *projection head*, we follow [Che+20a] in using two fully-connected layers. The first has a size of 2048 and is followed by a `ReLU`. The second has size of 128, which is the projection embedding size. Finally, for classification and regression downstream tasks we add one fully-connected `Linear` layer on top to perform the task. But for the *PALM segmentation* task, we add a U-Net [RFB15] decoder on top of the Resnet50 encoder. For upsampling layers in the decoder, we use transposed convolutional layers `ConvTranspose2d`.
- **Loss functions**: the used loss functions for each task are as follows:
  - ContIG: for training our method, we use a contrastive loss (`NTXentLoss`). This loss is implemented using a cross-entropy loss, where the model is trained to classify which sample is positive in each mini-batch. However, our version of the `NTXentLoss` only does inter-modal contrasting, and

not intra-modal. We set $\lambda = 0.75$ in this loss (Eq. 1 in the main paper), and the temperature $\tau = 0.1$. Note that a larger value of λ gives more importance to image features than genetic features.
- APTOS & RFMiD: we use the binary cross-entropy loss in both tasks.
- PALM: we use a weighted combined loss of Dice-loss [Sor] (weight=0.8) and binary cross-entropy (weight=0.2).
- Cardiovascular risk classification (sex & smoking status): we use a binary cross-entropy loss.
- Cardiovascular risk prediction (age & BMI & SBP & DBP): we use the Mean Square Error (MSE) loss.
- SimCLR [Che+20a]: this method uses the contrastive `NTXentLoss` too. We similarly set the temperature $\tau = 0.1$.
- NNCLR [Dwi+21]: this method uses the contrastive `NTXentLoss` too. We similarly set the temperature $\tau = 0.1$.
- Simsiam [CH21]: this method does not use negative sampling, and instead uses a Siamese network to minimize the similarity between two augmented views of the same image. Hence, the loss function used is the negative cosine similarity loss.
- BYOL [Gri+20]: this method has the same loss used in Simsiam, which is the negative cosine similarity.
- Barlow Twins [Zbo+21]: this method modifies the contrastive loss to compute the cross-correlation matrix between two sets of embeddings, which are for the same batch of images but with different image augmentations. Then, it tries to make this matrix close to the identity matrix.

C.1.5 Implementation Details

We implement all of our methods using `Python`. The libraries we rely on are `PyTorch v1.9.1`, `Pytorch-Lightning v1.4.8`, `torchvision v0.10.0`, `torchmetrics v0.4.0`, and `Lightly` [lig] (for baseline self-supervised implementations). We also follow the reproducibility instructions for `Pytorch-Lightning` [Lig], *i.e.* by setting a unified random seed of 42 for all scripts and workers, and by using `deterministic` algorithms. We attach our source code with this supplementary material submission.

C.2 GWAS Analysis Details

We produced feature vectors by computing the hidden-layer embedding for each image in the test-split of our dataset (10% of the whole dataset, 7,079 individuals). In contrast to the main training, we only used embeddings of the left eye and only included each individual once. Feature vectors were reduced to 10 dimensions using a PCA. Before computing the association results, we also used an inverse-normal transform [Sof+19] after conditioning on the potential confounders "sex", "age", as well as the first 15 genetic PCs. This ensures that the residuals of the marginal distributions are approximately normally distributed and outlier deviations from normality don't artificially inflate the type-1 error rate, leading to spurious correlations. We performed the genetic association study with the PLINK2 software [Cha+15], using a linear model for each of the ten dimensions individually. We again correct for the same confounders in the linear model. Finally, we aggregate the summary statistics of the ten individual features into a single p-value for each SNP by using a Bonferroni-correction of the factor 10, following [Kir+21].

Genetic variants are locally highly correlated. Therefore, we group significantly associated SNPs that are spatially close and in LD together using the PLINK [Pur+07] clumping functionality (using parameters `clump-p1` $= 5 \cdot 10^{-8}$, `clump-p2` $= 10^{-7}$, `clump-r2` $= 0.1$, `clump-kb` $= 150$). We reported the number of independent associated regions returned by this procedure in the main document.

C.3 Genetic Explanation Method Validation

We ran a baseline experiment to validate that our feature explanation method properly attributes to meaningful features. In this experiment, instead of genetic features, we use phenotypic covariates such as age, sex, systolic and diastolic blood pressure (SBP and DBP), which can be predicted reliably from retinal fundus images. Additionally, we include the first 40 principal components, which mostly capture population structure information. As control variables, we also feed five random noise variables into the training process, which have no association with the images at all. Fig. C.1 shows the aggregated feature explanations. As expected, the noise variables (`noise0`, ... `noise4`) get assigned very low explanation scores, while all other variables have considerable influence. This validates that our feature explanation approach can distinguish between variables that carry true information relevant to the network and variables that are unrelated to the images.

C.4 Multimodal Explanation Results

Finally, we consider the multi-modal model trained on all three genetic modalities, ContIG with the "Outer" training scheme. Fig. C.2 shows the aggregated attribution scores for each of the three modalities, Raw-SNPs, PGS, and Burdens, for ContIG with the "Outer" training scheme. Fig. C.2 (a) shows that PGS scores on average have more influence than individual SNPs or burden scores. However, Fig. C.2 (b) also shows that in aggregate, raw SNPs and burden scores have more total influence on the model. This is likely due to PGS only having 481 features, while raw SNPs and Burdens have 7,854 and 18,574 features, respectively. This may also explain the small but counterintuitive performance drop from ContIG (PGS) to ContIG (Outer RPB) in certain downstream tasks: the strongest signal – and probably the most relevant for those tasks –, PGS, gets "drowned out" by the less important but overabundant signal in the raw SNPs and burden scores.

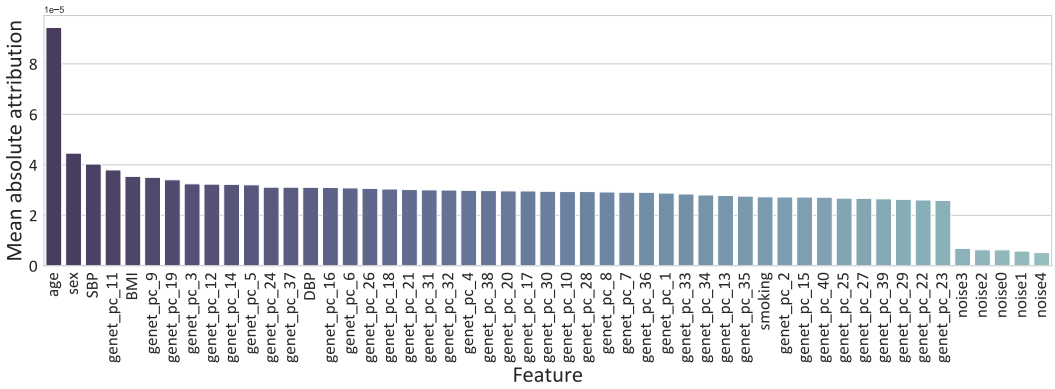
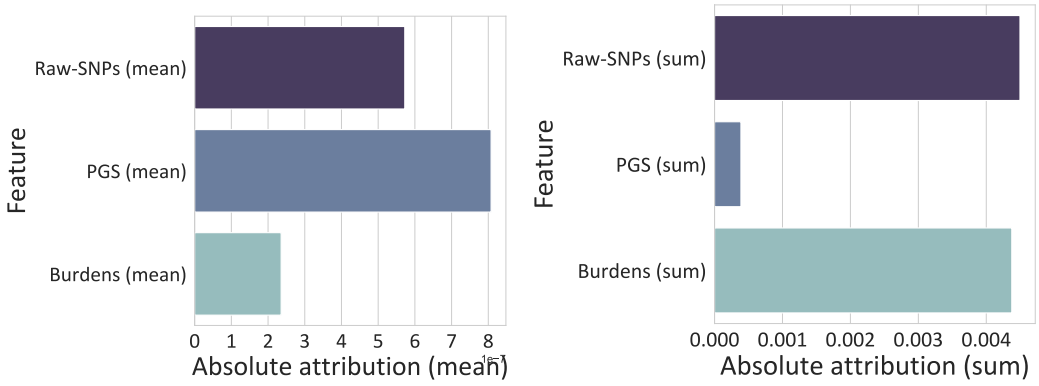


Figure C.1: Explanation method validation. Shown is the mean absolute attribution for each feature aggregated over a batch-size of 1,000 individuals. `noise0, ..., noise4` don't carry any information and also get downweighted by our attribution method. Figure Source [Tal+22a], reprinted with permission.



(a) Absolute attribution for each modality, aggregated by mean.

(b) Absolute attribution for each modality, aggregated by sum.

Figure C.2: Absolute attributions by modality for ContIG (Outer RPB). Figure Source [Tal+22a], reprinted with permission.

Bibliography

[19] *APTOS 2019*. https://www.kaggle.com/c/aptos2019-blindness-detection/. Accessed: 2020-11-02. 2019 (see pages 80, 81, 85, 119).

[AGG18] Tuka Al Hanai, Mohammad Ghassemi, and James Glass. **Detecting Depression with Audio/Text Sequence Modeling of Interviews**. In: *Proc. Interspeech 2018*. Graz, Austria: ISCOM, 2018, 1716–1720. DOI: 10.21437/Interspeech.2018-2522. URL: http://dx.doi.org/10.21437/Interspeech.2018-2522 (see pages 32, 74).

[Ala+20] 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 (2020), 7 (see page 74).

[Alb+18] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. **Emotion Recognition in Speech Using Cross-Modal Transfer in the Wild**. In: *Proceedings of the 26th ACM International Conference on Multimedia*. MM '18. Seoul, Republic of Korea: Association for Computing Machinery, 2018, 292–301. ISBN: 9781450356657. DOI: 10.1145/3240508.3240578. URL: https://doi.org/10.1145/3240508.3240578 (see page 32).

[Ali+21] Yamen Ali, Aiham Taleb, Marina M.-C. Höhne, and Christoph Lippert. **Self-Supervised Learning for 3D Medical Image Analysis using 3D SimCLR and Monte Carlo Dropout**. *CoRR* abs/2109.14288 (2021). arXiv: 2109.14288. URL: https://arxiv.org/abs/2109.14288 (see page 53).

[Alw+19] 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) (see page 74).

[AMZ21] Andrés Anaya-Isaza, Leonel Mera-Jiménez, and Martha Zequera-Diaz. **An overview of deep learning in medical imaging**. *Informatics in Medicine Unlocked* 26 (2021), 100723. ISSN: 2352-9148. DOI: https://doi.org/10.1016/j.imu.2021.100723. URL: https://www.sciencedirect.com/science/article/pii/S2352914821002033 (see pages 1, 18).

[Ant+15a] Grigory Antipov, Sid-Ahmed Berrani, Natacha Ruchaud, and Jean-Luc Dugelay. **Learned vs. hand-crafted features for pedestrian gender recognition**. In: *Proceedings of the 23rd ACM international conference on Multimedia*. 2015, 1263–1266 (see page 7).

[Ant+15b] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. **VQA: Visual Question Answering**. In: *The IEEE International Conference on Computer Vision (ICCV)*. Las Condes, Chile: IEEE, Dec. 2015, 2425–2433. DOI: 10.1109/ICCV.2015.536 (see page 32).

[Ard+19] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. **End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography**. *Nature medicine* 25:6 (2019), 954–961 (see page 1).

[Arm+19] Karim Armanious, Chenming Yang, Marc Fischer, Thomas Küstner, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. **MedGAN: Medical Image Translation using GANs**. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* 79 (2019) (see pages 19, 33).

[Asa+20] 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) (see page 74).

[Ash+21] 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 (2021), 1–12 (see pages 74, 86).

[AVT16] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. **SoundNet: Learning Sound Representations from Unlabeled Video**. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, 892–900. ISBN: 9781510838819 (see pages 32, 74).

[AYM14] Harris A Ahmad, Hui Jing Yu, and Colin G Miller. "Medical imaging modalities." In: *Medical imaging in clinical trials*. Springer, 2014, 3–26 (see page 17).

[Ayt+18] Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. **Cross-Modal Scene Networks**. *IEEE Trans. Pattern Anal. Mach. Intell.* 40:10 (Oct. 2018), 2303–2314. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2017.2753232. URL: https://doi.org/10.1109/TPAMI.2017.2753232 (see pages 32, 74).

[AZ11] Ryan Prescott Adams and Richard S. Zemel. **Ranking via Sinkhorn Propagation**. *CoRR* abs/1106.1925 (2011). arXiv: 1106.1925. URL: http://arxiv.org/abs/1106.1925 (see pages 31, 32).

[AZ17] Relja Arandjelovic and Andrew Zisserman. **Look, Listen and Learn**. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017 (see pages 32, 74).

[BA21] Yusuf Bayraktar and Enes Ayan. **Diagnosis of interproximal caries lesions with deep convolutional neural network in digital bitewing radiographs**. *Clinical oral investigations* (2021), 1–10 (see page 96).

[Bac+11] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. **Sequential deep learning for human action recognition**. In: *International workshop on human behavior understanding*. Springer. 2011, 29–39 (see page 13).

[Bai+18] Ujjwal Baid, Abhishek Mahajan, Sanjay Talbar, Swapnil Rane, Siddhesh Thakur, Aliasgar Moiyadi, Meenakshi Thakur, and Sudeep Gupta. **GBM Segmentation with 3D U-Net and Survival Prediction with Radiomics**. In: *International MICCAI Brainlesion Workshop*. Springer. 2018, 28–35 (see pages 66, 68).

[Bai+19] 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: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan. Cham: Springer International Publishing, 2019, 541–549. ISBN: 978-3-030-32245-8 (see page 24).

[Bak+17] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. **Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features**. *Scientific Data* 4 (Sept. 2017), 170117 (see pages 20, 34, 38, 65, 116).

[Bal+19] Guha Balakrishnan, Amy Zhao, Mert Sabuncu, John Guttag, and Adrian Dalca. **VoxelMorph: A Learning Framework for Deformable Medical Image Registration**. *IEEE Transactions on Medical Imaging* PP (Feb. 2019), 1–1. DOI: 10.1109/TMI.2019.2897538 (see page 43).

[BAM19] 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 (2019), 423–443 (see pages 17, 32, 33, 74).

[Bar+11] Caroline Barwood, Bruce Murdoch, B. Whelan, David Lloyd, Stephan Riek, John O'Sullivan, Alan Coulthard, Andrew Wong, Philip Aitken, and Graham Hall. **The effects of low frequency Repetitive Transcranial Magnetic Stimulation (rTMS) and sham condition rTMS on behavioural language in chronic non-fluent aphasia: Short term outcomes**. *NeuroRehabilitation* 28 (Mar. 2011), 113–28. DOI: 10.3233/NRE-2011-0640 (see page 53).

[Bay+21] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. **A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets**. *The Visual Computer* (2021), 1–32 (see page 17).

[BCV12] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. **Unsupervised Feature Learning and Deep Learning: A Review and New Perspectives**. *CoRR* abs/1206.5538 (2012). arXiv: 1206.5538. URL: http://arxiv.org/abs/1206.5538 (see page 7).

[BDD63] Arthur E Bryson Jr, Walter F Denham, and Stewart E Dreyfus. **Optimal programming problems with inequality constraints**. *AIAA journal* 1:11 (1963), 2544–2550 (see page 10).

[Bey+20] Thomas Beyer, Luc Bidaut, John Dickson, Marc Kachelriess, Fabian Kiessling, Rainer Leitgeb, Jingfei Ma, Lalith Kumar Shiyam Sundar, Benjamin Theek, and Osama Mawlawi. **What scans we will read: imaging instrumentation trends in clinical oncology**. *Cancer Imaging* 20:1 (2020), 1–38 (see pages 17, 18).

[BH92] Suzanna Becker and Geoffrey E Hinton. **Self-organizing neural network that discovers surfaces in random-dot stereograms**. *Nature* 355:6356 (1992), 161–163 (see page 97).

[Bioa] BioMe Biobank. *BioMe Biobank*. https://icahn.mssm.edu/research/ipm/programs/biome-biobank. Accessed: 2021-11-09 (see page 71).

[Biob] Estonia Biobank. *Estonia Biobank*. https://genomics.ut.ee/en. Accessed: 2021-11-09 (see page 71).

[Bioc] Nako Biobank. *Nako Biobank*. https://nako.de/. Accessed: 2021-11-09 (see page 71).

[Bis06] Christopher M. Bishop. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Springer-Verlag, 2006. ISBN: 0387310738 (see page 8).

[BL07] Yoshua Bengio and Yann Lecun. *Scaling learning algorithms towards AI*. 2007 (see page 10).

[BLK14] Joseph J Budovec, Cesar A Lam, and Charles E Kahn Jr. **Informatics in radiology: radiology gamuts ontology: differential diagnosis for the Semantic Web**. *Radiographics* 34:1 (2014), 254–264 (see page 71).

[BNH19] Maximilian Blendowski, Hannes Nickisch, and Mattias P. Heinrich. **How to Learn from Unlabeled Volume Data: Self-supervised 3D Context Feature Learning**. In: *MICCAI*. Ed. by Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan. Springer International Publishing, 2019, 649–657. ISBN: 978-3-030-32226-7 (see page 25).

[Bro+93] Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. **Signature verification using a "siamese" time delay neural network**. *International Journal of Pattern Recognition and Artificial Intelligence* 7:04 (1993), 669–688 (see page 97).

[BS20] Liviu Badea and Emil Stănescu. **Identifying transcriptomic correlates of histology using deep learning**. *PloS one* 15:11 (2020), e0242858 (see page 74).

[bus] businesswire. *Global Medical Imaging Market Report*. https://www.businesswire.com/news/home/20210608005582/en/Global-Medical-Imaging-Market-Report-2021-2026-Analysis-by-X-Ray-Ultrasound-MRI-CT-Scan-Nuclear-Imaging---ResearchAndMarkets.com. Accessed: 2022-11-22 (see page 1).

[CAF10] T. S. Cho, S. Avidan, and W. T. Freeman. **A probabilistic image jigsaw puzzle solver**. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010, 183–190 (see page 32).

[Can+20] Anselmo Garcia Cantu, Sascha Gehrung, Joachim Krois, Akhilanand Chaurasia, Jesus Gomez Rossi, Robert Gaudin, Karim Elhennawy, and Falk Schwendicke. **Detecting caries lesions of different radiographic extension on bitewings using deep learning**. *Journal of dentistry* 100 (2020), 103425 (see page 96).

[Car+18] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. **Deep Clustering for Unsupervised Learning of Visual Features**. In: *ECCV*. Munich, Germany: Springer, Sept. 2018 (see pages 23, 26).

[Car+19] Fabio M. Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. **Domain Generalization by Solving Jigsaw Puzzles**. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, June 2019 (see pages 40–44, 49).

[Car+20] 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) (see page 24).

[CD14] David Daniel Cox and Thomas Dean. **Neural networks and neuroscience-inspired computer vision**. *Current Biology* 24:18 (2014), R921–R929 (see page 7).

[CH21] Xinlei Chen and Kaiming He. **Exploring simple siamese representation learning**. In: *CVPR*. 2021, 15750–15758 (see pages 24, 81, 82, 84, 85, 88, 98, 124).

[Cha+15] 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 (2015), s13742–015 (see pages 86, 125).

[Cha+18a] Siddhartha Chandra, Maria Vakalopoulou, Lucas Fidon, Enzo Battistella, Theo Estienne, Roger Sun, Charlotte Robert, Eric Deutch, and Nikos Paragios. **Context Aware 3-D Residual Networks for Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*. Springer. 2018, 74–82 (see pages 66, 68).

[Cha+18b] Yi-Ju Chang, Zheng-Shen Lin, Tsai-Ling Yang, and Teng-Yi Huang. **Automatic segmentation of brain tumor from 3D MR images using a 2D convolutional neural networks**. In: *Pre-Conference Proceedings of the 7th MICCAI BraTS Challenge*. Granada, Spain: Springer, 2018 (see pages 39–41).

[Cha+18c] 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 (2018), 1201–1207 (see page 74).

[Cha+20] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. **Contrastive learning of global and local features for medical image segmentation with limited annotations**. *Advances in Neural Information Processing Systems* 33 (2020) (see page 25).

[Cha01] Monty Charles. **UNSCEAR Report 2000: sources and effects of ionizing radiation**. *Journal of Radiological Protection* 21:1 (2001), 83 (see page 95).

[Che+11] 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 (2011), 812–818 (see page 89).

[Che+19] 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 (2019), 101539. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2019.101539. URL: http://www.sciencedirect.com/science/article/pii/S1361841518304699 (see page 25).

[Che+20a] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. **A Simple Framework for Contrastive Learning of Visual Representations**. In: *Int. Conf. on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, 1597–1607. URL: https://proceedings.mlr.press/v119/chen20j.html (see pages 24, 27, 59, 61, 64, 78, 81–85, 88, 89, 97, 98, 100, 101, 103, 104, 115, 118, 123, 124).

[Che+20b] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. **Improved baselines with momentum contrastive learning**. *arXiv preprint arXiv:2003.04297* (2020) (see page 24).

[Cho+14] Kyunghyun Cho, Bart van Merriënboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. **Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation**. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, 1724–1734. DOI: 10.3115/v1/D14-1179. URL: https://www.aclweb.org/anthology/D14-1179 (see page 117).

[Cho+19] Dami Choi, Christopher J Shallue, Zachary Nado, Jaehoon Lee, Chris J Maddison, and George E Dahl. **On empirical comparisons of optimizers for deep learning**. *arXiv preprint arXiv:1910.05446* (2019) (see page 11).

[Chu+17] Joon S. Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. **Lip Reading Sentences in the Wild**. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, Hawaii, USA: IEEE, 2017, 3444–3453 (see page 32).

[Coh68] Jacob Cohen. **Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit.** *Psychological bulletin* 70:4 (1968), 213 (see page 81).

[Dai+21] 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 (2021), 1–10 (see page 74).

[DCK19] Kim Dahun, Donghyeon Cho, and Soo-Ok Kweon. **Self-Supervised Video Representation Learning with Space-Time Cubic Puzzles**. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), 8545–8552. DOI: 10.1609/aaai.v33i01.33018545 (see page 55).

[Den+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. **ImageNet: A Large-Scale Hierarchical Image Database**. In: *CVPR09*. Miami, FL, USA: IEEE, 2009 (see pages 2, 7, 14, 20, 24, 60, 105, 121).

[Dev+18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. **Bert: Pre-training of deep bidirectional transformers for language understanding**. *arXiv preprint arXiv:1810.04805* (2018) (see page 12).

[DGE15] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. **Unsupervised Visual Representation Learning by Context Prediction**. In: *ICCV*. ICCV '15. USA: IEEE Computer Society, 2015, 1422–1430. ISBN: 9781467383912. DOI: 10.1109/ICCV.2015.167. URL: https://doi.org/10.1109/ICCV.2015.167 (see pages 7, 22, 26, 29, 31, 62).

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. **Adaptive subgradient methods for online learning and stochastic optimization.** *Journal of machine learning research* 12:7 (2011) (see page 11).

[Dic03] Adam B Dickerson. **Kant on representation and objectivity**. Cambridge University Press, 2003 (see page 5).

[DKD16] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. **Adversarial feature learning**. *arXiv preprint arXiv:1605.09782* (2016) (see page 21).

[Doi07] Kunio Doi. **Computer-aided diagnosis in medical imaging: historical review, current status and future potential**. *Computerized medical imaging and graphics* 31:4-5 (2007), 198–211 (see pages 1, 109).

[Dos+14] Alexey Dosovitskiy, Jost T. Springenberg, Martin Riedmiller, and Thomas Brox. **Discriminative Unsupervised Feature Learning with Convolutional Neural Networks**. In: *Advances in Neural Information Processing Systems 27 (NIPS)*. 2014. URL: http://lmb.informatik.uni-freiburg.de/Publications/2014/DB14b (see pages 23, 24, 26, 64).

[Dos+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. **An image is worth 16x16 words: Transformers for image recognition at scale**. *arXiv preprint arXiv:2010.11929* (2020) (see page 12).

[DT05] Navneet Dalal and Bill Triggs. **Histograms of oriented gradients for human detection**. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. Ieee. 2005, 886–893 (see page 7).

[Dud13] Frank Dudbridge. **Power and predictive accuracy of polygenic risk scores**. *PLoS genetics* 9:3 (2013), e1003348 (see page 76).

[Dwi+21] 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*. Oct. 2021, 9588–9597 (see pages 24, 81, 82, 84, 85, 88, 124).

[EJ16] Aarthipoornima Elangovan and T Jeyaseelan. **Medical imaging modalities: a survey**. In: *2016 International Conference on emerging trends in engineering, technology and science (ICETETS)*. ieee. 2016, 1–4 (see page 17).

[EM11] Ronald Eisenberg and Alexander Margulis. **A Patient's Guide to Medical Imaging**. NY, USA: New York: Oxford University Press, 2011, 45–67 (see pages 17, 29, 33, 54).

[Era+19] Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J Theis. **Deep learning: new computational modelling techniques for genomics**. *Nature Reviews Genetics* 20:7 (2019), 389–403 (see page 74).

[Est+17] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. **Dermatologist-level classification of skin cancer with deep neural networks**. *nature* 542:7639 (2017), 115–118 (see pages 1, 19).

[Eve+10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. **The pascal visual object classes (voc) challenge**. *International journal of computer vision* 88:2 (2010), 303–338 (see page 24).

[Far+20] Abolfazl Farahani, Behrouz Pourshojae, Khaled Rasheed, and Hamid R Arabnia. **A concise review of transfer learning**. In: *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE. 2020, 344–351 (see page 105).

[FC18] Jonathan Frankle and Michael Carbin. **The lottery ticket hypothesis: Finding sparse, trainable neural networks**. *arXiv preprint arXiv:1803.03635* (2018) (see page 110).

[Fin] FinnGen. *FinnGen*. https://www.finngen.fi/en. Accessed: 2021-11-09 (see page 71).

[Fra+09] 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 (2009), 241–251 (see page 76).

[Fu+18] Chichen Fu, Soonam Lee, David Ho, Shuo Han, Paul Salama, Kenneth Dunn, and Edward Delp. **Three Dimensional Fluorescence Microscopy Image Synthesis and Segmentation**. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. June 2018. DOI: 10.1109/CVPRW.2018.00298 (see page 33).

[Fu+19] 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. DOI: 10.21227/55pk-8z03. URL: https://dx.doi.org/10.21227/55pk-8z03 (see pages 81, 83, 85, 120).

[Fu+20] Yabo Fu, Yang Lei, Tonghe Wang, Walter J Curran, Tian Liu, and Xiaofeng Yang. **Deep learning in medical image registration: a review**. *Physics in Medicine & Biology* 65:20 (2020), 20TR01 (see page 19).

[Fuj+21] 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) (see page 74).

[Gal12] Andrew C. Gallagher. **Jigsaw puzzles with pieces of unknown orientation**. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012, 382–389 (see page 32).

[Gal16] Yarin Gal. **Uncertainty in Deep Learning**. PhD thesis. University of Cambridge, 2016 (see pages 8, 14).

[Gau09] C.F. Gauss. **Theoria motus corporum coelestium in sectionibus conicis solem ambientium**. Carl Friedrich Gauss Werke. sumtibus F. Perthes et I. H. Besser, 1809. URL: https://books.google.be/books?id=ORUOAAAAQAAJ (see page 8).

[GB19] David Griffiths and Jan Boehm. **A review on deep learning techniques for 3D sensed data classification**. *Remote Sensing* 11:12 (2019), 1499 (see pages 53, 55).

[GBC16] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. **Deep Learning**. http://www.deeplearningbook.org. Cambridge, MA, USA: MIT Press, 2016 (see pages 7, 9, 12).

[GH10] Michael Gutmann and Aapo Hyvärinen. **Noise-contrastive estimation: A new estimation principle for unnormalized statistical models**. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR,

May 2010, 297–304. URL: http://proceedings.mlr.press/v9/gutmann10a.html (see pages 24, 58–60, 64, 97).

[Gia+14] 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 (2014), 873–882 (see page 88).

[Gia+18] Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. **Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data**. *JAMA Internal Medicine* 178:11 (Nov. 2018), 1544–1547. ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2018.3763. eprint: https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/2697394/jamainternal_gianfrancesco_2018_sc_180005.pdf. URL: https://doi.org/10.1001/jamainternmed.2018.3763 (see pages 102, 105).

[Gir+13] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. **Rich feature hierarchies for accurate object detection and semantic segmentation**. *CoRR* abs/1311.2524 (2013). arXiv: 1311.2524. URL: http://arxiv.org/abs/1311.2524 (see pages 7, 13, 15).

[Gon+14] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. **Multi-scale orderless pooling of deep convolutional activation features**. In: *European conference on computer vision*. Springer. 2014, 392–407 (see pages 13, 15).

[Goo+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. **Generative adversarial nets**. *Advances in neural information processing systems* 27 (2014) (see page 21).

[Gor+18] Yu Gordienko, Peng Gang, Jiang Hui, Wei Zeng, Yu Kochura, Oleg Alienin, Oleksandr Rokovyi, and Sergii Stirenko. **Deep learning with lung segmentation and bone shadow exclusion techniques for chest X-ray analysis of lung cancer**. In: *International conference on computer science, engineering and education applications*. Springer. 2018, 638–647 (see page 19).

[Goy+19] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. **Scaling and Benchmarking Self-Supervised Visual Representation Learning**. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019 (see pages 50, 65).

[Gri+20] 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. arXiv: 2006.07733 [cs.LG] (see pages 24, 64, 81, 82, 84–86, 88, 97, 98, 100, 101, 124).

[Gro+19] Aditya Grover, Eric Wang, Aaron Zweig, and Stefano Ermon. **Stochastic Optimization of Sorting Networks via Continuous Relaxations**. In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=H1eSS3CcKX (see page 32).

[Grü+17] Katharina Grünberg, Oscar Jimenez-del-Toro, Andras Jakab, Georg Langs, Tomàs Salas Fernandez, Marianne Winterstein, Marc-André Weber, and Markus Krenn, 45–67. In: *Cloud-Based Benchmarking of Medical Image Analysis*. Springer International Publishing, 2017 (see pages 1, 19).

[GSK18] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. **Unsupervised Representation Learning by Predicting Image Rotations**. *CoRR* abs/1803.07728 (2018). arXiv: 1803.07728. URL: http://arxiv.org/abs/1803.07728 (see pages 23, 26, 55, 63).

[Gu+18] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. **Recent advances in convolutional neural networks**. *Pattern recognition* 77 (2018), 354–377 (see page 13).

[Gul+16] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. **Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs**. *Jama* 316:22 (2016), 2402–2410 (see pages 1, 19).

[Gun+20] 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 (see page 74).

[Guo+20] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. **Deep learning for 3d point clouds: A survey**. *IEEE transactions on pattern analysis and machine intelligence* 43:12 (2020), 4338–4364 (see page 55).

[Has+17] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. **Neuroscience-inspired artificial intelligence**. *Neuron* 95:2 (2017), 245–258 (see page 7).

[He+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. **Deep Residual Learning for Image Recognition**. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (see pages 81, 102, 123).

[He+20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. **Momentum contrast for unsupervised visual representation learning**. In: *CVPR*. 2020, 9729–9738 (see pages 24, 64).

[Hen20] Olivier Henaff. **Data-efficient image recognition with contrastive predictive coding**. In: *International Conference on Machine Learning*. PMLR. 2020, 4182–4192 (see pages 23, 27, 58, 60, 65, 97).

[Hew11] Robert E Hewitt. **Biobanking: the foundation of personalized medicine**. *Current opinion in oncology* 23:1 (2011), 112–119 (see page 71).

[Hje+18] 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) (see page 24).

[HMD15] Song Han, Huizi Mao, and William J Dally. **Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding**. *arXiv preprint arXiv:1510.00149* (2015) (see page 110).

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. **Long short-term memory**. *Neural computation* 9:8 (1997), 1735–1780 (see page 12).

[HSS18] Jie Hu, Li Shen, and Gang Sun. **Squeeze-and-excitation networks**. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, 7132–7141 (see page 109).

[HTG16] David Harwath, Antonio Torralba, and James Glass. "Unsupervised Learning of Spoken Language with Visual Context." In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Barcelona, Spain: Curran Associates, Inc., 2016, 1858–1866. URL: http://papers.nips.cc/paper/6186-unsupervised-learning-of-spoken-language-with-visual-context.pdf (see page 32).

[Hu+19] 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 (2019), 132816–132827 (see page 25).

[Hua+17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. **Densely Connected Convolutional Networks**. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, 2261–2269 (see page 115).

[Hum03] David Hume. **A treatise of human nature**. Courier Corporation, 2003 (see page 5).

[Ian+16] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. **SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size**. *arXiv preprint arXiv:1602.07360* (2016) (see pages 109, 110).

[IHA18] Sheikh Muhammad Saiful Islam, Md Mahedi Hasan, and Sohaib Abdullah. **Deep Learning based Early Detection and Grading of Diabetic Retinopathy Using Retinal Fundus Images**. *CoRR* abs/1812.10595 (2018). arXiv: 1812.10595. URL: http://arxiv.org/abs/1812.10595 (see page 2).

[IK87] John Illingworth and Josef Kittler. **The adaptive Hough transform**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-9:5 (1987), 690–698 (see page 120).

[Insa] National Human Genome Research Institute. *DNA Sequencing*. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet. Accessed: 2021-10-31 (see page 76).

[Insb] 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 (see page 73).

[Insc] National Human Genome Research Institute. *Genome-Wide Association Studies Fact Sheet*. https://www.ebi.ac.uk/gwas/. Accessed: 2021-11-03 (see pages 73, 86).

[Ioa+17] Anastasia Ioannidou, Elisavet Chatzilari, Spiros Nikolopoulos, and Ioannis Kompatsiaris. **Deep Learning Advances in Computer Vision with 3D Data: A Survey**. *ACM Computing Surveys* 50 (June 2017). DOI: 10.1145/3042064 (see pages 53, 55, 109).

[Ise+18] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein. **No new-net**. In: *International MICCAI Brainlesion Workshop*. Granada, Spain: Springer, 2018, 234–244 (see pages 39–41, 66, 68).

[Iso+17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. **Image-to-Image Translation with Conditional Adversarial Networks**. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, Hawaii, USA: IEEE, 2017, 5967–5976 (see page 33).

[J] Leonardo Calderon J. *Leonardo Calderon J. - Linkedin Blog*. https://www.linkedin.com/pulse/activation-functions-neural-networks-leonardo-calderon-j-/. Accessed: 2022-07-21 (see page 10).

[JAF16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. **Perceptual Losses for Real-Time Style Transfer and Super-Resolution**. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Cham: Springer International Publishing, 2016, 694–711. ISBN: 978-3-319-46475-6 (see page 112).

[Ji+12] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. **3D convolutional neural networks for human action recognition**. *IEEE transactions on pattern analysis and machine intelligence* 35:1 (2012), 221–231 (see page 13).

[Ji+13] S. Ji, W. Xu, M. Yang, and K. Yu. **3D Convolutional Neural Networks for Human Action Recognition**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35:1 (2013), 221–231 (see page 55).

[Jia+20] 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)*. 2020, 1847–1850 (see page 25).

[JKZ17] 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*. Springer, Sept. 2017, 294–302. ISBN: 978-3-319-67557-2. DOI: 10.1007/978-3-319-67558-9_34 (see page 24).

[Joh+16] Justin Johnson, Andrej Karpathy, Fei Fei Li, and Djabeur Mohamed Seifeddine Zekrifa. **DenseCap: Fully Convolutional Localization Networks for Dense Captioning**. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, June 2016, 4565–4574 (see pages 32, 74).

[JRP19] Amir Jaberzadeh, Lokesh Rukmangadachar, and Daniel Pelletier. **3D MR image synthesis using unsupervised deep learning algorithm in MS patients**. *Neurology* 92:15 Supplement (2019), P5. 2–025. ISSN: 0028-3878. eprint: https://n.neurology.org/content (see pages 19, 33).

[JT18] Longlong Jing and Yingli Tian. **Self-supervised Spatiotemporal Feature Learning by Video Geometric Transformations**. *CoRR* abs/1811.11387 (2018). arXiv: 1811.11387. URL: http://arxiv.org/abs/1811.11387 (see page 55).

[JT20] Longlong Jing and Yingli Tian. **Self-supervised visual feature learning with deep neural networks: A survey**. *IEEE transactions on pattern analysis and machine intelligence* (2020) (see pages 21, 55).

[Kar+17] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. **Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion**. *ACM Trans. Graph.* 36:4 (July 2017). ISSN: 0730-0301. DOI: 10.1145/3072959.3073658. URL: https://doi.org/10.1145/3072959.3073658 (see page 32).

[Kas+17] Nicholas J Kassebaum, Amanda GC Smith, Eduardo Bernabé, Thomas D Fleming, Alex E Reynolds, Theo Vos, CJL Murray, W Marcenes, and GBD 2015 Oral Health Collaborators. **Global, regional, and national prevalence, incidence, and disability-adjusted life years for oral conditions for 195 countries, 1990–2015: a systematic analysis for the global burden of diseases, injuries, and risk factors**. *Journal of dental research* 96:4 (2017), 380–387 (see page 95).

[Kav+19] Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer. *CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data*. Version v1.03. Apr. 2019. DOI: 10.5281/zenodo.3362844. URL: https://doi.org/10.5281/zenodo.3362844 (see pages 37, 38).

[KB14a] Diederik P Kingma and Jimmy Ba. **Adam: A method for stochastic optimization**. *arXiv preprint arXiv:1412.6980* (2014) (see page 102).

[KB14b] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. 2014. URL: http://arxiv.org/abs/1412.6980 (see pages 11, 111, 115, 123).

[Ker+17] Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim. **Deep learning applications in medical image analysis**. *Ieee Access* 6 (2017), 9375–9389 (see pages 1, 18, 19).

[Kha+21] Sanjeev B. Khanagar, Ali Al-ehaideb, Prabhadevi C. Maganur, Satish Vishwanathaiah, Shankargouda Patil, Hosam A. Baeshen, Sachin C. Sarode, and Shilpa Bhandi. **Developments, application, and performance of artificial intelligence in dentistry – A systematic review**. *Journal of Dental Sciences* 16:1 (2021), 508–522. ISSN: 1991-7902. DOI: https://doi.org/10.1016/j.jds.2020.06.019. URL: https://www.sciencedirect.com/science/article/pii/S1991790220301434 (see page 95).

[Khe+18] 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 (2018), 1219–1224 (see page 76).

[Kim+19a] Jaeyoung Kim, Hong-Seok Lee, In-Seok Song, and Kyu-Hwan Jung. **DeNTNet: Deep Neural Transfer Network for the detection of periodontal bone loss using panoramic dental radiographs**. *Scientific reports* 9:1 (2019), 1–9 (see pages 95, 96).

[Kim+19b] Mingyu Kim, Jihye Yun, Yongwon Cho, Keewon Shin, Ryoungwoo Jang, Hyun-jin Bae, and Namkug Kim. **Deep learning in medical imaging**. *Neurospine* 16:4 (2019), 657 (see pages 1, 18, 19).

[Kir+21] 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). DOI: 10.1101/2021.10.22.465430. eprint: https://www.biorxiv.org/content/early/2021/10/24/2021.10.22.465430.full.pdf. URL: https://www.biorxiv.org/content/early/2021/10/24/2021.10.22.465430 (see pages 74, 86, 125).

[Kou20] Lefteris Koumakis. **Deep learning models in genomics; are we there yet?** *Computational and Structural Biotechnology Journal* 18 (2020), 1466–1473. ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2020.06.017. URL: https://www.sciencedirect.com/science/article/pii/S2001037020303068 (see page 74).

[Kov+16] Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, Kristen Grauman, et al. **Crowdsourcing in computer vision**. *Foundations and Trends® in computer graphics and Vision* 10:3 (2016), 177–243 (see pages 19, 20).

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Lake Tahoe, NV, USA: Curran Associates, Inc., 2012, 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf (see pages 7, 12, 14, 31).

[Kum+20] Ashnil Kumar, Michael Fulham, Dagan Feng, and Jinman Kim. **Co-Learning Feature Fusion Maps From PET-CT Images of Lung Cancer**. *IEEE Transactions on Medical Imaging* 39:1 (2020), 204–217 (see pages 32, 39, 41, 43).

[Kuz+20] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. **The open images dataset v4**. *International Journal of Computer Vision* 128:7 (2020), 1956–1981 (see page 105).

[KW13] Diederik P Kingma and Max Welling. **Auto-encoding variational bayes**. *arXiv preprint arXiv:1312.6114* (2013) (see page 21).

[KZB19] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. **Revisiting Self-Supervised Visual Representation Learning**. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (see page 115).

[Lam+21] Samuel A Lambert, Laurent Gil, Simon Jupp, Scott C Ritchie, Yu Xu, Annalisa Buniello, Aoife McMahon, Gad Abraham, Michael Chapman, Helen Parkinson, et al. **The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation**. *Nature Genetics* 53:4 (2021), 420–425 (see pages 80, 119, 121).

[LAS20] Lihao Liu, Angelica I Aviles-Rivero, and Carola-Bibiane Schönlieb. **Contrastive Registration for Unsupervised Medical Image Segmentation**. *arXiv preprint arXiv:2011.08894* (2020) (see page 25).

[Lea] Scikit Learn. *ROC AUC Micro Averaging*. https://scikit-learn.org/stable/modules/model_evaluation.html. Accessed: 2021-11-04 (see page 83).

[LeC+89a] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. **Backpropagation Applied to Handwritten Zip Code Recognition**. *Neural Computation* 1 (1989), 541–551 (see page 12).

[LeC+89b] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. **Handwritten digit recognition with a back-propagation network**. *Advances in neural information processing systems* 2 (1989) (see page 13).

[Lee+12] 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 (2012), 224–237 (see pages 77, 122).

[Lee+13] Dong-Hyun Lee et al. **Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks**. In: *Workshop on challenges in representation learning, ICML*. Vol. 3. 2. 2013, 896 (see page 22).

[LF18] 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)*. Washington, DC, USA: IEEE, Apr. 2018, 1075–1078 (see page 24).

[LH17] Ilya Loshchilov and Frank Hutter. *SGDR: Stochastic Gradient Descent with Warm Restarts*. 2017. arXiv: 1608.03983 [cs.LG] (see pages 102, 123).

[Li+19] 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) (see page 74).

[Li+20] Ying Li, Lingfei Ma, Zilong Zhong, Fei Liu, Michael A Chapman, Dongpu Cao, and Jonathan Li. **Deep learning for lidar point clouds in autonomous driving: A review**. *IEEE Transactions on Neural Networks and Learning Systems* 32:8 (2020), 3412–3432 (see pages 53, 55).

[Li18] Xiaochuan Li. **Fused U-Net for Brain Tumor Segmentation based on Multimodal MR Images**. In: *Pre-Conference Proceedings of the 7th MICCAI BraTS Challenge*. Granada, Spain: Springer, 2018 (see pages 39–41).

[Lig] PyTorch Lightning. *Reproducibility*. https://pytorch-lightning.readthedocs.io/en/latest/common/trainer.html#reproducibility. Accessed: 2021-11-19 (see page 124).

[lig] lightly.ai. *lightly*. https://github.com/lightly-ai/lightly. Accessed: 2021-11-20 (see pages 103, 124).

[Lin+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. **Microsoft coco: Common objects in context**. In: *European conference on computer vision*. Springer. 2014, 740–755 (see page 105).

[Lip+11] 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 (2011), 833–835 (see pages 77, 80, 86).

[Lip+17] 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 (2017), 10166–10171 (see page 72).

[Liu+17] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado, et al. **Detecting cancer metastases on gigapixel pathology images**. *arXiv preprint arXiv:1703.02442* (2017) (see page 19).

[Liu+18] 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). arXiv: 1806.09521. URL: http://arxiv.org/abs/1806.09521 (see page 24).

[Liu+21] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. **Self-supervised learning: Generative or contrastive**. *IEEE Transactions on Knowledge and Data Engineering* (2021) (see page 21).

[Lon+12] Dan Long, Jinwei Wang, Min Xuan, Quanquan Gu, Xiaojun Xu, Dexing Kong, and Minming Zhang. **Automatic Classification of Early Parkinson's Disease with Multi-Modal MR Imaging**. *PLOS ONE* 7:11 (Nov. 2012), 1–9. DOI: 10.1371/journal.pone.0047714. URL: https://doi.org/10.1371/journal.pone.0047714 (see page 29).

[LS17] Paras Lakhani and Baskaran Sundaram. **Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks**. *Radiology* 284:2 (2017), 574–582 (see page 19).

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. **Fully convolutional networks for semantic segmentation**. In: *CVPR*. 2015, 3431–3440 (see pages 13, 15).

[Lu+19] 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) (see page 74).

[LW16] Chuan Li and Michael Wand. **Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks**. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Cham: Springer International Publishing, 2016, 702–716. ISBN: 978-3-319-46487-9 (see page 112).

[Ma+20] 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 (2020), 450–469 (see pages 1, 18, 74).

[Man+09] 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 (2009), 747–753 (see page 73).

[Man10] Teri A Manolio. **Genomewide association studies and assessment of the risk of disease**. *New England journal of medicine* 363:2 (2010), 166–176 (see page 73).

[McK+20] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. **International evaluation of an AI system for breast cancer screening**. *Nature* 577:7788 (2020), 89–94 (see page 1).

[Men+15] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, and et al. **The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)**. *IEEE Transactions on Medical Imaging* 34:10 (2015), 1993–2024 (see pages 20, 34, 38, 65, 116).

[Men+18] Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. **Learning Latent Permutations with Gumbel-Sinkhorn Networks**. In: *International Conference on Learning Representations*. Vancouver, Canada: OpenReview, 2018. URL: https://openreview.net/forum?id=Byt3oJ-0W (see pages 31, 32, 36).

[mer] merriam-webster. *merriam-webster - Representationalism*. https://www.merriam-webster.com/dictionary/representationalism. Accessed: 2022-07-22 (see page 5).

[Mik+13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. **Efficient Estimation of Word Representations in Vector Space**. In: *1st International Conference on Learning Representations, ICLR 2013, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. Scottsdale, Arizona, USA: OpenReview, 2013. URL: http://arxiv.org/abs/1301.3781 (see pages 21, 22, 58, 62).

[MK20] L Megalan Leo and T Kalpalatha Reddy. **Dental Caries Classification System Using Deep Learning Based Convolutional Neural Network**. *Journal of Computational and Theoretical Nanoscience* 17:9-10 (2020), 4660–4665 (see page 96).

[MM20] Ishan Misra and Laurens van der Maaten. **Self-supervised learning of pretext-invariant representations**. In: *CVPR*. 2020, 6707–6717 (see pages 23, 26).

[Mon+21] Remo Monti, Pia Rautenstrauch, Mahsa L Ghanbari, Alva Rani James, Uwe Ohler, Stefan Konigorski, and Christoph Lippert. **Identifying interpretable gene-biomarker associations with functionally informed kernel-based tests in 190,000 exomes**. *bioRxiv* (2021) (see pages 80, 119, 122).

[Nes83] Yurii E Nesterov. **A method for solving the convex programming problem with convergence rate O (1/kˆ 2)**. In: *Dokl. akad. nauk Sssr*. Vol. 269. 1983, 543–547 (see page 11).

[NF16] Mehdi Noroozi and Paolo Favaro. **Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles**. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Cham: Springer International Publishing, 2016, 69–84. ISBN: 978-3-319-46466-4 (see pages 22, 26, 29, 31, 32, 49, 63, 117).

[Ngi+11] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. **Multimodal Deep Learning**. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Bellevue, Washington, USA: Omnipress, 2011, 689–696. ISBN: 9781450306195 (see pages 17, 32, 74).

[Nin+18] 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 (2018), 151–158 (see page 74).

[OE18] Andrew Owens and Alexei A. Efros. **Audio-Visual Scene Analysis with Self-Supervised Multisensory Features**. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018 (see pages 32, 74).

[OLV18] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. **Representation Learning with Contrastive Predictive Coding**. *CoRR* abs/1807.03748 (2018). arXiv: 1807.03748. URL: http://arxiv.org/abs/1807.03748 (see pages 23, 24, 58–60, 83, 97, 117).

[Oma+05] 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 (2005), 1620–1639 (see page 71).

[Oor+16] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu koray, Oriol Vinyals, and Alex Graves. "Conditional Image Generation with PixelCNN Decoders." In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc., 2016, 4790–4798. URL: http://papers.nips.cc/paper/6527-conditional-image-generation-with-pixelcnn-decoders.pdf (see page 117).

[Org] World Health Organization. *World Radiography Day*. https://www3.paho.org/hq/index.php?option=com_content&view=article&id=7410:2012-dia-radiografia-dos-tercios-poblacion-mundial-no-tiene-acceso-diagnostico-imagen&Itemid=0&lang=en#gsc.tab=0. Accessed: 2022-11-22 (see page 1).

[Owe+18] 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 (Oct. 2018), 1120–1137. ISSN: 0920-5691. DOI: 10.1007/s11263-018-1083-5. URL: https://doi.org/10.1007/s11263-018-1083-5 (see pages 32, 74).

[Pac+21] 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). ISSN: 2306-5729. DOI: 10.3390/data6020014. URL: https://www.mdpi.com/2306-5729/6/2/14 (see pages 80, 82, 85, 119).

[Pad+20] Sriramakrishnan Padmanaban, Kalaiselvi Thiruvenkadam, Padmapriya T., M. Thirumalaiselvi, and RAM KUMAR. **A Role of Medical Imaging Techniques in Human Brain Tumor Treatment**. 8 (Jan. 2020), 565–568. DOI: 10.35940/ijrte.D1105.1284S219 (see page 53).

[PAP18] Anmol Popli, Manu Agarwal, and G.N. Pillai. **Automatic Brain Tumor Segmentation using U-Net based 3D Fully Convolutional Network**. In: *Pre-Conference Proceedings of the 7th MICCAI BraTS Challenge*. Springer. 2018, 374–382 (see pages 66, 68).

[Pat+16] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. **Context Encoders: Feature Learning by Inpainting**. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (see pages 23, 26).

[Pat+21] 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*. 2021, 9577–9587 (see page 74).

[Per+16] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. **Brain tumor segmentation using convolutional neural networks in MRI images**. *IEEE transactions on medical imaging* 35:5 (2016), 1240–1251 (see page 19).

[PG16a] Senthil Purushwalkam and Abhinav Gupta. **Pose from Action: Unsupervised Learning of Pose Features based on Motion**. *CoRR* abs/1609.05420 (2016). arXiv: 1609.05420. URL: http://arxiv.org/abs/1609.05420 (see pages 32, 74).

[PG16b] Senthil Purushwalkam and Abhinav Gupta. **Pose from Action: Unsupervised Learning of Pose Features based on Motion**. *CoRR* abs/1609.05420 (2016). arXiv: 1609.05420. URL: http://arxiv.org/abs/1609.05420 (see page 55).

[phi] philosophybasics. *philosophybasics - Representationalism*. https://www.philosophybasics.com/branch_representationalism.html. Accessed: 2022-07-22 (see page 5).

[PK16] Sang Tae Park and Jayoung Kim. **Trends in next-generation sequencing and a new era for whole genome sequencing**. *International neurourology journal* 20 (2016), S76 (see page 71).

[PLA72] R. L. PLACKETT. **Studies in the History of Probability and Statistics. XXIX: The discovery of the method of least squares**. *Biometrika* 59:2 (Aug. 1972), 239–251. ISSN: 0006-3444. DOI: 10.1093/biomet/59.2.239. eprint: http://oup.prod.sis.lan/biomet/article-pdf/59/2/239/928516/59-2-239.pdf. URL: https://doi.org/10.1093/biomet/59.2.239 (see page 8).

[Pol64] Boris T Polyak. **Some methods of speeding up the convergence of iteration methods**. *Ussr computational mathematics and mathematical physics* 4:5 (1964), 1–17 (see page 11).

[Pop+18] 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 (2018), 158–164 (see pages 83, 88).

[PPT18] Marie-Morgane Paumard, David Picard, and Hedi Tabia. **Image Reassembly Combining Deep Learning and Shortest Path Problem**. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018 (see page 32).

[Pur+07] 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 (2007), 559–575 (see pages 86, 122, 125).

[Rad+21] 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) (see page 74).

[Rag+19] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. "Transfusion: Understanding Transfer Learning for Medical Imaging." In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, 3347–3357. URL: http://papers.nips.cc/paper/8596-transfusion-understanding-transfer-learning-for-medical-imaging.pdf (see page 2).

[Raj+17] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, Aarti Bagul, Curtis Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. **CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning**. *CoRR* abs/1711.05225 (2017). arXiv: 1711.05225. URL: http://arxiv.org/abs/1711.05225 (see page 2).

[RBZ22] Giulia Rizzoli, Francesco Barbato, and Pietro Zanuttigh. **Multimodal Semantic Segmentation in Autonomous Driving: A Review of Current Approaches and Future Perspectives**. *Technologies* 10:4 (2022), 90 (see page 109).

[Ree+16] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. **Generative Adversarial Text to Image Synthesis**. In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, June 2016, 1060–1069. URL: http://proceedings.mlr.press/v48/reed16.html (see page 32).

[Rei+01] David E Reich, Michele Cargill, Stacey Bolk, James Ireland, Pardis C Sabeti, Daniel J Richter, Thomas Lavery, Rose Kouyoumjian, Shelli F Farhadian, Ryk Ward, et al. **Linkage disequilibrium in the human genome**. *Nature* 411:6834 (2001), 199–204 (see page 121).

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. **U-Net: Convolutional Networks for Biomedical Image Segmentation**. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Cham: Springer International Publishing, 2015, 234–241. ISBN: 978-3-319-24574-4 (see pages 19, 39, 66, 83, 111, 115, 123).

[RH05] Ralph Rapley and Stuart Harbron. **Molecular analysis and genome discovery**. John Wiley & Sons, 2005 (see page 76).

[RHS05] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. **Semi-supervised self-training of object detection models** (2005) (see page 21).

[RHW88] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. **Learning Internal Representations by Error Propagation**, 673–695. In: Cambridge, MA, USA: MIT Press, 1988. ISBN: 0-262-01097-6. URL: http://dl.acm.org/citation.cfm?id=65669.104449 (see pages 8, 12).

[RM51] Herbert Robbins and Sutton Monro. **A stochastic approximation method**. *The annals of mathematical statistics* (1951), 400–407 (see page 11).

[Roß+17] 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 (Nov. 2017). DOI: 10.1007/s11548-018-1772-0 (see page 24).

[Sah+19] Jaakko Sahlsten, Joel Jaskari, Jyri Kivinen, Lauri Turunen, Esa Jaanio, Kustaa Hietala, and Kimmo Kaski. **Deep Learning Fundus Image Analysis for Diabetic Retinopathy and Macular Edema Grading**. *Scientific Reports* 9 (Dec. 2019). DOI: 10.1038/s41598-019-47181-w (see page 2).

[San+17] Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, and Stephen Gould. **DeepPermNet: Visual Permutation Learning**. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017 (see page 32).

[San+19] Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. **Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks**. *Scientific reports* 9:1 (2019), 1–9 (see page 33).

[Sat+14] 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 (2014), 350–355 (see page 88).

[SBO18] Nawid Sayed, Biagio Brattoli, and Björn Ommer. **Cross and Learn: Cross-Modal Self-Supervision**. *CoRR* abs/1811.03879 (2018). arXiv: 1811.03879. URL: http://arxiv.org/abs/1811.03879 (see pages 32, 74).

[SDN14] Dror Sholomon, Eli David, and Nathan Netanyahu. **A Generalized Genetic Algorithm-Based Solver for Very Large Jigsaw Puzzles of Complex Types**. In: *Proceedings of the National Conference on Artificial Intelligence*. Vol. 4. Jan. 2014, 2839–2845 (see page 32).

[Set+20] Frank Setzer, Katherine Shi, Zhiyang Zhang, Hao Yan, Hyunsoo Yoon, Mel Mupparapu, and Jing Li. **Artificial Intelligence for the Computer-aided Detection of Periapical Lesions in Cone-beam Computed Tomographic Images**. *Journal of Endodontics* 46 (May 2020). DOI: 10.1016/j.joen.2020.03.025 (see page 95).

[SGK17] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. **Learning important features through propagating activation differences**. In: *Int. Conf. on Machine Learning*. PMLR. 2017, 3145–3153 (see page 79).

[SHC14] Kilho Son, James Hays, and David B. Cooper. **Solving Square Jigsaw Puzzles with Loop Constraints**. In: *Computer Vision – ECCV 2014*. Ed. by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars. Cham: Springer International Publishing, 2014, 32–46 (see page 32).

[Sim+19a] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert J. S. Litjens, Bjoern H. Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick Ferdinand Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan Heckers, William R. Jarnagin, Maureen McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. **A large annotated medical image dataset for the development and evaluation of segmentation algorithms**. *CoRR* abs/1902.09063 (2019). arXiv: 1902.09063. URL: http://arxiv.org/abs/1902.09063 (see pages 30, 38).

[Sim+19b] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert J. S. Litjens, Bjoern H. Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick Ferdinand Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan Heckers, William R. Jarnagin, Maureen McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. **A large annotated medical image dataset for the development and evaluation of segmentation algorithms**. *CoRR* abs/1902.09063 (2019). arXiv: 1902.09063. URL: http://arxiv.org/abs/1902.09063 (see page 67).

[Sin64] Richard Sinkhorn. **A relationship between arbitrary positive matrices and doubly stochastic matrices**. *The annals of mathematical statistics* 35:2 (1964), 876–879 (see pages 31, 32).

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. **FaceNet: A unified embedding for face recognition and clustering**. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, 815–823 (see page 64).

[Sof+19] Tamar Sofer, Xiuwen Zheng, Stephanie M Gogarten, Cecelia A Laurie, Kelsey Grinde, John R Shaffer, Dmitry Shungin, Jeffrey R O'Connell, Ramon A Durazo-Arvizo, Laura Raffield, et al. **A fully adjusted two-stage procedure for rank-normalization in genetic association studies**. *Genetic epidemiology* 43:3 (2019), 263–275 (see page 125).

[Son+16] Kilho Son, daniel Moreno, James Hays, and David B. Cooper. **Solving Small-Piece Jigsaw Puzzles by Growing Consensus**. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (see page 32).

[Sor] Sorensen-Dice. *Dice Score*. https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient. Accessed: 2021-11-05 (see pages 83, 124).

[Spi+18] Hannah Spitzer, Kai Kiwitz, Katrin Amunts, Stefan Harmeling, and Timo Dickscheid. **Improving Cytoarchitectonic Segmentation of Human Brain Areas with Self-supervised Siamese Networks**. In: *MICCAI*. Ed. by Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger. Springer International Publishing, 2018, 663–671. ISBN: 978-3-030-00931-1 (see pages 24, 25).

[Sto+15] Marijn F. Stollenga, Wonmin Byeon, Marcus Liwicki, and Jürgen Schmidhuber. **Parallel Multi-Dimensional LSTM, With Application to Fast Biomedical Volumetric Image Segmentation**. *CoRR* abs/1506.07452 (2015). arXiv: 1506.07452. URL: http://arxiv.org/abs/1506.07452 (see page 59).

[STP15] Falk Schwendicke, Markus Tzschoppe, and Sebastian Paris. **Radiographic caries detection: a systematic review and meta-analysis**. *Journal of dentistry* 43:8 (2015), 924–933 (see page 96).

[STY17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. **Axiomatic attribution for deep networks**. In: *Int. Conf. on Machine Learning*. PMLR. 2017, 3319–3328 (see page 79).

[Su+17] Hao Su, Leonidas Guibas, Michael Bronstein, Evangelos Kalogerakis, Jimei Yang, Charles Qi, and Qixing Huang. *3D Deep Learning*. 2017. URL: http://3ddl.stanford.edu/ (see pages 53, 55).

[Sud+15] 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 (Mar. 2015), 1–10. DOI: 10.1371/journal.pmed.1001779. URL: https://doi.org/10.1371/journal.pmed.1001779 (see pages 65, 71, 80, 83, 85, 116, 120).

[Sum+21] Jabeen Summaira, Xi Li, Amin Muhammad Shoib, Songyuan Li, and Jabbar Abdul. **Recent Advances and Trends in Multimodal Deep Learning: A Review**. *arXiv preprint arXiv:2105.11087* (2021) (see page 17).

[Sun+19a] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. **Learning video representations using contrastive bidirectional transformer**. *arXiv preprint arXiv:1906.05743* (2019) (see page 74).

[Sun+19b] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. **Videobert: A joint model for video and language representation learning**. In: *ICCV*. 2019, 7464–7473 (see page 74).

[Sut+18] Yannick Suter, Alain Jungo, Michael Rebsamen, and Mauricio Reyes. **End-to-End Deep Learning versus Classical Regression for Brain Tumor Patient Survival Prediction**. In: *Pre-Conference Proceedings of the 7th MICCAI BraTS Challenge*. Granada, Spain: Springer, 2018 (see pages 44, 45).

[SVL14] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. **Sequence to sequence learning with neural networks**. *Advances in neural information processing systems* 27 (2014) (see page 12).

[SWS05] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. **Early Versus Late Fusion in Semantic Video Analysis**. In: *Proceedings of the 13th Annual ACM International Conference on Multimedia*. MULTIMEDIA '05. Hilton, Singapore: ACM, 2005, 399–402. ISBN: 1-59593-044-2. DOI: 10.1145/1101149.1101236. URL: http://doi.acm.org/10.1145/1101149.1101236 (see page 116).

[SZ14] Karen Simonyan and Andrew Zisserman. **Two-Stream Convolutional Networks for Action Recognition in Videos**. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'14. Montreal, Canada: MIT Press, 2014, 568–576 (see pages 32, 74).

[Szu+21] Joseph D Szustakowski, Suganthi Balasubramanian, Erika Kvikstad, Shareef Khalid, Paola G Bronson, Ariella Sasson, Emily Wong, Daren Liu, J Wade Davis, Carolina Haefliger, et al. **Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank**. *Nature genetics* 53:7 (2021), 942–948 (see page 122).

[Taj+19] 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)*. 2019, 1251–1255 (see page 24).

[Taj+20] 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 (2020), 101693. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2020.101693. URL: http://www.sciencedirect.com/science/article/pii/S136184152030058X (see page 24).

[Tal+19] Aiham Taleb, Christoph Lippert, Moin Nabi, and Tassilo Klein. **Multimodal Self-Supervised Learning for Medical Image Analysis**. In: *Proceedings of 33rd Conference on Neural Information Processing Systems (NeurIPS) - Medical Imaging Workshop*. Vancouver, Canada, 2019 (see page 29).

[Tal+20] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. **3D Self-Supervised Methods for Medical Imaging**. In: *Proceedings of 34rd Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada, 2020 (see pages 53, 118).

[Tal+21] Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. **Multimodal Self-supervised Learning for Medical Image Analysis**. In: *Information Processing in Medical Imaging (IPMI)*. Ed. by Aasa Feragen, Stefan Sommer, Julia Schnabel, and Mads Nielsen. Springer International Publishing, 2021, 661–673 (see pages 29, 30, 33, 34, 37, 47–49).

[Tal+22a] Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. **ContIG: Self-Supervised Multimodal Contrastive Learning for Medical Imaging With Genetics**. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, 20908–20921 (see pages 71, 72, 74, 75, 80, 87, 92, 93, 127).

[Tal+22b] Aiham Taleb, Csaba Rohrer, Benjamin Bergner, Guilherme De Leon, Jonas Almeida Rodrigues, Falk Schwendicke, Christoph Lippert, and Joachim Krois. **Self-Supervised Learning Methods for Label-Efficient Dental Caries Classification**. *Diagnostics* 12:5 (2022). ISSN: 2075-4418. URL: https://www.mdpi.com/2075-4418/12/5/1237 (see pages 22, 95, 96, 99, 101).

[Tan+19] Youbao Tang, Yuxing Tang, Jing Xiao, and Ronald M. Summers. **XLSor: A Robust and Accurate Lung Segmentor on Chest X-Rays Using Criss-Cross Attention and Customized Radiorealistic Abnormalities Generation**. *CoRR* abs/1904.09229 (2019). arXiv: 1904.09229. URL: http://arxiv.org/abs/1904.09229 (see page 33).

[TB19] Hao Tan and Mohit Bansal. **Lxmert: Learning cross-modality encoder representations from transformers**. *arXiv preprint arXiv:1908.07490* (2019) (see page 74).

[TE11] Antonio Torralba and Alexey A. Efros. **Unbiased look at dataset bias**. In: *CVPR 2011*. 2011, 1521–1528 (see page 2).

[ten20] tensorflow.org. *Tensorflow v2.1*. 2020. URL: https://www.tensorflow.org/versions/r2.1/api_docs/python/tf (see page 115).

[TH12] Tijmen Tieleman and Geoffrey Hinton. **Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning**. *COURSERA Neural Networks Mach. Learn* (2012) (see page 11).

[TKI20] Yonglong Tian, Dilip Krishnan, and Phillip Isola. **Contrastive multiview coding**. In: *ECCV*. Springer. 2020, 776–794 (see pages 74, 107).

[Tou+21] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. **Training data-efficient image transformers & distillation through attention**. In: *International Conference on Machine Learning*. PMLR. 2021, 10347–10357 (see page 12).

[Tsa] Sik-Ho Tsang. *SimCLR Review - Medium*. https://sh-tsang.medium.com/review-simclr-a-simple-framework-for-contrastive-learning-of-visual-representations-5de42ba0bc66. Accessed: 2022-07-27 (see page 27).

[Val+18] V. V. Valindria, N. Pawlowski, M. Rajchl, I. Lavdas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker. **Multi-modal Learning from Unpaired Images: Application to Multi-organ Segmentation in CT and MRI**. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018, 547–556 (see page 113).

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. **Attention is all you need**. *Advances in neural information processing systems* 30 (2017) (see page 12).

[Ven+21] 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 (2021), 1–13 (see page 74).

[Ver+21] 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* (2021), 1–14 (see page 71).

[Vin+08] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. **Extracting and composing robust features with denoising autoencoders**. In: *Proceedings of the 25th international conference on Machine learning*. 2008, 1096–1103 (see page 21).

[VKK16] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. **Pixel Recurrent Neural Networks**. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, 2016, 1747–1756 (see page 59).

[Von+18] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. **Tracking Emerges by Colorizing Videos**. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018 (see page 55).

[VPT15] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. **Anticipating the future by watching unlabeled video**. *CoRR* abs/1504.08023 (2015). arXiv: 1504.08023. URL: http://arxiv.org/abs/1504.08023 (see page 55).

[Wai+21] 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* (2021), 588020 (see page 77).

[Wal+21] Tanya Walsh, Richard Macey, Philip Riley, Anne-Marie Glenny, Falk Schwendicke, Helen V Worthington, Janet E Clarkson, David Ricketts, Ting-Li Su, and Anita Sengupta. **Imaging modalities to inform the detection and diagnosis of early caries**. *Cochrane Database of Systematic Reviews*: 3 (2021) (see pages 96, 105).

[Wan+17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. **ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases**. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, 3462–3471 (see page 2).

[Wer88] Paul J Werbos. **Generalization of backpropagation with application to a recurrent gas market model**. *Neural networks* 1:4 (1988), 339–356 (see page 12).

[WG15a] Xiaolong Wang and Abhinav Gupta. **Unsupervised Learning of Visual Representations Using Videos**. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, 2794–2802 (see page 55).

[WG15b] Xiaolong Wang and Abhinav Gupta. **Unsupervised Learning of Visual Representations Using Videos**. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, 2794–2802 (see page 64).

[WGH15] Jacob Walker, Abhinav Gupta, and Martial Hebert. **Dense Optical Flow Prediction from a Static Image**. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, 2443–2451 (see page 55).

[WK] Lilian Weng and Jong Wook Kim. *Self-Supervised Learning: Self-prediction and Contrastive Learning - NeurIPS 2021 Tutorial*. https://nips.cc/media/neurips-2021/Slides/21895.pdf. Accessed: 2022-07-27 (see pages 21, 23).

[Wol+17] Jelmer M. Wolterink, Anna M. Dinkla, Mark H. F. Savenije, Peter R. Seevinck, Cornelis A. T. van den Berg, and Ivana Išgum. **Deep MR to CT Synthesis Using Unpaired Data**. In: *Simulation and Synthesis in Medical Imaging*. Ed. by Sotirios A. Tsaftaris, Ali Gooya, Alejandro F. Frangi, and Jerry L. Prince. Cham: Springer International Publishing, 2017, 14–23. ISBN: 978-3-319-68127-6 (see pages 19, 33).

[Woo+09] Mark W. Woolrich, Saad Jbabdi, Brian Patenaude, Michael Chappell, Salima Makni, Timothy Behrens, Christian Beckmann, Mark Jenkinson, and Stephen M. Smith. **Bayesian analysis of neuroimaging data in FSL**. *NeuroImage* 45:1, Supplement 1 (2009). Mathematics in Brain Imaging, S173–S186. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2008.10.055. URL: http://www.sciencedirect.com/science/article/pii/S1053811908012044 (see page 66).

[Wu+17] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. **Sampling matters in deep embedding learning**. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, 2840–2848 (see page 98).

[Wu+18a] Bichen Wu, Alvin Wan, Xiangyu Yue, Peter Jin, Sicheng Zhao, Noah Golmant, Amir Gholaminejad, Joseph Gonzalez, and Kurt Keutzer. **Shift: A Zero FLOP, Zero Parameter Alternative to Spatial Convolutions**. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018 (see page 109).

[Wu+18b] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. **Unsupervised feature learning via non-parametric instance discrimination**. In: *CVPR*. 2018, 3733–3742 (see page 24).

[Wu+21] Di Wu, Deepti S. Karhade, Malvika Pillai, Min-Zhi Jiang, Le Huang, Gang Li, Hunyong Cho, Jeff Roach, Yun Li, and Kimon Divaris, 163–181. In: *Machine Learning in Dentistry*. Springer International Publishing, 2021. ISBN: 978-3-030-71881-7. DOI: 10.1007/978-3-030-71881-7_13. URL: https://doi.org/10.1007/978-3-030-71881-7_13 (see page 74).

[WVW14] 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 (2014), 765–776 (see page 72).

[Xie+20] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. **Self-training with noisy student improves imagenet classification**. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, 10687–10698 (see page 22).

[Xu+15] 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: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, 2015, 2048–2057 (see pages 32, 74).

[Xu21] Jiashu Xu. **A Review of Self-supervised Learning Methods in the Field of Medical Image Analysis**. *Int. J. Image Graph. Signal Process.(IJIGSP)* 13 (2021), 33–46 (see page 24).

[Yan+10] 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 (2010), 565–569 (see page 76).

[Yan+18] Heran Yang, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Zongben Xu, and Jerry Prince. **Unpaired Brain MR-to-CT Synthesis Using a Structure-Constrained CycleGAN**. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Ed. by Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi. Cham: Springer International Publishing, 2018, 174–182. ISBN: 978-3-030-00889-5 (see pages 19, 33).

[Yan+19] Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam P. Harrison, Mohammadhadi Bagheri, and Ronald M. Summers, 413–435. In: *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*. Cham: Springer International Publishing, 2019. ISBN: 978-3-030-13969-8. DOI: 10.1007/978-3-030-13969-8_20. URL: https://doi.org/10.1007/978-3-030-13969-8_20 (see page 24).

[Yan+20] 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 (2020), 187–214 (see pages 72, 110).

[YBJ18] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. **Multimodal Speech Emotion Recognition Using Audio and Text**. In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. Athens, Greece: IEEE, Dec. 2018, 112–118. DOI: 10.1109/SLT.2018.8639583 (see pages 32, 74).

[Ye+17] 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*. June 2017, 27–28. DOI: 10.31256/HSMR2017.14 (see page 24).

[Yi+17] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. **DualGAN: Unsupervised Dual Learning for Image-to-Image Translation**. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: IEEE, 2017, 2868–2876 (see page 33).

[Yua+21] 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*. 2021, 6995–7004 (see page 74).

[Zbo+21] 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) (see pages 24, 81, 82, 84, 85, 88, 97, 98, 100, 101, 124).

[ZC88] Yi Tao Zhou and Rama Chellappa. **Computation of optical flow using a neural network**. *IEEE 1988 International Conference on Neural Networks* (1988), 71–78 vol.2 (see page 13).

[ZDG19] Martin Zlocha, Qi Dou, and Ben Glocker. **Improving RetinaNet for CT lesion detection with dense masks from weak RECIST labels**. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2019, 402–410 (see page 19).

[ZF13] Matthew D. Zeiler and Rob Fergus. **Visualizing and Understanding Convolutional Networks**. *ArXiv* abs/1311.2901 (2013) (see pages 13, 15).

[Zha+19] Yingying Zhang, Shengsheng Qian, Quan Fang, and Changsheng Xu. **Multi-Modal Knowledge-Aware Hierarchical Attention Network for Explainable Medical Question Answering**. In: *Proceedings of the 27th ACM International Conference on Multimedia*. MM '19. Nice, France: Association for Computing Machinery, 2019, 1089–1097. ISBN: 9781450368896. DOI: 10.1145/3343031.3351033. URL: https://doi.org/10.1145/3343031.3351033 (see page 32).

[Zha+20a] 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) (see page 74).

[Zha+20b] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. **Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding**. *NeurIPS* 33 (2020), 18123–18134 (see page 74).

[Zho+19a] 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 (2019), 1001–1016 (see page 74).

[Zho+19b] 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: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Cham: Springer International Publishing, 2019, 384–393. ISBN: 978-3-030-32251-9 (see pages 25, 40–44, 55).

[Zho+21] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. **A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises**. *Proceedings of the IEEE* 109:5 (2021), 820–838 (see page 71).

[Zhu+17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. **Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks**. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: IEEE, 2017, 2242–2251 (see pages 31, 33, 37, 112).

[Zhu+19] 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: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan. Cham: Springer International Publishing, 2019, 420–428. ISBN: 978-3-030-32251-9 (see pages 25, 40–44, 49, 55).

[Zhu+20a] Jiuwen Zhu, Yuexiang Li, Yifan Hu, Kai Ma, S. Kevin Zhou, and Yefeng Zheng. **Rubik's Cube+: A self-supervised feature learning framework for 3D medical image analysis**. *Medical Image Analysis* 64 (2020), 101746. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2020.101746. URL: http://www.sciencedirect.com/science/article/pii/S1361841520301109 (see page 55).

[Zhu+20b] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. **A comprehensive survey on transfer learning**. *Proceedings of the IEEE* 109:1 (2020), 43–76 (see page 105).

[ZIE16] Richard Zhang, Phillip Isola, and Alexei A. Efros. **Colorful Image Colorization**. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Cham: Springer International Publishing, 2016, 649–666. ISBN: 978-3-319-46487-9 (see pages 21, 23, 26, 83).

[Zit+19] 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 (2019), 71–91 (see page 74).

[Zou+19] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. **A primer on deep learning in genomics**. *Nature genetics* 51:1 (2019), 12–18 (see page 74).

[Zwi+20] 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 (2020), 2295–2312 (see page 72).

[ZWZ17] 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)*. Melbourne, Australia: IEEE, Apr. 2017, 578–582 (see page 24).

[ZYZ18] Zizhao Zhang, Lin Yang, and Yefeng Zheng. **Translating and Segmenting Multimodal Medical Volumes with Cycle- and Shape-Consistency Generative Adversarial Network**. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, United States: IEEE, 2018, 9242–9251 (see pages 19, 33).

List of Figures

List of Tables

List of Publications

Articles in Refereed Journals

[1] **Self-Supervised Learning Methods for Label-Efficient Dental Caries Classification**. *Diagnostics* 12:5 (2022). ISSN: 2075-4418. URL: https://www.mdpi.com/2075-4418/12/5/1237. Joint work with Aiham Taleb, Csaba Rohrer, Benjamin Bergner, Guilherme De Leon, Jonas Almeida Rodrigues, Falk Schwendicke, Christoph Lippert, and Joachim Krois.

Articles in Refereed Conference Proceedings

[2] **Multimodal Self-Supervised Learning for Medical Image Analysis**. In: *Proceedings of 33rd Conference on Neural Information Processing Systems (NeurIPS) - Medical Imaging Workshop*. Vancouver, Canada, 2019. Joint work with Aiham Taleb, Christoph Lippert, Moin Nabi, and Tassilo Klein.

[3] **3D Self-Supervised Methods for Medical Imaging**. In: *Proceedings of 34rd Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada, 2020. Joint work with Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert.

[4] **Multimodal Self-supervised Learning for Medical Image Analysis**. In: *Information Processing in Medical Imaging (IPMI)*. Ed. by Aasa Feragen, Stefan Sommer, Julia Schnabel, and Mads Nielsen. Springer International Publishing, 2021, 661–673. Joint work with Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi.

[5] **ContIG: Self-Supervised Multimodal Contrastive Learning for Medical Imaging With Genetics**. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, 20908–20921. Joint work with Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert.