Advancing Deep Learning in Medical Imaging through Generative Modeling and Representation Learning

Tobias Weber

10.01.2025

Advancing Deep Learning in Medical Imaging through Generative Modeling and Representation Learning

Tobias Weber

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

Eingereicht von
Tobias Weber
München, den 10.01.2025

Erster Berichterstatter: Prof. Dr. Bernd Bischl
Zweiter Berichterstatter: Prof. Dr. Michael Ingrisch
Dritter Berichterstatter: Prof. Dr. Christian Wachinger

Tag der mündlichen Prüfung: 15.04.2025

Acknowledgements

I am deeply grateful to the many individuals whose support, guidance, and encouragement made this dissertation possible. I would like to express my sincere thanks to ...

... *Prof. Dr. David Rügamer, whose unwavering support and infectious drive for academic excellence pushed me to strive for more. David's enthusiasm and belief in my work helped me stay motivated throughout this journey and I am truly fortunate to have had the opportunity to learn from such an academic powerhouse. Thanks for all the advice, discussions, and nightly review sessions.*

... *Prof. Dr. Michael Ingrisch for embracing my experimental and sometimes weird research ideas. Despite having little immediate practical clinical application, he encouraged me to pursue them nevertheless and saw the potential in my work. His willingness to always take time to provide guidance and genuine enthusiasm for clinical research made him an extraordinary mentor.*

... *Prof. Dr. Bernd Bischl for providing the overarching guidance and funding that made this research possible. I am grateful for his role in shaping my understanding of research practice and what it means to be a good researcher. His support laid the foundation for my academic journey, and I am sincerely thankful for his contributions.*

... *Prof. Dr. Christian Wachinger for his willingness to invest his valuable time as the third reviewer of this thesis.*

... *the Clinical Data Science group, who first welcomed me during my master's thesis and continued to provide an incredible environment throughout my PhD. The group's supportive and collaborative spirit made my time here truly special. Sorry for not learning how to play Schafkopf.*

... *the AK7 for the fun times and the great work environment, even though I have mostly been a home office drone throughout my PhD. I have always felt welcomed and I appreciate the camaraderie and emotional support they have provided along the way.*

... *my wife, Laura, who has been a constant source of support throughout this journey. Her patience, encouragement, and belief in me helped me navigate both the challenges and triumphs of this PhD. I couldn't have done it without her.*

Zusammenfassung

In den letzten Jahren hat sich Deep Learning (DL) in vielen Bereichen als bahnbrechend erwiesen, auch im Bereich der medizinischen Bildgebung. Die Anwendung neuronaler Netze und anderer lernfähiger Algorithmen hat sich erheblich auf den medizinischen Bereich ausgewirkt und verspricht, die Diagnosegenauigkeit zu erhöhen, die Ergebnisse für die Patienten zu verbessern und die klinischen Arbeitsabläufe zu optimieren. Das Aufkommen großer Datensätze und Fortschritte bei der Rechenleistung haben die Entwicklung anspruchsvoller DL-Modelle erleichtert, die in der Lage sind, komplexe medizinische Bilder zu analysieren und zu interpretieren. Der Umfang dieser Arbeit konzentriert sich auf einen Teil des gesamten DL-Spektrums, insbesondere auf die aufstrebenden Bereiche der generativen Modellierung und des Representation Learnings, die eng miteinander verflochten sind. Die angefügten Publikationen zielen darauf ab, die Grenzen etablierter medizinischer Bild-DL-Methoden zu erweitern und in experimentellere Forschungsbereiche vorzustoßen.

Die generative Modellierung zielt darauf ab, die Datenverteilung selbst zu erlernen, mit dem Ziel, neue ungesehene Daten zu erzeugen, die eine Vielzahl nützlicher nachgelagerter Aufgaben ermöglichen. Der erste Beitrag dieser Arbeit untersucht das Gebiet der Unterabtastung in der Magnetresonanztomographie (MRI). Anstatt die unterabgetasteten Bilder zu verfeinern, wird direkt eine diskrete Verteilung erlernt, die eine aufgabenspezifische Undersampling-Maske mit einer Sparsity-Beschränkung erzeugt. Im zweiten und dritten Beitrag wird ein Variational Autoencoder (VAE) im Zusammenhang mit der Überlebensanalyse von CT-Scans von Lebermetastasen eingesetzt. Durch die Einbeziehung eines Labels in den rekonstruktionsbasierten VAE wird der latente Raum mit Time-to-Event-Informationen angereichert. Diese Anreicherung bietet ein höheres Maß an Erklärbarkeit und erleichtert die Modellierung von Risiken in Datenproben via gradientengesteuertes Durchqueren des latenten Raums. Der vierte Beitrag invertiert ein vortrainiertes Generative Adversial Network (GAN), um Embeddings von Thorax-Röntgenbildern (CXR) zu erzeugen, die zur Entdeckung zusätzlicher Muster oder zur Modellierung des Krankheitsverlaufs verwendet werden. Aufgrund der unzureichenden Qualität der synthetischen CXR des GANs wird im fünften Beitrag ein neuer, moderner Ansatz für die CXR-Synthese entwickelt. Die Verwendung eines Konglomerats aus mehreren großen Datensätzen und kaskadierten Diffusionsmodellen ermöglicht die Erstellung von hochauflösenden CXRs. Darüber hinaus leistet dieser Beitrag Pionierarbeit bei der textbasierten CXR-Synthese mit Konditionierung auf der Grundlage von Radiologieberichten.

Durch die Manipulation des latenten Raums mittels generativer Modelle sind die bisherigen Beiträge eng mit Representation Learning verbunden. Dieser Bereich zielt auf die Umwandlung von Rohdaten ab, aus denen wichtige Informationen extrahiert, strukturiert und zu einem kompakten Embedding verdichtet werden, die für eine beliebige nachfolgende Aufgabe geeignet ist. Der sechste Beitrag analysiert CXR-Embeddings, die einen inhärenten Bias in Richtung sensibler, geschützter Merkmale enthalten. Mittels Post-hoc-Orthogonalisierung wird die entsprechende Information entfernt, wodurch die Vorhersage nicht mehr möglich ist, die

Klassifikationsleistung aber erhalten bleibt. Der siebte und letzte Beitrag befasst sich mit dem Rechenaufwand für große volumetrische Segmentierungsmodelle. Unter Verwendung der Tucker-Zerlegung werden die Gewichte von 3D Convolutions in kleinere, effiziente Darstellungen faktorisiert. Diese Methode reduziert einen großen Teil der Parameter des Modells, während die Vorhersagequalität der Segmentierung erhalten bleibt.

Summary

In recent years, deep learning (DL) has proven to be a disruptive enabler in many domains, including the realm of medical imaging. The application of neural networks and other learnable algorithms has substantially impacted the medical field, promising to improve diagnostic accuracy, enhance patient outcomes, and streamline clinical workflows. The advent of large-scale datasets and advancements in computational power have facilitated the development of sophisticated DL models capable of analyzing and interpreting complex medical images. The scope of this thesis concentrates on a subset of the full DL spectrum, specifically the uprising areas of generative modeling and representation learning, which are closely interleaved with each other. The proposed contributions aim to push the boundaries of established medical image DL methods, venturing into more experimental research areas.

Generative modeling targets to learn the data distribution itself with the ultimate goal of producing new unseen data samples, enabling a variety of useful downstream tasks. The first contribution of this thesis explores the area of undersampling in magnetic resonance imaging (MRI). Specifically, rather than refining the undersampled images, a discrete distribution that generates a task-specific undersampling mask with a sparsity constraint is learned directly. The second and third contributions utilize a variational autoencoder (VAE) in the context of survival analysis on CT scans of liver metastases. By incorporating a survival objective into the reconstruction-based VAE, the latent space becomes enriched with time-to-event information. This enhancement offers a higher degree of explainability and facilitates counterfactual modeling of hazard in data samples through gradient-guided traversal of the latent space. In contrast, the fourth contribution proposes to invert a pre-trained generative adversarial network (GAN) to produce chest X-ray (CXR) embeddings, which are used to discover additional patterns or model pathology progress. Intrigued by the insufficient quality of synthetic GAN CXR samples, the fifth contribution forms a new state-of-the-art approach for CXR synthesis. Using a conglomerate of multiple large-scale datasets and cascaded diffusion models enables the generation of high-resolution CXRs. Moreover, this work pioneered text-based CXR synthesis with conditioning based on radiology reports.

Through engaging heavily in manipulating the latent space of generative models, the previous contributions are closely related to representation learning. As the name suggests, this area aims to transform raw data, where important information is extracted, structured, and condensed into an often concise embedding fit for an arbitrary downstream task. The sixth contribution analyzes CXR embeddings, which were found to contain inherent biases towards sensitive protected features. Using post-hoc orthogonalization the respective feature information is removed, rendering its prediction infeasible but preserving downstream classification performance. The seventh and last contribution tackles the computational efforts of large volumetric segmentation models. Utilizing Tucker decomposition, 3D convolution weight kernels are factorized into smaller, efficient representations. This method reduces a large fraction of the models' parameters by maintaining its predictive segmentation quality.

LIST OF ABBREVIATIONS

Abbreviation	Meaning
2D	Two-Dimensional
3D	Three-Dimensional
AI	Artificial Intelligence
CT	Computed Tomography
CXR	Chest X-ray
DL	Deep Learning
DM	Diffusion Model
ELBO	Evidence Lower Bound
FLOP	Floating Point Operation
GAN	Generative Adversarial Network
GPU	Graphical Processing Unit
HOSVD	Higher-Order Singular Value Decomposition
HU	Hounsfield Units
KL	Kullback-Leibler
LDM	Latent Diffusion Model
LLM	Large Language Model
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
PCA	Principal Component Analysis
RF	Radiofrequency
RL	Representation Learning
SSL	Self-Supervised Learning
SVD	Singular Value Decomposition
TS	TotalSegmentator
VAE	Variational Autoencoder

LIST OF CONTRIBUTIONS

This cumulative PhD thesis consists of the following contributions:

[*C1*] **Tobias Weber**, Michael Ingrisch, Bernd Bischl, and David Rügamer. "Constrained Probabilistic Mask Learning for Task-specific Undersampled MRI Reconstruction". In: *Proceedings of the IEEE/ CVF Winter Conference on Applications of Computer Vision, WACV*. 2024, pp. 7665–7674

[*C2*] **Tobias Weber**, Michael Ingrisch, Matthias Fabritius, Bernd Bischl, and David Rügamer. "Survival-Oriented Embeddings for Improving Accessibility to Complex Data Structures". In: *Bridging the Gap: From Machine Learning Research to Clinical Practice, NeurIPS Workshops*. 2021

[*C3*] **Tobias Weber**, Michael Ingrisch, Bernd Bischl, and David Rügamer. "Towards Modelling Hazard Factors in Unstructured Data Spaces Using Gradient-Based Latent Interpolation". In: *Deep Generative Models and Downstream Applications, NeurIPS Workshops*. 2021

[*C4*] **Tobias Weber**, Michael Ingrisch, Bernd Bischl, and David Rügamer. "Implicit Embeddings via GAN Inversion for High Resolution Chest Radiographs". In: *Medical Applications with Disentanglements, MICCAI Workshops*. 2022, pp. 22–32

[*C5*] **Tobias Weber**, Michael Ingrisch, Bernd Bischl, and David Rügamer. "Cascaded Latent Diffusion Models for High-Resolution Chest X-ray Synthesis". In: *Advances in Knowledge Discovery and Data Mining: 27th Pacific-Asia Conference, PAKDD*. 2023

[*C6*] **Tobias Weber**, Michael Ingrisch, Bernd Bischl, and David Rügamer. "Post-hoc Orthogonalization for Mitigation of Protected Feature Bias in CXR Embeddings". In: *arXiv preprint arXiv:2311.01349*. 2023

[*C7*] **Tobias Weber**, Jakob Dexl, David Rügamer, and Michael Ingrisch. "Post-Training Network Compression for 3D Medical Image Segmentation: Reducing Computational Efforts via Tucker Decomposition". In: *Radiology: Artificial Intelligence*. Vol. 7. 2. Radiological Society of North America, 2025

Additional Contributions

During this PhD program, the thesis author contributed to several additional publications as a co-author.

[*1*] Anna Theresa Stüber, Stefan Coors, Balthasar Schachtner, **Tobias Weber**, David Rügamer, Andreas Bender, Andreas Mittermeier, Osman Öcal, Max Seidensticker, Jens Ricke, et al. "A comprehensive machine learning benchmark study for radiomics-based survival analysis of CT imaging data in patients with hepatic metastases of CRC". in: *Investigative Radiology* 58.12 (2023), pp. 874–881

[*2*] Ludwig Bothmann, Lisa Wimmer, Omid Charrakh, **Tobias Weber**, Hendrik Edelhoff, Wibke Peters, Hien Nguyen, Caryl Benjamin, and Annette Menzel. "Automated wildlife image classification: An active learning tool for ecological applications". In: *Ecological Informatics* 77 (2023), p. 102231

[*3*] Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, **Tobias Weber**, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, et al. "ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports". In: *European Radiology* 34.5 (2024), pp. 2817–2825

[*4*] David Rügamer, Chris Kolb, **Tobias Weber**, Lucas Kook, and Thomas Nagler. "Generalizing Orthogonalization for Models with Non-linearities". In: *International Conference on Machine Learning, ICML* (2024)

[*5*] Chris Kolb, **Tobias Weber**, Bernd Bischl, and David Rügamer. "Deep Weight Factorization: Sparse Learning Through the Lens of Artificial Symmetries". In: *International Conference on Learning Representations, ICLR* (2025)

Conference Talks

[*1*] **Tobias Weber**, Michael Ingrisch, Bernd Bischl, and David Rügamer. "Exploring Latent Spaces: Manipulating Medical Data Through Image Editing". In: *International Conference on Computational and Methodological Statistics, CMStatistics* (2023)

CONTENTS

CHAPTER 1

INTRODUCTION

> I think if you work as a radiologist, you're like the coyote that's already over the edge of the cliff but hasn't yet looked down, so it doesn't realize there is no ground underneath. People should stop training radiologists now. It's just completely obvious that, within five years, deep learning is going to do better than radiologists.
>
> *Geoffrey Hinton, 2016*

> Will AI ever replace radiologists?
> I say the answer is no — but radiologists who use AI will replace radiologists who don't.
>
> *Curtis Langlotz, 2017*

The presented quotes reflect two diverse and contradictory approaches to the swiftly evolving field of Artificial Intelligence (AI) and its future in radiology. Geoffrey Hinton – often proclaimed as the godfather of AI – foresees a radical shift in modern medicine, where medical professionals are rendered obsolete due to the superiority of deep learning (DL). The qualified physician and professor of radiology Curtis Langlotz offers a more synergistic vision. AI may not serve as a general replacement but rather paves the way for cooperation between human expertise and artificial assistance. These opinions represent two opposing viewpoints. Hinton is manifesting AI as a technical disruptor of the field. In contrast, Langlotz promotes the evolution of the traditional radiologist by appreciating the human factor and enhancing the

role of radiologists. Both quotes state the undeniable presence of AI in the realm of medical imaging and demand a paradigm shift in radiological practice. The presented thesis examines the current status quo of *Deep Learning* (DL) research as well as its application in medical imaging and evaluates how or whether the prognoses of Hinton and Langlotz became a reality.

Already nearly eight years have passed since the release of the statements at the time of writing[1], which resembles an eon in the fast-paced machine learning research environment. General developments in a research field can be identified and analyzed by tracking trends at premier conference venues, e.g. the MICCAI conference (Medical Image Computing and Computer Assisted Intervention) for medical imaging. By investigating the number of MICCAI conference submissions over the years in Figure 1.1, it can be seen that the time points of the quotes mark the start of a nearly exponential growth of submitted research papers, with a small dip during the global pandemic. This reflects a rising and continuing interest of the

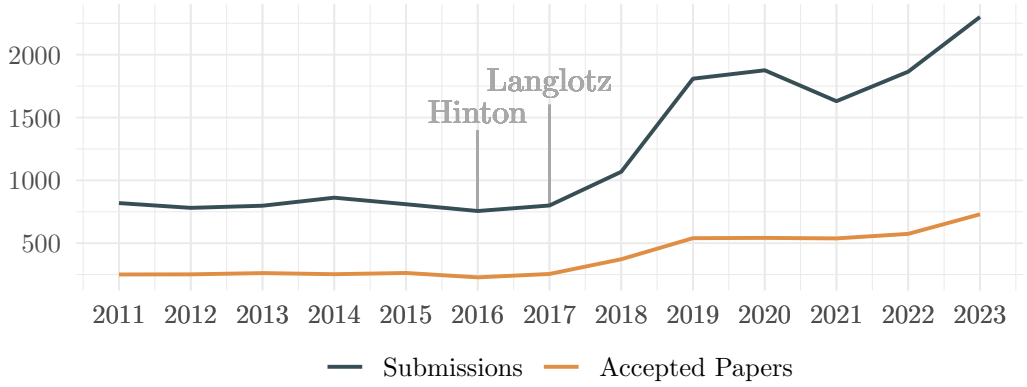


Figure 1.1. Yearly number of submissions and accepted papers for the MICCAI conference from 2011 until 2023. The markers indicate the dates of the analyzed quotes from Hinton and Langlotz.

community to contribute and advance the boundaries of medical imaging techniques. Employing a quick web crawling routine of the MICCAI 2023 proceedings paired with zero-shot GPT-4 classification results in a vague and rough estimate of around ≈97% of the accepted papers utilizing DL techniques or are related to DL. While this is merely a snapshot of the full spectrum of medical imaging, it indicates that DL is a major driver for current scientific and methodological progress. Similar developments can be observed in new journals being issued by medical imaging publishers, which are fully dedicated to AI - often used as a synonym for DL-powered methods. Examples are *Radiology: Artificial Intelligence* by RSNA first issued in 2019 and *BJR|Artificial Intelligence* of the British Institute of Radiology as well as *NEJM AI*, a spin-off from the New England Journal of Medicine, both founded in 2024.

The observed rise in submissions is a mere indicator of the hype around DL in combination with medical imaging. Starting with the increased possibility of utilizing GPUs for the efficient

[1] Hinton: `https://www.youtube.com/watch?v=2HMPRXstSvQ`
Langlotz: `https://aimi.stanford.edu/news/rsna-2017-rads-who-use-ai-will-replace-rads-who-dont-0`

training of convolutional neural networks in the 2010s and the surprising dominance of AlexNet [1] in the 2012 ImageNet [2] challenge, the interest to utilize DL for computer vision surged. Afterward, the basic processing methods evolved, and groundbreaking architectures like the U-Net [3] and ResNet [4] in 2015 built a foundation for a wide range of practical applications including the medical domain. The interest in medical imaging can be attributed to multiple reasons. To name just two, there is the noble intention of improving human life and the fascination with modern medical technology. Additionally, DL algorithms require data and the medical field can in theory provide this data, with some hospitals stockpiling patient records for decades. The amount of available medical data is steadily growing as a result of an increasingly digital clinical practice paired with multi-centric collaborations.

Aside from the developments in medical image research, the possibility of improving clinical practice comes with opportunities for investors following a quote of Walter Wriston: "*Capital goes where it is welcome and stays where it is well treated*". Medical professionals work in a high-stress environment where mistakes can cost human lives, creating a critical need for innovative solutions and advanced technologies. This demand presents a significant potential for the development of new products and advancements in medical imaging, with deep learning as a disruptive technology forming an optimal breeding ground for industrial growth. To name just one of the hundreds of innovating ideas, the Munich start-up *deepc* operates a unified platform for radiological AI, which allows other vendors to promote their commercial models. Already, there are dozens of licensed models available for radiologists to use, including *Chest X-ray* (CXR) classifiers, fracture detection, bone age assessment, analysis of brain scans, and localization of strokes or lung nodules, among others. Productionalization of conceptual research may ultimately be another important factor for the field to mature and establish itself.

An important aspect is the divergence of DL in general computer vision and medical imaging. The differences between those domains are large. Whereas current natural image datasets contain nearly a billion images with captions and myriads of concepts and variation, a medical dataset may contain only a few dozen or hundreds of samples for a rare pathology. Further, in medical imaging, the samples are often 3D images and grayscale instead of RGB. Each medical image modality also presents distinct challenges, such as differentiating tissue densities, managing contrast variations, addressing artifacts, etc. Importantly, a point in a medical image is not a simple pixel value but represents complex physical properties. This raises the question of whether architectures or algorithms designed for natural computer vision can be effectively translated to the medical domain, or what modifications are required to make them usable.

Scope. While the general field of medical imaging is wide, the present thesis focuses on the branch of radiology. As the name suggests, the term *radiology* originates from radiation-based imaging methods but nowadays includes most imaging modalities, such as X-ray imaging, computed tomography (CT), magnetic resonance imaging (MRI), or ultrasound. From the perspective of DL, the focus lies on the subfield of generative modeling and its symbiosis with *Representation Learning* (RL). Figure 1.2 visualizes the specific topics as well as methods grazed in the thesis and indicates the respective contributions. Conceptually, this thesis is on an

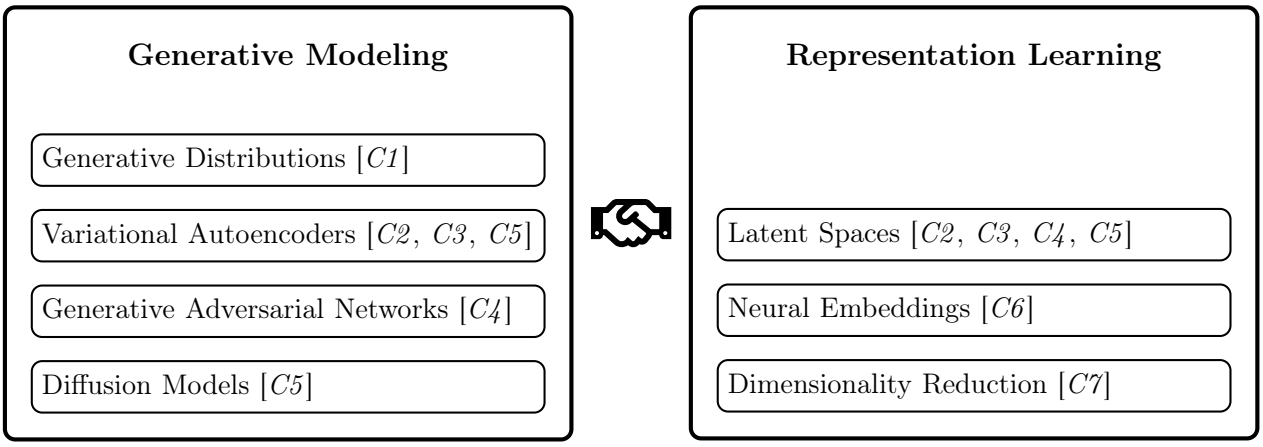


Figure 1.2. Overview over this thesis' contributions and their relation.

explorative level, investigating new experimental domains and searching for a novel perspective on established problems. Thus, its contributions are less concerned with an accurate clinical evaluation and deployment but more with elaborating and exploring the boundaries of current medical image research.

With this motto in mind, the first block of this thesis' contributions concerns a variety of generative models. The first contribution *C1* delves into the problem statement of MRI undersampling. Instead of enhancing an image based on a given undersampling pattern, *C1* proposes a fully differentiable and probabilistic optimization routine capable of directly generating discrete undersampling masks. Given a specific optimization objective, the method produces task- and data-specific undersampling masks, highlighting the benefit of tailoring the acquisition process toward the respective use case. Contribution *C2* and *C3* deviate into the field of survival analysis and employ an extension of a *Variational Autoencoder* (VAE) with an additional supervised survival head. Hereby, the data concerns CTs of colorectal cancer patients with liver metastases. The two contributions discuss the merits of employing such a routine in a clinical context and elaborate on the guided traversal in the VAE's latent space to model survival hazards in unstructured data spaces. Contribution *C4* analyzes whether embeddings of CXR radiographs can be obtained by inverting a *Generative Adversarial Network* (GAN). This analysis discovers some inherent limitations of the used GAN when generating synthetic CXR images. Subsequently, contribution *C5* investigates the efficacy of a cascaded diffusion model to eliminate this issue and improve the quality of the CXR synthesis process. Moreover, contribution *C5* pioneers the utilization of radiological reports for guiding CXR generation.

All previously proposed contributions powered by generative models share a strong connection to RL, as each generative method gains unique benefits when operating in a latent space. The concept of representations is further intensified in contribution *C6*, where orthogonaliza-

tion is used to remove sensible and hidden information from vectorized embeddings of CXR images, providing further insights on biases occurring in CXR classifiers. Lastly, contribution *C7* tackles the problem of complex 3D segmentation models having a high computational effort. Through a post-hoc Tucker decomposition of a model's weights, the method proposed in contribution *C7* substantially reduces the number of arithmetic operations required during inference while preserving the model's predictive accuracy.

Outline. The structure of this thesis is as follows. Chapter 2 contains the methodology foundational to the proposed contributions. Section 2.1 defines the utilized notation. Section 2.2 introduces the unfamiliar reader to the medical imaging modalities of X-ray imaging, CT, and MRI. Next, Section 2.3 handles the topic of generative modeling. Generative distributions are defined in Section 2.3.1 and the methods of VAE (2.3.2), GAN (2.3.3) as well as diffusion models (2.3.4) are elaborated in the subsequent sections. Section 2.4 highlights different aspects of RL. Hereby, Section 2.4.1 showcases the bond of generative models with RL through the concept of latent space. Section 2.4.2 introduces neural feature representations. Section 2.4.3 explains methods for dimensionality reduction via tensor decomposition under the umbrella of RL. Chapter 3 contains the proposed scientific contributions. This thesis concludes with an outlook and conclusion in Chapter 4.

CHAPTER 2

METHODOLOGICAL BACKGROUND

2.1 Notation

A sample $\boldsymbol{x}$ is defined as an element of the data space $\mathcal{X} \subseteq \mathbb{R}^d$ and is generated by the data distribution $p(\boldsymbol{x})$, i.e, $\boldsymbol{x} \sim p(\boldsymbol{x})$. An unsupervised dataset is a collection of multiple samples $\boldsymbol{x}$, organized into a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, where each of the n rows corresponds to one sample $\boldsymbol{x}$. In a supervised learning setting, every $\boldsymbol{x}$ is accompanied by a label, where $y \in \{0, 1\}$ indicates a binary label. It is important to note that the data discussed in this thesis is typically multi-dimensional, such as 2D and 3D images with additional channel information. However, for the purpose of this methodological background, the spatial tensor structure is not relevant. Thus, every multi-dimensional data sample is conceptually unfolded and treated as the vector sample $\boldsymbol{x}$.

Similarly, a latent variable $\boldsymbol{z}$ is denoted as a member of the latent space $\mathcal{Z} \subseteq \mathbb{R}^k$, where $\boldsymbol{z} \sim p(\boldsymbol{z})$. The distributions of $\boldsymbol{x}$ and $\boldsymbol{z}$ can be conditionally dependent on each other. For example, $p(\boldsymbol{z} \mid \boldsymbol{x})$ represents the posterior distribution of $\boldsymbol{z}$ after observing $\boldsymbol{x}$.

Generally, all vectors are column vectors and are denoted in boldface Roman letters. Matrices and Tensors are defined in uppercase boldface Roman letters, e.g., $\boldsymbol{A}$. A superscript $\top$ indicates the transpose of a vector or matrix. $\boldsymbol{I}$ denotes the identity matrix, characterized by 1 entries on the diagonal and 0 elsewhere. The shape of $\boldsymbol{I}$ should be inferred from the context in which it is used. Scalar values, which are elements of $\mathbb{R}$, are denoted using lowercase Greek letters, e.g. α.

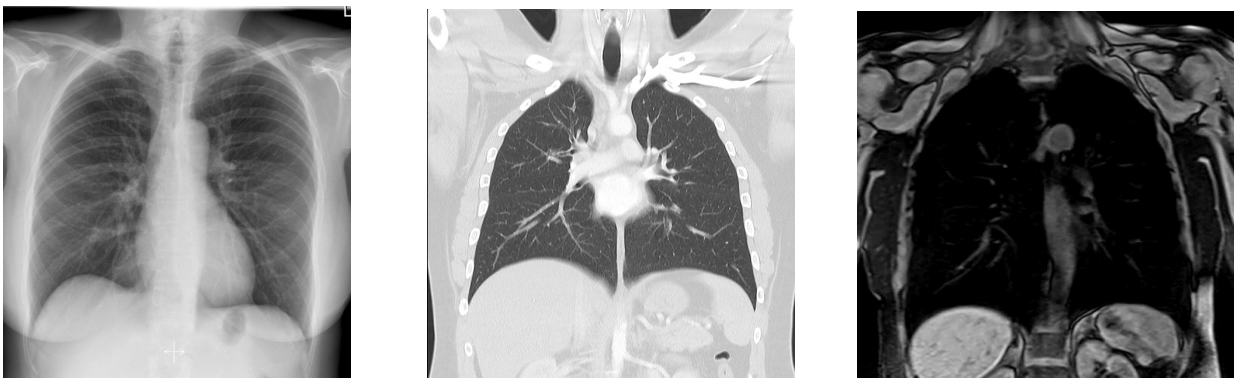
A neural network is defined as a deterministic mapping from one space to another. For example, $\mathcal{E} : \mathcal{X} \to \mathcal{Z}$ denotes a network that projects data samples $\boldsymbol{x}$ into the latent space. Networks are denoted using calligraphic letters, with the exception of the network used in diffusion models, which is denoted as $\boldsymbol{\epsilon}_\theta$ for reasons of consistency.

2.2 A Primer on Medical Imaging

This section provides a brief overview of the medical imaging modalities utilized in this thesis: X-ray Imaging, Computed Tomography (CT) Scan, and Magnetic Resonance Imaging (MRI). As shown in the coronal view of a chest in Figure 2.1 every imaging modality has a distinct display. X-ray imaging is generally the fastest and most cost-effective of the three options. It is commonly used for diagnosing bone fractures, examining dental and chest areas as well as screening purposes due to its ability to produce 2D images quickly. CT and MRI produce volumetric images, capturing data in 3D, resulting in superior image quality but at a higher cost and longer acquisition time in comparison to X-rays. A CT scan is readily available and still magnitudes faster than an MRI and is therefore used for rapid detection in emergencies, like strokes and internal bleeding. Additionally, a CT excels in visualizing bone structures, blood vessels, and a range of lung pathologies and tumors. The main strength of the MRI lies in its high-detail soft tissue and variable image contrasts, produced by variations in the scanning protocol, and its ability to provide fine-grained images of organs, muscles, and other soft tissues without utilizing ionizing radiation.

Note that the following section is intended as a mere high-level introduction to the complex topic of medical imaging and does in no way reflect a holistic description of the modalities' inner workings and applied physics. Thus, the presented information relies on concepts rather than mathematical formulations. The avid reader is referred to "*The Essential Physics of Medical Imaging*" by Bushberg and Boon [5] as well as Buzug's "*Computed Tomography*" [6] for a detailed tutorial. Unless stated otherwise, the subsequent introduction is based on these two recommendations.

The Electromagnetic Spectrum. A method to group different image modalities is by categorizing them according to the energy source of the image. One such source is electromagnetic radiation, whose full range is portrayed in the electromagnetic spectrum (see Figure 2.2).



(a) X-ray. Source: radiopaedia.org **(b)** CT. Source: radiopaedia.org **(c)** MRI. Source: mrimaster.com

Figure 2.1. A coronal view of a chest using the image modalities X-ray, CT, and MRI.

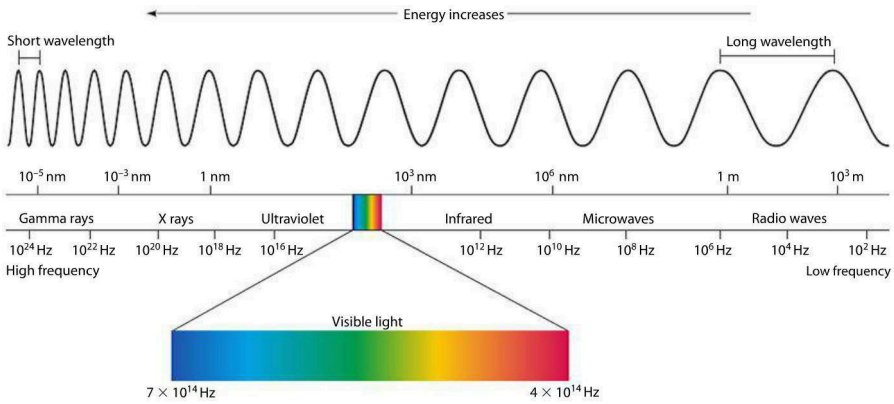


Figure 2.2. The electromagnetic spectrum. Electromagnetic radiation can be categorized based on its wavelength, energy, and frequency. Image from [7].

Radiation can be described either as a propagating sinusoidal wave or as a stream of massless particles traveling in wave-like patterns. Another term for such a particle is *photon*. The categorization of radiation is based on its attributes: its energy measured in electron volts (eV), its frequency measured in Hertz (Hz), or its wavelength measured in meters (m). Hereby, a high energy level coincides with high frequency and short wavelengths, while low energy corresponds to low frequency and long wavelengths.

Each category within the spectrum has specific use cases in computer vision, with some categories being more prominent in medical imaging than others. These categories have no clear boundaries but transition seamlessly into one another. Radio waves, with wavelengths spanning several meters, play a crucial role in the MRI process, even though they are not directly responsible for image creation. The light visible to the human eye is also part of the electromagnetic spectrum, covering wavelengths from 380 to approximately 700 nanometers. Highly energetic waves like Gamma rays and X-rays are particularly suited for medical imaging, due to their ability to penetrate human tissue easily. In contrast, lower energy waves can interact with tissues through heating, similar to the mechanism in a microwave oven. Additionally, high-energy waves, characterized by short wavelengths, exhibit small diffraction that benefits sharp and detailed recordings. For example, gamma rays are fundamental to nuclear imaging, particularly in *Positron Emission Tomography* (PET). In PET, a patient is administered a radioactive isotope (tracer) that emits positrons upon decay. When these positrons encounter electrons, gamma rays are produced and subsequently detected. Tumors can be identified due to their higher metabolic activity, leading to increased tracer consumption and distinctive emission of gamma rays [8]. The principle of X-ray imaging is elaborated in the next section.

Both gamma and X-ray radiation are ionizing, i.e., they have the ability to remove electrons from the atomic shell and thus generate positively charged atoms (ions). This ionization can cause stochastic and deterministic radiation damage, which may ultimately lead to cellular

damage and mutations, which is why the usage of ionizing radiation is highly regulated and conducted with protective measures such as limiting exposure time and shields [9]. Aside from electromagnetic radiation, other noteworthy sources for imaging include sound waves in ultrasound imaging and electron beams in electron microscopy, among others.

X-Ray Imaging. X-ray imaging is one of the oldest and most frequently used medical imaging techniques. In a nutshell, to create an image, a beam of X-rays penetrates the object of interest, and the X-ray attenuation, i.e., the reduction in wave intensity as the X-ray passes through tissue, is measured on a detector. The first step involves generating X-ray radiation using an X-ray tube, as illustrated in Figure 2.3. This vacuum-sealed chamber contains a negatively charged electrode, the cathode, and a positively charged electrode, the anode. To enable electron flow from the cathode to the physically detached anode, the cathode is heated to induce thermionic emission. For medical X-ray imaging, a voltage of 20keV to 150keV is usually applied to accelerate the free electrons [9]. When these electrons collide with the anode, radiation is produced based on two effects. The first effect denotes as *bremsstrahlung*, or braking radiation. If electrons are deflected by the atoms in the element of the anode, their loss of kinetic energy is converted into radiation. The second effect is the characteristic radiation, which is unique to every anode material. This phenomenon describes the emission of radiation that occurs when an inner-shell electron within an atom is ejected upon collision and is subsequently replaced by an electron from a higher energy level. Typical materials for the anode are tungsten, molybdenum, or rhodium, chosen for their extremely high melting points to mitigate damage due to heat. An example of an X-ray spectrum using a tungsten anode is demonstrated in Figure 2.4, which shows the distinct differences between bremsstrahlung and characteristic radiation. To shape the emitted radiation into a beam, the anode is angled. Further, in an effort to reduce the strain on the anode material, the anode is often in the form of a rotating disc.

On its way to the patient, the generated X-ray beam passes through a collimator, also known as a beam-restrictor. These lead shutters concentrate the X-ray beam on an area of interest, reducing the radiation dose and wave scattering. As the X-ray passes through

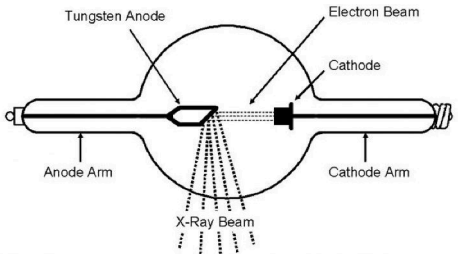


Figure 2.3. Schema of a X-ray tube. The electrons flow from the heated cathode to the angled anode, which upon collision emits a beam of X-rays caused by bremsstrahlung and characteristic radiation. Image from [10].

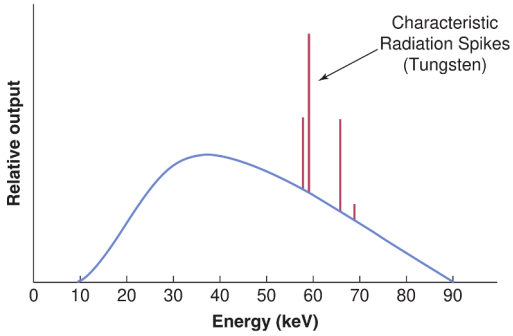


Figure 2.4. Spectrum of a tungsten anode with an applied voltage of 90keV. The unique characteristic radiation spikes (**red**) are superimposed on the bremsstrahlung (**blue**). Image adapted from [5].

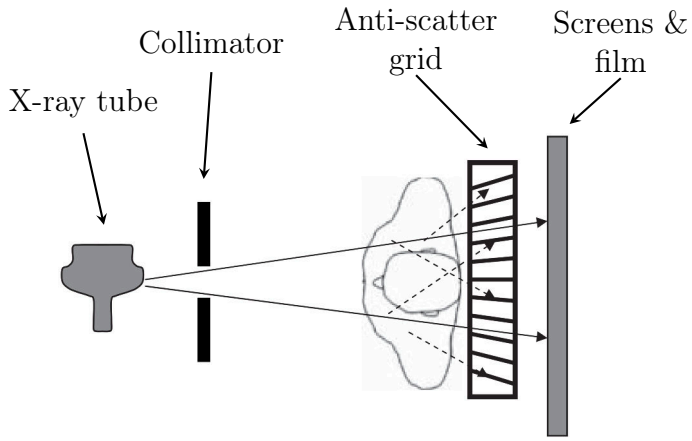


Figure 2.5. The setup for conducting an X-ray scan. The X-ray beam emitted by the X-ray tube is shaped by the collimator before passing through the patient. To mitigate the impact of scattered X-rays, an anti-scatter grid is placed before the detector. In analog X-ray imaging, the detector consists of a film positioned between two intensifying screens. Image adapted from [11].

the patient, it is attenuated depending on the type of tissue it encounters. Dense tissues, like bones, absorb and deflect more radiation than less dense tissues, like organs or muscles. The attenuation for a material or tissue being penetrated can be expressed as a coefficient, whose computation is a combination of various absorption and scattering principles, such as Rayleigh scattering, photoelectric absorption, Compton scattering, and pair production. It is thus possible to determine the type of tissue based on the measured X-ray attenuation. After passing through the patient, an anti-scatter grid mitigates scatter radiation resulting from the interaction with the patient, enhancing the overall image quality. Eventually, the attenuated invisible X-rays are captured by a detecting measure to convert them into a visible image. Figure 2.5 showcases the full acquisition process.

Originally, films contained between two intensifying screens were used. Dense tissues, such

as bones, appear white in the resulting image because these areas have less exposure to X-rays compared to regions with, e.g., air, such as the lungs, which exhibit minimal attenuation and thus strong exposure. Nowadays, digital alternatives are used instead of film. In digital radiography, possible detectors include *photostimulable phosphor plates* (PSPs) and flat panel *Thin-Film-Transistor* (TFT) arrays, among others.

Computed Tomography. Planar X-ray imaging works based on measuring the attenuation of the X-ray radiation along one direction and is thus a 2D projection containing no information about depth. In contrast, CT can capture a volumetric 3D view of the patient, allowing for detailed depictions of internal structures. Like X-ray imaging, CT imaging utilizes X-rays to record an image. Instead of capturing just one scene, CT acquires multiple scans from various angles around the patient. The original CT method segments the patient or the object of interest into axial slices. Hereby, a typical value for the thickness of such a slice is around 1.5 to 3 millimeters. For each slice, an X-ray tube emitting a narrowed-down beam is traversed linearly (translated) across the subject. The attenuation of this so-called pencil beam is then measured by a detection unit. This process is also illustrated in Figure 2.6.

After finishing a full translation, the same measurement procedure is repeated at a different angle, i.e., the X-ray tube and detector are rotated. The number of recordings and the angle for rotation may vary depending on the protocol and used scanner. This translation and rotation process is repeated for every axial slice. Modern CT scanners can obtain more than 1000 measurements over a full 360° rotation [12]. Furthermore, recent technology employs multiple simultaneous beams, cone-shaped beams, or spiral patterns instead of pencil beams to scan

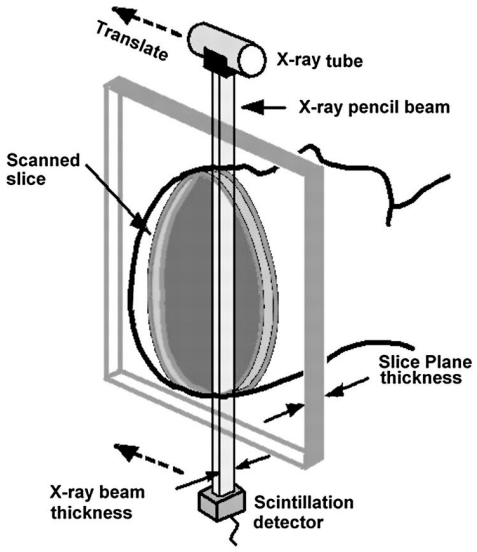


Figure 2.6. Recording of a CT slice using a pencil beam. The collimated beam is translated across the subject, hence creating a narrow axial X-ray scan of the patient. Image from [12].

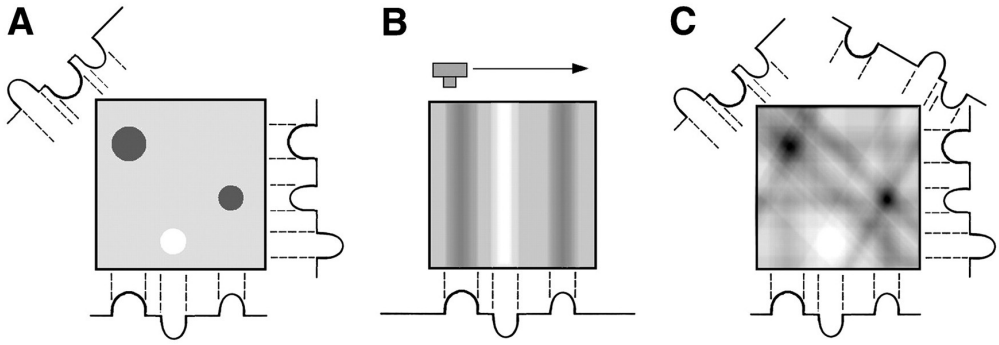


Figure 2.7. The backprojection process for image reconstruction in CT imaging. For every slice, X-ray attenuation values are measured for multiple angles (**A**). The obtained attenuation values are projected backward along the trajectory of the originating X-ray (**B**). This backprojection step is done for every collected angle. Superimposing all backprojections creates a cross-sectional image (**C**). Image from [12].

the patient, enhancing both recording speed and image quality.

The challenging part is now to reconstruct 3D cross-sectional images of the patient from all of the previously obtained measurements. A fundamental concept for this image reconstruction is *backprojection* as demonstrated in Figure 2.7. For every recorded angle, the scan is equivalent to the sum of the X-ray attenuation values along the path of the X-rays, i.e., a forward projection. Next, the measured attenuation values are projected back along the path of the X-ray in image space. All collected angle projections are then superimposed on each other, creating a cross-sectional view. The data from all angles can also be visualized in a plot known as a sinogram. Each row of the sinogram represents the data collected at a specific angle, and each column corresponds to a position during translation.

Unfortunately, utilizing the raw attenuation values for backprojection results in blurry images and distinct artifacts. This is due to backprojection being a global procedure, where contributions to attenuation are projected linearly across the entire image. As each projection is backprojected, the overlapping of data from multiple angles causes blurring because each voxel in the image receives contributions from multiple rays, leading to an accumulation of errors and noise. Convoluting the measurements with a filter prior to backprojection substantially mitigates this issue. Typically, the chosen filters, such as the Ram-Lak filter or the Shepp-Logan filter, aim to emphasize high-frequency features like contours and edges while reducing low-frequency blurring (cf. Figure 2.8). This process is known as *filtered backprojection*. Ultimately, all slices are concatenated to form a full volume of the scanned subject.

Following the process of backprojection, each voxel in the resulting 3D scan is assigned a *CT number*, which represents the average X-ray attenuation value at that specific location in *Hounsfield Units* (HU). Specific HU values indicate the presence of different types of matter. For example, -1000 HU represents air, 0 HU represents water, and values greater than 250

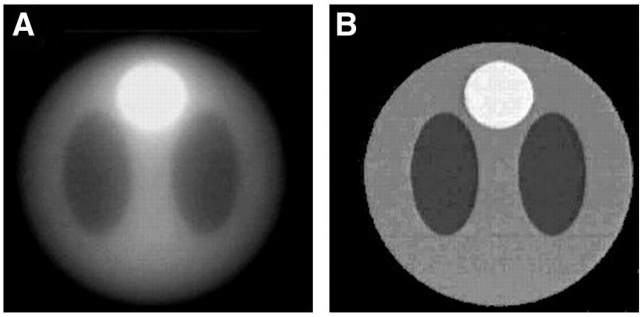


Figure 2.8. Comparison of backprojection without (**A**) and with (**B**) filtering. The applied filter substantially reduces blurring artifacts while emphasizing high-frequency features. Image from [12]

HU represent bone. Soft tissues like fat approximately range from −200 to 50 HU, tumors range from 20 to 50 HU, and the liver has an average value of around 60 HU.

Magnetic Resonance Imaging. While CT allows the recording of high-quality images of a patient, it is still based on ionizing radiation, thus posing risks on increased or multiple usage [9]. On the contrary, MRI observes the behavior of hydrogen atoms in a strong magnetic field without the influence of harmful radiation. As the name suggests, MRI comprises three distinct components, the first of which is *magnetization.* One of the MRI's main components is a strong magnetic field with the strength of 1.5 to 3 Tesla, which is up to 30,000 times stronger than the earth's magnetic field. The MRI's magnetic field is created by passing current through superconducting coils, which exhibit zero electrical resistance when cooled below a critical temperature. Thus, the MRI machine needs a cooling system, typically powered with liquid helium. Given the complexities of establishing the magnetic field with cooled superconductors,

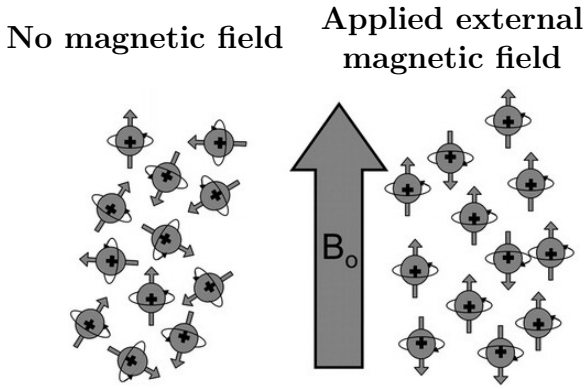


Figure 2.9. Proton alignment in a magnetic field. The orientation of protons is naturally random (**left**). If a magnetic field B_0 is applied, the protons are aligned parallel or anti-parallel to the direction of B_0 (**right**). Image adapted from [13].

the MRI's magnetic field is always on and will never be turned off, requiring constant careful handling of magnetizable objects in the machine's proximity. The magnetic coils are the main sources of the MRI machine's weight, which can be up to 15 tons, setting additional requirements for the room or floor where it is accommodated.

The application of a strong magnetic field, further abbreviated as B_0 has a distinct effect on the atoms of matter in it. It leads to aligning the spins of the nuclei in the atoms, causing precession. The frequency of this spin is known as the *Larmor frequency*, which depends entirely on the strength of B_0 and an atom-specific constant. Hydrogen atoms, which consist of a single proton and an electron, are particularly responsive to this effect. Hydrogen is a fundamental component of the human body, present in nearly all tissues, fats, and fluids. Moreover, B_0 induces the alignment of hydrogen protons either parallel or anti-parallel to the direction of B_0 (cf. Figure 2.9). Notably, there is a slight excess of protons that align parallel to B_0, resulting in a magnetization effect within the matter.

The second part of MRI concerns with the phenomenon of *resonance*. When a radiofrequency (RF) pulse at the Larmor frequency is applied, it provides enough energy to disturb the alignment of hydrogen protons in a small volume – known as the *isochromat* – leading to a net magnetization. As shown in Figure 2.10, the longitudinal magnetization along the direction of B_0 is tipped by the RF pulse into the transverse plane. If the RF pulse is turned off, the magnetization begins to relax and continue its previous alignment with B_0. During this relaxation, the isochromats emit RF waves by themselves, which are detected by a receiver coil. These RF waves contain information about the so-called *T1* and *T2* relaxation times, essential for ultimately constructing an image.

T1 relaxation describes the transition of the transversal magnetization back to the original longitudinal magnetization level, as further illustrated in Figure 2.11. The T1 time is defined as the time needed for the tissue to recover around 63% of its original magnetization along B_0. Each kind of tissue has a characteristic T1 time, where for example fat has a short and fluids have a long T1 time.

Conversely, T2 relaxation describes the decay of the transverse magnetization after the RF pulse. As shown in Figure 2.12 the RF pulse also leads to a temporary magnetization on the

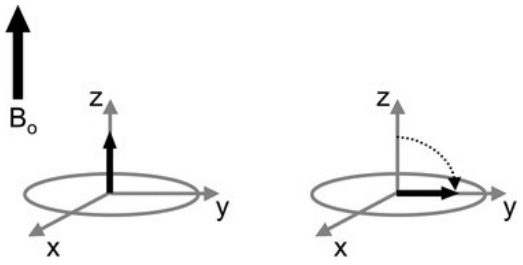


Figure 2.10. Magnetization before (**left**) and after (**right**) the application of an 90° RF pulse. Due to B_0 the hydrogen protons have a longitudinal net magnetization along the z-axis. The RF pulse tips this magnetization into the transverse xy-plane. Image adapted from [13].

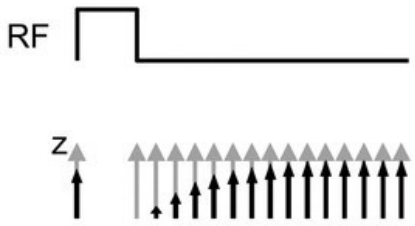


Figure 2.11. Effect of the RF pulse on the longitudinal magnetization. During the activated RF pulse (**top**) the net magnetization along the z-axis is zero (**bottom**). After turning off the signal, the magnetization steadily recovers. Image adapted from [13].

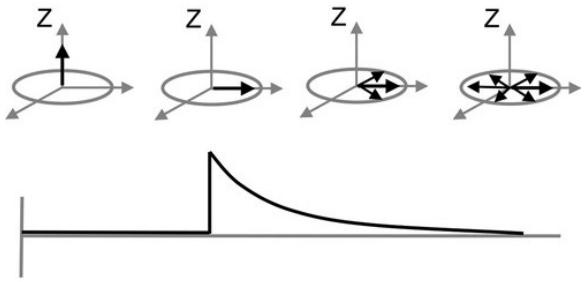


Figure 2.12. Visualization of T2 relaxation. The RF pulse leads to a transverse magnetization, demonstrated in the plot. After turning off the pulse, the protons dephase and the strength of the transverse magnetic field decays. Image adapted from [13].

transverse plane. The T2 time is defined as the time required for the transverse magnetic field to decay to 37% of its initial strength. Typically, fluid-type tissue is equipped with longer T2 times than dense tissues such as bone. Same as in T1, the T2 relaxation can be detected using receiver coils.

The third and last component in MRI is the mechanic of actually transforming the obtained T1 and T2 signals into an *image*. Intuitively, an MRI measures the sources of magnetization and water content within small regions of the subject's body. This process involves the application of gradient fields, known as *gradient coils*, which are responsible for spatial encoding by varying the magnetic field along its axes, which enables the localization of the signal in three dimensions. The acquired signal is stored in the *k-space*, a matrix where each entry represents a specific spatial frequency component of the image. Eventually, the fully sampled k-space matrix can be translated into the image domain using the inverse Fourier transform.

2.3 Generative Modeling

Generative Modeling, as a broad subfield of statistical modeling and machine learning, is equipped with various topics and nuances. The following section provides a detailed introduction to the domain and a general definition of generative models (Section 2.3.1.) Subsequently, the full generative modeling landscape is narrowed down to three foundational algorithms, each representing a different perspective on modeling a target distribution: *Variational Autoencoders* (VAEs; Section 2.3.2), *Generative Adversarial Networks* (GANs; Section 2.3.3), and *Diffusion Models* (DMs; Section 2.3.4).

The presented selection of generative approaches forms a trilemma (cf. Figure 2.13). A VAE is able to generate new samples quickly and with high diversity but lacks sufficient synthesis quality. In contrast, a GAN produces high-quality samples rapidly but is not suited for diverse and multi-modal data. The DM, however, achieves high-fidelity and quality sampling at the cost of a long inference time. In the following sections, a deep dive into these methods examines their nature and the reasons why this trilemma exists. Furthermore, efforts to mitigate specific weaknesses are discussed.

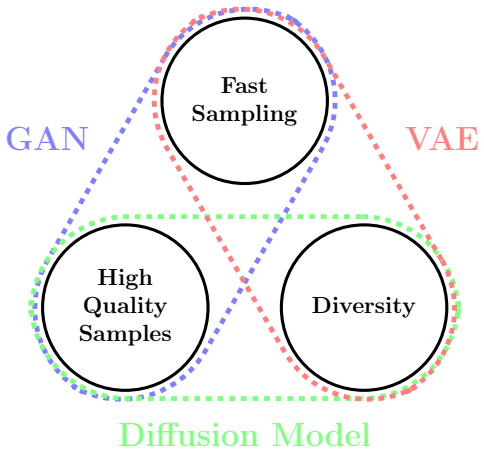


Figure 2.13. The trilemma present in the generative methods of VAE, GAN and DM. Image adapted from [14].

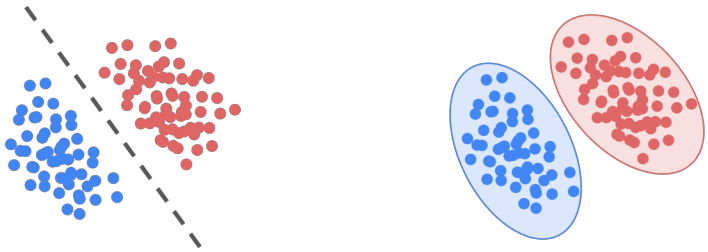
2.3.1 Generative Distributions

Despite being a specific subfield, generative modeling is still a very broad and vague term. The literature provides a variety of definitions that converge in a similar direction, though each offers a nuanced perspective on the topic. For example, *Foster* [15] claims that a generative model is a probabilistic model that describes how a dataset is generated. According to *Bishop* [16] a generative approach implicitly or explicitly models the distribution of inputs $\boldsymbol{x}$ and outputs

y, i.e., the joint distribution $p(\boldsymbol{x}, y)$. These definitions originate from the idea of utilizing generative models for supervised classification.

Discriminative versus Generative Classifiers. *Bishops* definition describes a classic generative classifier. Using the chain rule for $p(\boldsymbol{x}, y) = p(y \mid \boldsymbol{x})p(\boldsymbol{x})$, the conditional $p(y \mid \boldsymbol{x})$ is derived from the joint data distribution allowing the prediction of output targets y. Typical generative classifiers include e.g. Naive Bayes and Gaussian mixture models. In contrast, discriminative classifiers directly learn $p(y \mid \boldsymbol{x})$, equivalent to modeling a decision boundary between classes. A selection of models enforcing this concept includes logistic regression, support vector machines, or decision trees. Figure 2.14 highlights the difference between a discriminative and a generative model.

[17] illustrates one of the decisive advantages of using a generative classifier: given an exemplary inference data point $\boldsymbol{x}^*$ in Figure 2.14, which is situated far to the right of the decision boundary but also distant from the cluster of red samples, the discriminative approach would classify it with high confidence as belonging to the red class. On the contrary, recall that the generative classifier approximates $p(\boldsymbol{x}, y) = p(y \mid \boldsymbol{x})p(\boldsymbol{x})$. While $p(y = \text{red} \mid \boldsymbol{x} = \boldsymbol{x}^*)$ will be very high, reasoned by its distance to the separating hyperplane, this does not necessarily imply the correctness of the decision. The generative model has information about the data distribution itself. Since $\boldsymbol{x}^*$ is distant from the cluster of red samples, $p(\boldsymbol{x} = \boldsymbol{x}^*)$ will be notably low. Thus, $p(\boldsymbol{x} = \boldsymbol{x}^*, y = \text{red})$ is not equipped with high confidence, allowing a measure of uncertainty or quantification of belief in the decision. Depending on the situation, $\boldsymbol{x}^*$ might belong to a newly discovered class or could assigned to be an outlier. However, without considering $p(\boldsymbol{x})$ the discriminative classifier has no utility to give an according estimation, raising the question, of why discriminative models are used at all when generative models give such an advantage. This phenomenon was investigated by e.g. [18], which found, despite the presented toy example, discriminative models have a generally lower asymptotic error than the generative version. Additionally, discriminate models tend to be more data-efficient and do not require modeling an extremely high-dimensional $p(\boldsymbol{x})$ in the case of x being image, video,



(a) A discriminative model. **(b)** A generative model.

Figure 2.14. Conceptual comparison of a discriminative and generative model in a two-class setting. While the discriminative approach targets to differentiate between the two classes, the generative approach approximates the joint data distribution $p(\boldsymbol{x}, y)$.

or audio data [16].

Deep Generative Models. While the above definition of generative models is correct, the recent object of interest has changed. In the earlier days, it was seen primarily as a byproduct but is now a major reason for the widespread adoption of generative models: By learning $p(\boldsymbol{x})$, new, unseen data samples can be generated by sampling from this distribution. This alters the general understanding of the generative framework. For example, *Murphy* [19] recently defined a generative model solely as being the data distribution $p(\boldsymbol{x})$ omitting the need for label information y. Further introduced was the existence of a conditional generative model $p(\boldsymbol{x} \mid \boldsymbol{c})$, where $\boldsymbol{c}$ corresponds to covariates guiding the generative process. Current generative models are predominantly powered by DL, hence the prefix *deep*. Through their capacity, they succeeded in entering complex data regimes with mixed data modalities [20, 21], meeting high publicity and an engaging community. Various algorithms were proposed to create such a deep generative distribution. This thesis focuses on VAEs [22, 23], GANs [24], and DMs [25, 26]. Additional widely adopted methods include normalizing flows [27] and autoregressive plus energy-based models, among others. Each category possesses distinct strengths and weaknesses, influencing both sampling speed and quality.

Contributions

Contribution *C1* is located in the domain of undersampling MRI, which is a method to reduce acquisition time by subsampling the number of obtained data points in k-space. A previously defined pattern - a so-called undersampling mask - is employed to determine the sampling process and acceleration factor. Reducing the number of acquired data results in decreasing image quality, exhibiting blurring and other degradation artifacts. Hereby, a typical use case of DL is to apply image restoration and refinement models to enhance the undersampled images [28, 29, 30, 31, 32]. The approach of contribution *C1* reverses the process: Instead of optimizing images obtained from a pre-determined undersampling mask, the mask itself is optimized yielding customized masks for different downstream tasks and datasets. This perspective is an experimental and nearly unrevealed research area, encompassing only marginal literature [33, 34]. *C1* tackled the problem from a probabilistic point of view, optimizing a generative mask distribution that fulfills a convex sparsity constraint to enforce a desired acceleration factor, hence pushing the boundaries of the generative modeling framework. Despite its discrete nature, through a series of relaxation and reparametrization procedures, the novel optimization process is end-to-end differentiable and model-free. The results indicate that different anatomic regions have distinct optimal undersampling patterns and that visual quality is not the only measure to effectively perform a downstream task, such as segmentation.

2.3.2 Variational Autoencoders

A basic autoencoder can be described as chain of two networks: an encoder and a decoder. The encoder is defined as a mapping $\mathcal{E} : \mathcal{X} \to \mathcal{Z}$ where $\mathcal{X}$ is the data space and $\mathcal{Z}$ is the so-called *latent space*. In contrast, the decoder is a mapping $\mathcal{G} : \mathcal{Z} \to \mathcal{X}$. Note the unusual notation of $\mathcal{G}$, which is intended to emphasize its similarity to the generator used in GANs, as described in detail in Section 2.3.3. The general goal of the full autoencoder is to encode and reconstruct a sample $\boldsymbol{x} \in \mathcal{X}$. An overview of the architecture is displayed in Figure 2.15. For a reconstruction $\hat{\boldsymbol{x}} = \mathcal{G}(\mathcal{E}(\boldsymbol{x}))$ this could be perfectly resolved when choosing the identity function as a mapping for $\mathcal{E}$ and $\mathcal{G}$. However, usually $\dim(\mathcal{Z}) \ll \dim(\mathcal{X})$. In other words, $\mathcal{Z}$ serves as a bottleneck, forcing $\mathcal{E}$ to learn a meaningful representation of $\boldsymbol{x}$ to ensure optimal recovery. For an elaborate introduction to latent spaces in generative models, the avid reader is referred to Section 2.4.1.

The training of an autoencoder is traditionally done via the squared error. Interestingly, when optimizing an autoencoder with the squared error as well as linear mappings $\mathcal{E}$ and $\mathcal{G}$, it is equivalent to performing *Principal Component Analysis* (PCA) [35]. Recent iterations include the application of multiple loss functions simultaneously, including the binary cross-entropy loss, deep perceptual losses [36, 37], and adversarial objectives [38, 39], which has substantially boosted the visual reconstruction quality.

Autoencoder Challenges. Albeit the autoencoder framework is able to obtain good representations from the dataset and reconstructs $\boldsymbol{x}$ samples well, it still has a major downside: In its base formulation, the autoencoder is not yet a generative model. Aside from missing a probabilistic component, the autoencoder is not able or extremely limited to sample $\boldsymbol{x} \sim p(\boldsymbol{x})$ [19]. Generally, in an autoencoder one would sample a $\boldsymbol{z} \in \mathcal{Z}$ and then target to create a new $\boldsymbol{x} \approx \mathcal{G}(\boldsymbol{z})$. However, this is not tractable in most cases, as $\boldsymbol{z}$ does not follow a specific distribution. The model is trained fully unconstrained and thus gaps in the latent space occur. Taking into account the *curse of dimensionality* [40], which refers to the phenomenon where high-dimensional spaces, such as those in image data, are sparse, finding a $\boldsymbol{z}$ that maps to a valid $\boldsymbol{x} \in \mathcal{X}$ is akin to searching for a needle in a haystack. This challenge is addressed by the probabilistic VAE, which imposes the framework of variational inference on the autoencoder.

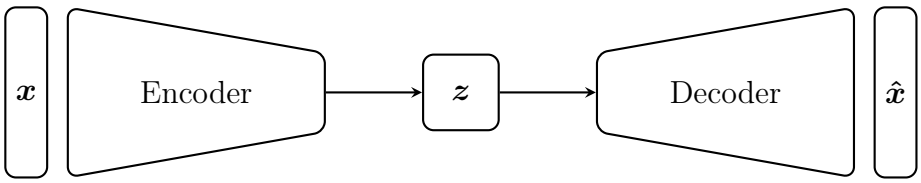


Figure 2.15. Architecture of an autoencoder. The encoder and decoder are connected via a typically low-dimensional bottleneck. The target is to reconstruct $\boldsymbol{x}$, requiring the model to learn a meaningful latent representation $\boldsymbol{z}$.

Variational Inference. Following [41] variational inference is defined as approximating a conditional density of latent variables given observed variables with a variational density. In the present case, variational inference strives to estimate a distribution over $\boldsymbol{z}$ in the latent space after having observed a data sample $\boldsymbol{x}$, i.e., the posterior distribution $p(\boldsymbol{z} \mid \boldsymbol{x})$. The need for a variational framework becomes evident when expanding $p(\boldsymbol{z} \mid \boldsymbol{x})$:

$$p(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{\int p(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z}) \, d\boldsymbol{z}} \,. \tag{2.1}$$

Estimating the posterior requires computing $p(\boldsymbol{x})$, the marginal density or *evidence*. Evaluating the integral within the evidence requires exponential computational effort and is often intractable [41, 16]. As an alternative, $p(\boldsymbol{z} \mid \boldsymbol{x})$ is replaced by an approximate density from a known family – the variational density $q(\boldsymbol{z}|\boldsymbol{x})$. A tractable training objective can be formulated by introducing a variational lower bound on $p(\boldsymbol{x})$ via the approximate $q(\boldsymbol{z} \mid \boldsymbol{x})$:

$$\log p(\boldsymbol{x}) = \log \int q(\boldsymbol{z} \mid \boldsymbol{x}) \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x})} \, dz \geq \mathbb{E}_{q(\boldsymbol{z} \mid \boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x})} \right] , \tag{2.2}$$

which is called the *evidence lower bound* or ELBO. Equation (2.2) allows the application of Jensens' inequality due to the convexity of the logarithm function. The gap between $p(\boldsymbol{x})$ and the ELBO can be explained by the divergence between the true posterior $p(\boldsymbol{z} \mid \boldsymbol{x})$ and the approximate posterior $q(\boldsymbol{z} \mid \boldsymbol{x})$, quantified by the Kullback-Leibler (KL) divergence:

$$\log p(\boldsymbol{x}) = \text{ELBO} + \mathbb{KL}(q(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z} \mid \boldsymbol{x})) \,. \tag{2.3}$$

By treating $p(\boldsymbol{x})$ as a constant with respect to the variational parameters, maximizing the tractable ELBO is equivalent to minimizing the KL divergence between the posteriors. Thus, the equality $\log p(\boldsymbol{x}) = \text{ELBO}$ holds only when $p(\boldsymbol{z} \mid \boldsymbol{x}) = q(\boldsymbol{z} \mid \boldsymbol{x})$.

From Variational Inference to Autoencoders. The previously elaborated variational inference framework is concerned with fitting a latent variable model to estimate $p(\boldsymbol{z} \mid \boldsymbol{x})$. This does not yet directly correspond to a generative model, though. Independently proposed by [22] and [23], the VAE combines variational inference with the architecture of the autoencoder, providing a method to jointly optimize $q(\boldsymbol{z} \mid \boldsymbol{x})$ and $p(\boldsymbol{x} \mid \boldsymbol{z})$. In that adaptation, $q(\boldsymbol{z} \mid \boldsymbol{x})$ is represented by the encoder $\mathcal{E}$, whereas $p(\boldsymbol{x} \mid \boldsymbol{z})$ is realized by the decoder $\mathcal{G}$. Starting from the ELBO, a fitting training objective can be derived by considering that $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z})$:

$$\text{ELBO} = \mathbb{E}_{q(\boldsymbol{z} \mid \boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x})} \right] = \underbrace{\mathbb{E}_{q(\boldsymbol{z} \mid \boldsymbol{x})} \left[\log p(\boldsymbol{x} \mid \boldsymbol{z}) \right]}_{\text{Reconstruction likelihood}} - \underbrace{\mathbb{KL}(q(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z}))}_{\text{Deviation from prior}} \,. \tag{2.4}$$

The first term in the resulting Equation (2.4) corresponds to what is essentially a reconstruction error. The second term denotes the deviation of the variational density from a prior distribution

over the latents $\boldsymbol{z}$. Essentially, penalizing the divergence between $q(\boldsymbol{z} \mid \boldsymbol{x})$ and $p(\boldsymbol{z})$ leads to imposing a constraint on the latent space $\mathcal{Z}$. The prior distribution usually stems from a known and tractable family. Thus, as $p(\boldsymbol{z})$ is aligned with the encoding distribution $q(\boldsymbol{z} \mid \boldsymbol{x})$, sampling $\boldsymbol{z} \sim p(\boldsymbol{z})$ enables the generation of latent points that are mapped by $\mathcal{G}$ to valid data points. This solves the problem of an autoencoder having gaps in its latent space. In the VAE framework, the latent space is now *smooth*, enabling the creation of new data samples and hence functioning as a generative model.

Stochastic Encoding. The definition of $\mathcal{E}$ as a deterministic mapping from the data to the latent space does not meet the probabilistic requirements outlined in the generative model's definition. A proposed solution is to redefine $\mathcal{E}$'s output as the distribution $p(\boldsymbol{z} \mid \boldsymbol{x})$ rather than a single latent data point. While the choice of distribution is arbitrary, an isotropic Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\Sigma = \operatorname{diag}(\boldsymbol{\sigma}^2)$ parametrized by a mean $\boldsymbol{\mu} = (\mu_1, \mu_2, ...\mu_k)^\top$ and variance $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, ...\sigma_k^2)^\top$ is typically utilized. Technically, $\mathcal{E}$ remains deterministic but, instead of predicting $\boldsymbol{z}$ directly, it predicts $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, allowing for subsequent stochastic sampling of $\boldsymbol{z}$. In practice, $\mathbb{E}_{q(\boldsymbol{z} \mid \boldsymbol{x})}$ is approximated via Monte Carlo sampling. This probabilistic perspective introduces the challenge of backpropagating through stochastic nodes during the training of $\mathcal{E}$ and $\mathcal{G}$. The solution involves reparametrization: A noise variable $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, with $\boldsymbol{\epsilon} \in \mathbb{R}^k$ serves as the source of stochasticity. Random sampling is then simulated with $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, which allows unhindered gradient flow.

KL Divergence in Latent Space. Given that $q(\boldsymbol{z} \mid \boldsymbol{x})$ is Gaussian, a natural choice for $p(\boldsymbol{z})$ is also a Gaussian distribution. Following [19], for $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, the divergence between the posterior and prior distribution can be directly estimated:

$$\mathbb{KL}(q(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z})) = \frac{1}{2}\left(-\sum_i^k \log \sigma_i^2 - k + \sum_i^k \sigma_i^2 + \sum_i^k \mu_i^2\right). \tag{2.5}$$

Utilizing VAEs as a generative model has a distinct downside: Due to the Gaussian prior, the reconstructed and generated outputs tend to be considerably blurry (c.f. Figure 2.16). In its extreme form, a phenomenon known as *mode collapse* occurs when $q(\boldsymbol{z} \mid \boldsymbol{x}) \approx p(\boldsymbol{z})$ due

Figure 2.16. Samples from a VAE trained on CelebA. The generative model exhibits severe blurriness and a lack of high-frequency details due to the strong regularization of $q(\boldsymbol{z} \mid \boldsymbol{x})$. Image from [42].

to excessively strong penalization. Here, the model completely ignores the covariates $\boldsymbol{x}$ and produces only a single outcome – the mean of the dataset. A simple but powerful solution is achieved by balancing the prior penalization term with a parameter β [43]:

$$\mathbb{E}_{q(\boldsymbol{z} \mid \boldsymbol{x})} [\log p(\boldsymbol{x} \mid \boldsymbol{z})] - \beta \mathbb{KL}(q(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z})) \,. \tag{2.6}$$

This formulation is generally denoted as β-*VAE*. The choice of β is a trade-off in itself. Choosing a large $\beta > 1$ results in strong latent representations with high disentanglement, but visual reconstruction quality plummets substantially. Modern approaches that focus on high-quality image outputs, such as the VAE backbone in Stable Diffusion [39], use β values of considerably smaller magnitudes (e.g., 1e−6). In these cases, the smoothness of the latent space is less important than image quality and is offset by the sheer amount of training data.

Contributions

The VAE framework is the foundation for contributions *C2* and *C3*. In its base formulation, the model is trained in an unsupervised fashion, with the latent space being formed by distinct image features. The Gaussian assumptions in the VAE tend to disregard high-frequency features. However, in the present case of CT scans with liver metastases, the object of interest is precisely such a small-detail feature. To counter this aspect, an additional head is added to the latent space, which is not optimized for visual reconstruction quality but via a supervised survival loss. This newly added head aids in extracting the information necessary for subsequent survival analysis by altering the vanilla VAE definition.
In contribution *C5* a VAE serves as a backbone in the cascaded diffusion architecture for synthesizing high-resolution CXRs. More precisely, the VAE acts as a bridge from the generated latent points to the data space before the sample is upscaled by a super-resolution diffusion model.

2.3.3 Generative Adversarial Networks

Generative Adversarial Networks (GANs), proposed by [24], mark another important milestone in developing deep generative models. In contrast to the VAE framework, the likelihood-free GAN does not have the inherent limitation of blurriness in the synthesized outputs.

Adversarial Objective. The architecture of the GAN, as illustrated in Figure 2.17, consists of two models: A generator $\mathcal{G}$ and a discriminator $\mathcal{D}$. The definition of the generator aligns with that of the encoding network in the autoencoder, as both are mappings, denoted $\mathcal{G} : \mathcal{Z} \to \mathcal{X}$. New samples $\hat{\boldsymbol{x}}$ can be generated by sampling $\boldsymbol{z}$ from a predefined prior $p(\boldsymbol{z})$, usually $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. On its own, $\mathcal{G}$ has no tractable solution for training to assimilate the data distribution. A

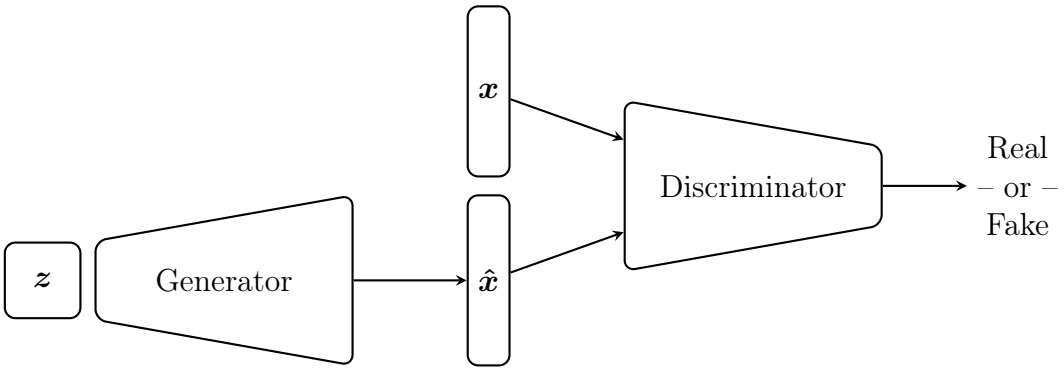


Figure 2.17. Architecture of a GAN. The generator and discriminator are engaged in an adversarial objective. The discriminator aims to differentiate between real and synthesized samples, while the generator is optimized to deceive the discriminator.

fitting signal is produced by engaging a discriminator $\mathcal{D}$, defined as a mapping $\mathcal{D} : \mathcal{X} \to [0, 1]$ [44]. The discriminator's role is to distinguish between real samples from the data distribution p_{data} and fake samples from the generator distribution $p_{\mathcal{G}}$. During training, $\mathcal{G}$ and $\mathcal{D}$ compete in a two-player minimax game. $\mathcal{D}$ aims to maximize the probability of correctly identifying real and fake samples, while $\mathcal{G}$ seeks to minimize this probability. This relationship can be expressed in a value function

$$\min_{\mathcal{G}} \max_{\mathcal{D}} V(\mathcal{G}, \mathcal{D}) = \mathbb{E}_{p_{\text{data}}(\boldsymbol{x})}[\log \mathcal{D}(\boldsymbol{x})] + \mathbb{E}_{p(\boldsymbol{z})}[\log(1 - \mathcal{D}(\mathcal{G}(\boldsymbol{z})))] \,. \tag{2.7}$$

By maximizing V with respect to $\mathcal{D}$, the optimum value of 0 is achieved when $\mathcal{D}$ correctly predicts all true samples $\boldsymbol{x} \sim p_{\text{data}}$ with a 1 and all fake samples $\hat{\boldsymbol{x}} \sim p_{\mathcal{G}}$ with a 0. In contrast, $\mathcal{G}$ can only influence the second term, which, when minimized, incentivizes the creation of samples that the discriminator cannot distinguish from real data. This effectively pushes $\mathcal{G}$ to imitate the underlying data distribution p_{data}. The two models are trained in an alternating fashion. The key is to find an optimal balance, as a good result can only be achieved when one model does not overpower the other.

Optimization Challenges. The first difficulty arises when initializing the training. It is extremely easy for $\mathcal{D}$ to differentiate between real and fake samples, as $\mathcal{G}$ has received few or no weight updates, resulting in largely disjoint distributions. In this scenario, $\mathcal{G}$ suffers from poor gradients. This issue can be mitigated by changing the objective for $\mathcal{G}$ from minimizing $\log(1 - \mathcal{D}(\mathcal{G}(\boldsymbol{z})))$ to maximizing $\mathcal{D}(\mathcal{G}(\boldsymbol{z}))$. Additionally, it has been proven beneficial for $\mathcal{D}$ to receive multiple weight updates before each iteration of $\mathcal{G}$ [24]. These adjustments provide an initial improvement to the original objective. Nonetheless, training a GAN remains an unstable task, and many further improvements were proposed to alleviate the optimization process.

Further Developments. The GAN is a popular and widely studied subject with a myriad of research literature. This section highlights key contributions that have substantially impacted the development of GANs. The GAN itself is an unsupervised algorithm, and sampling is unconditional. [45] introduced the ability to steer the generation process by providing $\mathcal{G}$ and $\mathcal{D}$ access to label information. The GAN is still subject to severe mode collapse. Thus a lot of effort has gone into stabilizing the training process. For example, [46] proposed to replace the sigmoidal activation in $\mathcal{D}$ with a real-valued output and the cross-entropy error function with a squared error function. [47] discovered that optimizing the *earth mover's* or *Wasserstein* distance, rather than the original GAN objective, results in much more stable training. However, the formulation of the *Wasserstein GAN* (WGAN) requires $\mathcal{D}$ to satisfy Lipschitz continuity. WGAN enforces this requirement by limiting $\mathcal{D}$ via weight-clipping, which "*is a clearly terrible way to enforce a Lipschitz constraint*" [47]. WGAN-GP [48] enhances this approach by incorporating a gradient penalty, which penalizes $\mathcal{D}$ when the gradient norm deviates from 1. Another regularization technique is spectral normalization [49], which scales the weights of $\mathcal{D}$ by their largest singular value. GANs traditionally require massive amounts of training data. By applying augmentation to both real and fake images, GANs can be trained with significantly less data [50].

For the domain of imaging, DCGAN [51] introduced architectural changes to convolutional layers and compiled best training practices. Training a GAN to produce images becomes increasingly challenging as the target resolution increases. The *Progressive Growing GAN* (PGAN) [52] addresses this issue by employing an iteratively expanding architecture. This approach begins optimization at a low resolution and incrementally extends it until the target resolution is achieved. A major milestone is the introduction of *StyleGAN* [53]. Instead of directly feeding the seed $\boldsymbol{z}$ to $\mathcal{G}$, StyleGAN applies a mapping network for a non-linear projection of $\boldsymbol{z}$, which is subsequently passed to all layer blocks of $\mathcal{G}$ via adaptive instance normalization [54]. [53] found that earlier layers encode coarse features while later layers capture fine-grained image details, enabling the combination and style-based manipulation of the generated images by injecting modified latent code. Its successor, *StyleGAN2* [55], improved upon this architecture by further refining the generator architecture and utilizing path length regularization that benefits a well-behaved latent space. In the next iteration, *StyleGAN3*, [56] argued that the StyleGAN synthesis process relies too much on absolute pixel coordinates. This dependency causes objects (e.g., eyes or teeth) to be fixed to specific pixel locations, resulting in unusual effects during interpolation. By treating signals in the network as continuous, the model became equivariant to translation and rotation, effectively resolving this issue. In recent years, the popularity of GANs has diminished slightly with the rise of diffusion models. Current approaches strive to demonstrate the competitiveness of GANs in comparison to diffusion models [57, 58].

GAN Inversion. The latent space $\mathcal{Z}$ implicitly formed by a GAN has some intriguing properties. For examples, if the samples $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{Z}$ are in close distance, the synthesized images $\mathcal{G}(\boldsymbol{z}_1), \mathcal{G}(\boldsymbol{z}_2) \in \mathcal{X}$ are similar as well [59]. Further, $\mathcal{Z}$ encodes rich semantic features

of $\mathcal{X}$ [60, 53, 61, 62]. $\mathcal{G}$ provides a one-way mapping from the latent to data space but the reverse direction is not possible in the GAN base architecture. Building a connection $\mathcal{X} \to \mathcal{Z}$ is referred to as *GAN inversion*. This objective concerns finding a latent point

$$\boldsymbol{z}^* = \arg\min_{\boldsymbol{z} \in \mathcal{Z}} \mathcal{L}(\mathcal{G}(\boldsymbol{z}), \boldsymbol{x}) \,, \tag{2.8}$$

where $\mathcal{L}$ defines an arbitrary similarity criterion. Methods for performing this inversion can be categorized into three groups [59, 63]:

- **Learning-based:** An encoder is trained to project $\boldsymbol{x}$ into $\mathcal{Z}$.
- **Optimization-based:** Optimizes Equation (2.8) directly by backpropagation.
- **Hybrid:** Engages an encoder to initialize an optimization-based approach.

Learning-based approaches require training an additional network, but inference is done with a single forward pass. Hence, this allows for the rapid inversion of large data sets, but the encoder's approximation of $\boldsymbol{z}^*$ is notorious for neglecting small details. The optimization-based approach is more precise, but it requires up to thousands of iterations to converge for a single sample, making it very slow. Hybrid methods combine the best of both approaches. By using the encoded $\boldsymbol{x}$ in the proximity of $\boldsymbol{z}^*$ as a starting point, they substantially accelerate the optimization process.

The previously described categories solve the inversion problem post-hoc. Outside of this taxonomy, approaches such as [64, 65] jointly optimize a VAE and GAN to directly provide a mapping to the latent space. Among other domains, the application of GAN inversion is an important backbone in adversarial anomaly detection (e.g. [63, 66, 67]). Here, the foundational assumption is that anomalous samples cannot be reconstructed by a $\mathcal{G}$ trained solely on regular data. The loss criterion $\mathcal{L}$ should then exceed a certain threshold, indicating the presence of an anomaly.

Image Editing. The release of StyleGAN has promoted a research movement towards utilizing GAN inversion for image editing, dominantly in the area of facial image manipulation. Notable methods for inverting the StyleGAN encoder include *encoder4editing* [36], *pixel2style2pixel* [62] and *ReStyle* [69]. The *HyperInverter* [70] does not only apply an encoder, but an additional auxiliary network that manipulates $\mathcal{G}$ to produce the target image.

After obtaining a latent point $\boldsymbol{z}^*$ where $\mathcal{G}(\boldsymbol{z}^*) \approx \boldsymbol{x}$, the original image itself can now be manipulated by transforming $\boldsymbol{z}^*$ and projecting the result back to $\mathcal{X}$ using $\mathcal{G}$. A simple transformation can be, for example, a linear interpolation between two samples. Due to the smoothness in $\mathcal{Z}$, every change in $\boldsymbol{z}^*$ translates to a semantic variation in the typically sparse $\mathcal{X}$ in image space. In practice, the Gaussian $p(\boldsymbol{z})$ is still subject to the *curse of dimensionality* [40] such that most of its probability mass lies on a hypersphere. To account for this aspect, spherical linear interpolation (SLERP) [71] is commonly used to enhance the quality of the reconstructed images during interpolation.

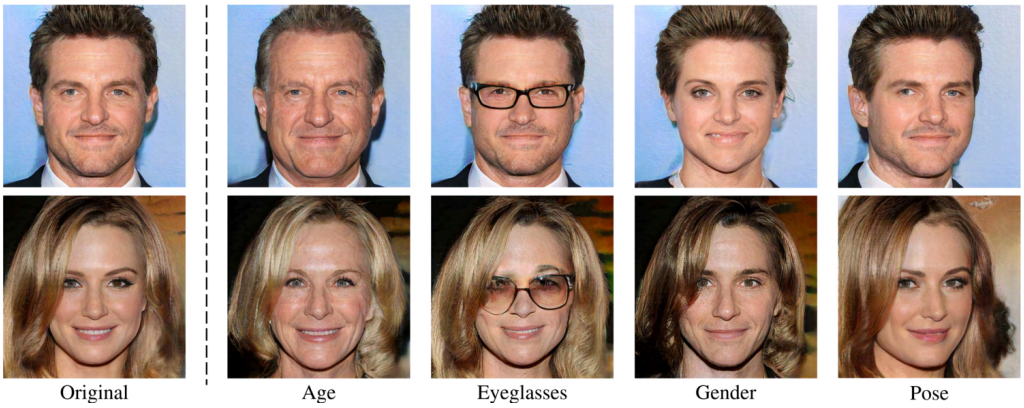


Figure 2.18. Manipulation of real images using the *InterFaceGAN* method on a GAN trained on the CelebA dataset. Image from [68].

Of particular interest are guided transformations based on target attributes (c.f. Figure 2.18). The attributes that match the subject of interest are often part of a larger dataset like in *CelebA* [72]. *InterFaceGAN* [68] is a simple but effective method for image manipulation: When interpolating between two latent samples with different attributes (e.g., *glasses* and *no glasses*), there must be a boundary where one attribute transitions into the other, i.e., the attribute turns into the opposite when crossing the boundary. This boundary, or separating hyperplane, is created by inverting a labeled dataset and then training a supervised classifier on the latent codes. In the case of InterFaceGAN a linear boundary is estimated by a support vector machine [73]. Its weights form a normal vector $\boldsymbol{n}$ perpendicular to the separating hyperplane between two attributes. Thus, one can manipulate a latent sample $\boldsymbol{z}$ using $\boldsymbol{z} + \alpha\boldsymbol{n}$ where $\boldsymbol{n}$ serves as an interpolation axis and α defines the step size. Choosing $\alpha > 0$ will leverage increasing semantic features of the positive class and vice versa. While the method is targeted towards facial editing, its formulation allows for general application. A similar interpolation technique is portrayed in [60] in an attempt to quantify memorability, aesthetics, and emotional valence. A different procedure that does not require inverting the GAN is *GANSpace* [61], which identifies latent directions via PCA to steer the synthesis process.

Contributions

Contribution *C4* builds on the PGAN generator of [74], which evaluate the clinical realism of synthetic CXR. *C4* analyzes whether post-hoc GAN inversion can be effectively applied to high-resolution CXR, providing several key contributions. Firstly, *C4* proposes a novel multi-stage hybrid approach for inverting GANs, enabling the mapping of CXR into the latent space of the generator. This approach includes bootstrapped pretraining to align the encoder directly to the generator's distribution, followed by fine-tuning with real data and iterative optimization to enhance inversion quality. Secondly, the study demonstrates that $\mathcal{Z}$ encodes semantically meaningful features of CXR, which allows for various applications such as image compression, guided image manipulation, and the creation of stylized samples, Moreover, GAN inversion enables the application of generative techniques to actual patient images. The findings conclude that the quality of the inversion process is upper-bounded by the capacity of the utilized generator. This manifests in inverted samples lacking details like medical devices and annotation, which the used $\mathcal{G}$ network cannot produce.

2.3.4 Diffusion Models

Despite the difficulty in training GANs, they were considered the state-of-the-art method in generative modeling until the advent of *Diffusion Models* (DMs). Propelled by papers with compelling titles such as "*Diffusion Models Beat GANs on Image Synthesis*" [75] and groundbreaking open-source models, the attention shifted and persisted until the time of writing this section. Interestingly enough, the original idea of DM stems from the field of thermodynamics [25]. The architecture and training objective were refined by [26] creating a fully-fledged generative model capable of high-quality content synthesis.

The upcoming introduction to DMs follows the definition of *Denoising Diffusion Probabilistic Models* (DDPM) [26] and the tutorial of [76]. Alternatively, DMs can be defined equivalently from a score-matching perspective [77, 78], which is not covered in this thesis. On a conceptual level, as shown in Figure 2.19, the DM forms a Markov chain, where states are connected through a series of transitions that gradually add (forward diffusion process q) or remove noise (reverse diffusion process p_θ). Note that the definitions of q, p_θ and some other notations deviate from the general notation used in this thesis but are retained to maintain consistency with the standard terminology in DM literature.

Forward Diffusion. The forward diffusion process q intuitively describes the corruption of a data sample $\boldsymbol{x}_0 \sim q(\boldsymbol{x})$ via some noise distribution, which is chosen to be Gaussian in the DM base formulation. The sample $\boldsymbol{x}_0$ is corrupted over T timesteps. According to the Markovian property, the next state depends solely on the present state and not the preceding sequence.

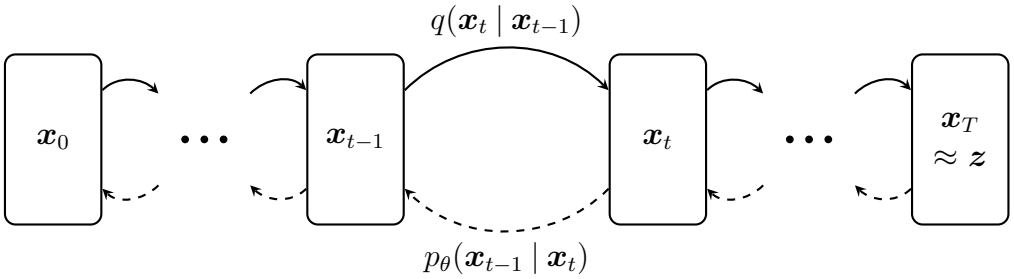


Figure 2.19. Architecture of a diffusion model. In the forward diffusion process q, data $\boldsymbol{x}_0 \sim q(\boldsymbol{x})$ is progressively noised until it matches a simple noise distribution, e.g. $\boldsymbol{x}_T \approx \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The reverse diffusion process p_θ transforms noise back into complex data through a series of denoising steps, each refining the generated sample.

Thus, for sampling an arbitrary timestep t with $0 < t \leq T$ the transition

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I})\,. \tag{2.9}$$

The step size between two states is moderated by a variance β_t that is part of a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^T$. Practically, T ranges in the magnitudes of thousands, whereas β is chosen to be small. For example, [26] employ $T = 1000$ with a linear interpolation from $\beta_1 = 1\text{e}{-4}$ to $\beta_T = 0.02$. The difference between two neighboring states is marginal, indicating a consistent and seamless transformation where increased corruption is applied to larger T. Eventually, the final $\boldsymbol{x}_T$ resembles an isotropic Gaussian distribution. Having declared the timesteps and variance schedule, the trajectory towards $\boldsymbol{x}_T$ starting from an uncorrupted sample $\boldsymbol{x}_0$ is defined as

$$q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})\,. \tag{2.10}$$

However, reaching T implies that every state in this trajectory must be evaluated, which is computationally intensive and time-consuming. This limitation can be circumvented by applying a reparametrization trick similar to the one used in VAEs, allowing sampling an arbitrary timestep t in a tractable closed-form. With the substitution $\alpha_t = 1 - \beta_t$ the computation of $\boldsymbol{x}_t$ can be recursively unfolded as follows:

$$\begin{aligned}
\boldsymbol{x}_t &= \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\boldsymbol{\epsilon} \\
&= ... \\
&= \sqrt{\bar{\alpha}}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}}\boldsymbol{\epsilon}
\end{aligned}$$

Here, $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $\bar{\alpha} = \prod_{i=1}^{t} \alpha_i$. This formulation is one of the crucial success factors of DMs, as now $q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}}\boldsymbol{x}_0, \sqrt{1-\bar{\alpha}}\boldsymbol{I})$ allows to create noised training data for all time steps in a single computation.

Reverse Diffusion. Adding noise to data samples is a trivial procedure, whereas this could not be stated about the opposite. Sampling $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, reversing the diffusion process, and thereby generating new data is somewhat more challenging. This starts by the true posterior $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ being intractable. It is thus replaced with an approximate distribution

$$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)) \tag{2.11}$$

parametrized by θ. The trajectory of deriving $\boldsymbol{x}_0$ from $\boldsymbol{x}_T$ is then defined as

$$p_\theta(\boldsymbol{x}_{0:T}) = p_\theta(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)\,. \tag{2.12}$$

Training Objective. A suitable training objective can be formulated over the evidence lower bound (ELBO) similar to the VAE. Intuitively, the approaches of VAE and DM have substantial parallels. Encoding data corresponds to transforming the data into noise, while the denoising aspect is akin to decoding. Following [79], the ELBO can be decomposed as:

$$\begin{aligned}
\log p_\theta(\boldsymbol{x}) &\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}\left[\log p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_{1:T})\right] - \mathbb{KL}(q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0) \parallel p_\theta(\boldsymbol{x}_{1:T})) \\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1 \mid \boldsymbol{x}_0)}\left[\log p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)\right]}_{\mathcal{L}_0} - \underbrace{\mathbb{KL}(q(\boldsymbol{x}_T \mid \boldsymbol{x}_0) \parallel p_\theta(\boldsymbol{x}_T))}_{\mathcal{L}_T} \\
&\qquad - \sum_{t=2}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)}\left[\mathbb{KL}(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t))\right]}_{\mathcal{L}_{t-1}}\,.
\end{aligned} \tag{2.13}$$

The results contain three distinct types of terms, each requiring a different mode of interpretation. Similar as in the VAE objective, the first term, $\mathcal{L}_0$, can be interpreted as a type of reconstruction error when traversing from $\boldsymbol{x}_1$ to $\boldsymbol{x}_0$. [26] model $\mathcal{L}_0$ as a separate discrete decoder, which maps the continuous signal to discrete pixel values. The second term $\mathcal{L}_\mathcal{T}$ measures the distance from the corrupted $\boldsymbol{x}_T$ to the prior distribution of $\boldsymbol{x}_T$. Since both expressions resemble an isotropic Gaussian and lack trainable parameters, this term remains constant and is not relevant for training. The last term $\mathcal{L}_{t-1}$ denotes the similarity of the true versus the approximated reverse posterior for every state > 0 in the Markov chain. It can be observed that for $T = 1$, the objective in Equation (2.13) is equivalent to that used in VAEs.

By applying the Bayes rule and the closed form of the forward process, the mean of the posterior can be expressed as

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right)\,. \tag{2.14}$$

This notation indicates a valuable discovery. The neural network used for the approximation $p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ does not directly predict the mean of its distribution. Instead, it predicts the noise applied to $\boldsymbol{x}_0$. The network has the same weights across all timesteps and is equipped with $\boldsymbol{x}_t$ and the information about the current t. As $\dim(\boldsymbol{x}_t) = \dim(\boldsymbol{\epsilon}_t)$, this is similar to an image-to-image translation problem. The standard architectural choice is hence a U-Net [3] with time-conditioned encodings.

One remaining task is to parametrize the variances, $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$. [26] set $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t) = \sigma_t^2 \boldsymbol{I}$ to time-dependent constants, where experiments showed that $\sigma_t^2 = \beta_t$ and $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}$ as the upper and lower bounds of the reverse process entropy led to similar results. In contrast, other work (cf. [80]) opted for learning the variance parameters.

Through a series of rearrangements and substitutions (see [79] for full derivation), the final loss objective then boils down to

$$\mathbb{E}_{t\sim[1,T],\boldsymbol{x}_0,\boldsymbol{\epsilon}_t}\left[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)\left\|\boldsymbol{\Sigma}_\theta\right\|^2}\left\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right\|^2\right] \overset{\text{simple}}{\approx} \mathbb{E}_{t\sim[1,T],\boldsymbol{x}_0,\boldsymbol{\epsilon}_t}\left[\left\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right\|^2\right], \quad (2.15)$$

where $\boldsymbol{x}_t = \sqrt{\bar{\alpha}}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}}\boldsymbol{\epsilon}_t$. Empirical evidence has shown that using the simplified version of the loss, which treats the weighting term as constant, yields superior results [26]. In summary, by the merits of the closed-form forward noising and one unified network for all timesteps, a DM can be trained rather efficiently across all Markov states in parallel. To generate a new data sample, the base DDPM must traverse every state from $\boldsymbol{x}_T$ to $\boldsymbol{x}_0$, necessitating thousands of network passes. This results in slow and computationally intensive inference.

Accelerating Diffusion. With the intention of mitigating the inference bottleneck of DMs, *Diffusion Denoising Implicit Models* (DDIM) [81] formulate DM as a non-Markovian process with the same training procedure as DDPMs but a substantially faster generation process. The key idea is to skip certain steps when generating a new data sample from $\boldsymbol{x}_T$. In the original DDPM, one step is being computed by sampling from the predicted posterior distribution. In contrast, the DDIM has no stochastic element:

$$\boldsymbol{x}_{t-1} = \underbrace{\sqrt{\bar{\alpha}_{t-1}}\left(\frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)}{\sqrt{\bar{\alpha}_t}}\right)}_{\text{Prediction of “}\boldsymbol{x}_0\text{”}} + \underbrace{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)}_{\text{Deterministic “noise”}}. \quad (2.16)$$

The inference step consists of two components. First, given $\boldsymbol{x}_t$, the model directly estimates an approximation of $\boldsymbol{x}_0$. Second, to account for the correct amount of noise corruption in the target $\boldsymbol{x}_{t-1}$, a predicted but rescaled $\boldsymbol{\epsilon}$ is added to the initial estimate. This has two major effects. On one hand, the absence of stochasticity due to deterministic model forward passes renders the full DM deterministic, meaning one specific $\boldsymbol{x}_T$ maps to one $\boldsymbol{x}_0$. On the other hand, arbitrary steps can be skipped by adjusting the scaling term $\sqrt{\bar{\alpha}_{t-1}}$ to earlier timesteps than $t-1$ in the second term. The number of inference steps represents a quality trade-off. Typically,

around 100 DDIM steps are used, which maintains generation quality while accelerating the DM by multiple factors.

[82] argue that DDIM is a variant of pseudo numerical methods and improve the sampling quality with even less steps. From another point of view, DMs can be solved as ordinary differential equations, giving rise to a new class of samplers [83, 84]. More approaches to boost the inference speed of DMs include the parallel sampling of steps [85] or progressive distillation [86], among others.

Latent Diffusion Models. A notable drawback of DDPM is that the dimensionality of $\mathcal{Z}$ matches that of $\mathcal{X}$, which may be extremely large and has a high sparsity. This limitation is circumvented, for example, by the VAE framework, where the latent space is usually a semantic bottleneck. However, it has been demonstrated that simply sampling from a Gaussian prior does not achieve the desired fidelity and perplexity for complex $\mathcal{X}$. One possible mitigation is to employ an autoregressive transformer to generate new $\boldsymbol{z}$ instead of sampling from a basic prior distribution [87, 38]. The architecture of a latent diffusion model (LDM) [39] transports this idea into the realm of DMs (see Figure 2.20). Intuitively, an LDM is a VAE in which, during inference, latent samples are generated by a DM before being translated into data space. In other words, the forward and reverse diffusion processes are applied to a semantic meaningful $\mathcal{Z}$. One of the major benefits of conducting diffusion in latent space is the reduced computational effort as $\dim(\mathcal{Z})$ is usually magnitudes smaller than $\dim(\mathcal{X})$. Thus, the LDM comprises two neural networks, the VAE and DM, which are trained separately in two stages. The VAE operates independently of the DM and can be reused across various DM engines. To

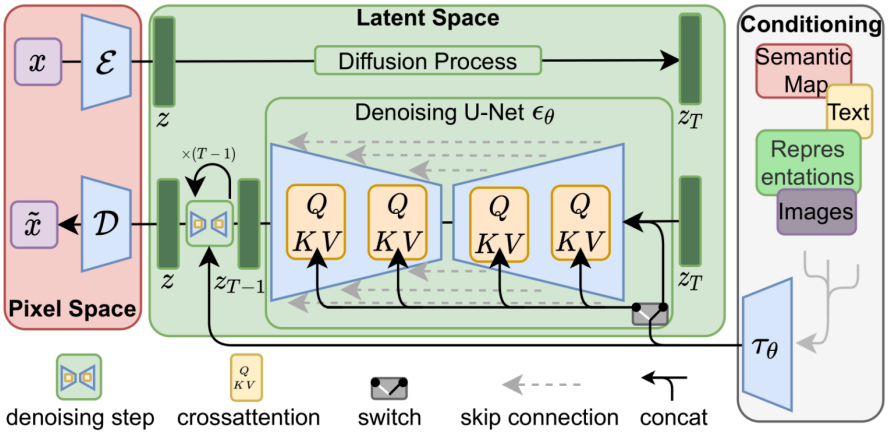


Figure 2.20. Architecture of a latent diffusion model (LDM). Instead of applying diffusion in the data space $\mathcal{X}$, the LDM utilizes a VAE to conduct the process in the latent space $\mathcal{Z}$. Additionally, cross-attention serves as a conditioning mechanism to guide the generation toward a target of choice. Figure from [39].

account for the diffusion in latent space, the DM objective is modified accordingly as follows:

$$\mathbb{E}_{t\sim[1,T],\boldsymbol{z}_0\sim\mathcal{E}(\boldsymbol{x}),\boldsymbol{\epsilon}_t}\left[\left\|\boldsymbol{\epsilon}_t-\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t,t)\right\|^2\right], \tag{2.17}$$

using an encoder $\mathcal{E}$ and $\boldsymbol{z}_t=\sqrt{\bar{\alpha}}\boldsymbol{z}_0+\sqrt{1-\bar{\alpha}}\boldsymbol{\epsilon}_t$.

Moreover, the LDM framework proposed a new method to steer the diffusion process toward some conditioning $\boldsymbol{c}$. This conditioning can be any concept that should be included in the generated data, be it text, images, representations, or scenes. Previous methods were limited to utilizing the gradients of an auxiliary classifier [75] or applying *classifier-free guidance* [88]. Alternatively, using the same mechanism as in encoder-decoder transformer architectures [89], the information can be fed to the underlying diffusion network via cross-attention. For this, an embedding head τ_θ is used to process $\boldsymbol{c}$ into a suitable representation before being passed to the model.

Further Developments. Propelled by the generation quality and capability of conditioning on text inputs, DMs dominate the generative computer vision community. With exponentially growing datasets, large-scale models have been developed by major private industry players capable of affording the necessary GPU resources and advancing the state-of-the-art. Examples include *GLIDE* [90], *DALL·E2* [91], and *DALL·E3* [92] from OpenAI or *Imagen* [93], *SR3* [94], and *Palette* [95] by Google. A significant contribution to the open-source community was made by the *Stable Diffusion* model series, supported by Stability AI. So far, its releases include *Stable Diffusion* 1, 2 [39], XL [96], and 3 [97, 98].

Given the widespread open access to foundational diffusion models, a new field within generative modeling focuses on adapting existing models to meet personal needs with limited resources. For example, *textual inversion* [99] and *DreamBooth* [100] allow to adapt a DM to custom objects or concepts given only a handful of images. *Custom Diffusion* [101] fine-tunes only the cross-attention layer in the DM. *DiffEdit* [102] facilitates semantic image editing through content preservation via automatically generated masks. The *LoRA* [103] method, also often used in large language models, learns only additive low-rank approximations of the model weights. Similarly, *ControlNet* [104] is an auxiliary network integrated into the original model, enabling new modes of conditioning, such as canny edges or poses. The training-free framework *FABRIC* [105] guides the diffusion process based on user preferences over multiple iterations.

Further effort has gone into improving the diffusion process in general. A non-exhaustive list includes, e.g., experimenting with transformers as a backbone [106], using an ensemble of expert denoisers for given timestep range [107], or applying rectified flows for optimal transport between diffusion states [108]. To accelerate the inference speed, *FlashAttention* [109, 110] revises the implementation of the attention mechanism with the DM and provides GPU-specific kernels for increased efficiency. The deployment of DMs on edge devices without access to GPU architectures can be facilitated with weight quantization [111, 112]. While the center of attention is still on image synthesis, the application of diffusion spreads across a myriad of

other modalities and disciplines, for example, videos [113, 114], audio [115], time series [116], anomaly detection [117], learning representations [118], generating gene sequences [119], or MRI-to-CT synthesis [120].

Contributions

The previously described methods are typically applied and validated by training models on extremely large datasets, often comprising tens of millions of natural images or more. Contribution *C5* examines whether these approaches can be adapted to the medical domain, using CXR images as an example. An additional motivation for *C5* was the deficiency of the analyzed GAN generator discovered in contribution *C4*, which potentially can be eliminated by a more complex model. The first part of the proposed contribution involves compiling a collection of open-sourced CXR datasets, referred to as *MaCheX* (massive chest X-ray dataset), encompassing roughly $650,000$ CXR images preprocessed for training. MaCheX was expanded after the paper's publication and now contains nearly 1 million entries. The goal of MaCheX was to provide an optimal basis for training a large-scale foundational CXR DM, which is the second part of *C5*. The model consists of three stages: the first two stages employ a LDM, and the final stage is a super-resolution diffusion model that upscales the target to one megapixel. Empirical results showed that this model provides high-detail and -resolution samples, forming a new state-of-the-art in the domain of CXR synthesis. Moreover, *C5* was the first peer-reviewed publication that investigated radiology-report-to-CXR synthesis, hence, pioneering a new research area. The model is publicly available and equipped with an open-source license.

2.4 Representation Learning

While *Representation Learning* (RL) is a subfield on its own, its concepts pervade most of the DL domain. According to [35] RL is a "*set of techniques that allow automatic construction of data representations needed for machine learning*". This definition aligns with one of the ultimate goals of DL: eliminating the need for manual feature selection and preparation and thus allowing the network to process raw or minimally preprocessed data directly. RL is implicitly or explicitly part of nearly every neural network. For example, a simple image classifier network can be decomposed into two components: the main body and the head. The main body employs a series of linear and non-linear transformations to map the input data into a feature space. This process extracts valuable semantic concepts from the data while removing redundant information. The head, typically a simple linear classifier, then makes decisions based on the features produced by the body. In this end-to-end setting, the main body did indeed learn a representation of the input data.

The subsequent section serves as a glimpse into the RL domain by focusing on three of its subfields. First, Section 2.4.1 demonstrates how RL is embedded as a fundamental concept in generative models. Second, Section 2.4.2 investigates the application of RL for creating semantic embeddings. Third, Section 2.4.3 shows how dimensionality reduction and tensor decomposition are related to RL.

2.4.1 Latent Spaces in Generative Models

Formally, a latent space $\mathcal{Z}$ is an abstract representation or encoding of typically high-dimensional and complex data [15]. This space captures the underlying factors or features that characterize the data but are not directly observable. $\mathcal{Z}$ is also often called *feature* or *embedding space*. The concept of latent spaces is closely related to manifold learning that assumes the observed data reside on a low-dimensional manifold embedded within a high-dimensional space [121]. A latent space possesses several desirable properties [35]. One such property is *smoothness*, which indicates that $\mathcal{Z}$ is continuous, ensuring that small variations within it lead to proportional changes in the output. Another property is *disentanglement*, where distinct concepts from $\mathcal{X}$ are represented in $\mathcal{Z}$ as separate factors or dimensions. For example, in an ideal scenario, a latent space for images of black circles might consist of only two factors: the location of the center and the radius.

Generative Seeds and Latent Spaces. In generative modeling, the latent space plays a central role and is often the "*fuel*" for inducing a generative process. Most, if not all, generative models employ some tractable prior distribution $p(\boldsymbol{z})$ to sample a $\boldsymbol{z} \in \mathcal{Z}$ that connects to a new synthetic sample $\boldsymbol{x} \in \mathcal{X}$. Figure 2.21 illustrates how the different concepts of VAEs, GANs, and DMs are related. The VAE has the most accessible latent space of the three showcased models. Real data samples can be directly converted to a latent representation using the encoder, which facilitates the conversion of large datasets with subsequent analysis of their latent structure.

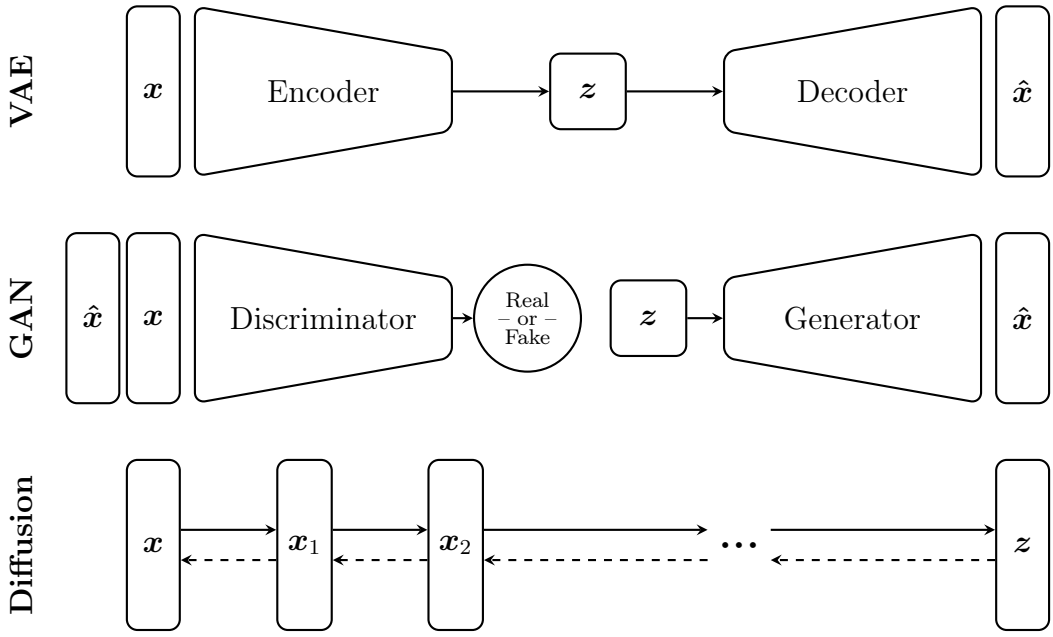


Figure 2.21. A holistic high-level overview of the generative models: GAN, VAE, and DM. Albeit having separate frameworks and concepts, every framework has some connection to a latent space that enables the generative process. Figure inspired by [76].

Conversely, the synthesis process is more complex. While the decoder can readily project a latent sample back to the data space, identifying a valid latent point is non-trivial. Sampling from $\boldsymbol{z} \sim p(\boldsymbol{z})$ does not ensure that $\boldsymbol{z}$ lies on the manifold of $q(\boldsymbol{z} \mid \boldsymbol{x})$. Although the ELBO loss objective minimizes the distance between $p(\boldsymbol{z})$ and $q(\boldsymbol{z} \mid \boldsymbol{x})$, it does not completely align them, as doing so would result in posterior collapse. The strength of this regularization determines the structure of the spanned latent space. A stark alignment to the isotropic Gaussian prior implies a strong factor of disentanglement, yielding favorable representations. The close distance to the prior also aids in finding a $\boldsymbol{z}$ that translates to a meaningful data sample. In practice, typical training data do not follow a Gaussian distribution (e.g. binary images) and this regularization leads to a substantial loss in generation quality. Thus, finding the optimal balance between disentanglement and synthetic quality is a key challenge in fitting a VAE architecture. Aside from pure sample synthesis, the VAE is particularly suited for image manipulation. Through the encoder's translation to feature space, a data point can be modified using various techniques, including interpolations, traversals, and projections, among others. However, given the limited reconstruction quality of the decoder, these are more beneficial for feature discovery or counterfactual explanations rather than for photo-realistic image editing.

The GAN shares some concepts with the VAE architecture. Both have a generating architecture, the generator and the decoder, which take a latent sample and convert it to a data point. Due to the absence of a latent posterior, sampling from $p(\boldsymbol{z})$ in a GAN does not result in a distribution discrepancy as it does in a VAE. Further, the two models utilize a compressing

architecture. While the VAE's encoder projects data into latent space, the GAN's discriminator outputs a scalar indicating the authenticity of the input. Unfortunately, the GAN lacks a direct access point to its latent space. Building the required bridge follows the technique of GAN inversion, described in detail in Section 2.3.3. Fascinatingly, the GAN's prior distribution does form a fully-fledged latent space. It implicitly encodes semantically meaningful features, despite having no explicit encoding process in its training and merely serving as a seed for the generator. With inversion, the GAN can be utilized for latent space editing and manipulation similarly to the VAE, but with a significant advantage: The generated samples are of superior quality compared to those produced by the VAE.

The latent space of the DM exhibits distinct differences compared to that of the VAE and GAN. Firstly, the DM does not involve compression, as the dimensionality of $\boldsymbol{z}$ matches that of $\boldsymbol{x}$. Consequently, there is no feature extraction. Due to the probabilistic nature of the DM, a single $\boldsymbol{z}$ can yield infinitely different $\boldsymbol{x}$ instances. Intuitively, a latent point provides a starting point for a trajectory whose course has not yet been determined. Secondly, as the encoding and decoding process between $\mathcal{X}$ and $\mathcal{Z}$ is based on a stochastic process, new complexity in navigating the latent space is introduced. While VAEs and GANs allow relatively straightforward manipulations of the latent vectors, the DM requires careful control over the diffusion steps to ensure meaningful modifications. This process can be more computationally intensive and less intuitive in comparison to the direct latent space manipulations available in VAEs and GANs. The application of DDIM [81] with a deterministic sampling process mitigates this limitation to a certain degree, allowing for example the interpolation between two $\boldsymbol{z}$. Another viable strategy is masking, where certain parts or areas, such as in an image, are masked during each denoising step. The DM is then tasked with filling in only the unmasked areas, often guided by a conditioning mechanism that steers the diffusion process toward the desired changes.

Unifying Architectures. Each architecture has its advantages and disadvantages. To mitigate the weaknesses of one model and enhance it with the strengths of another, unified architectures attempt to integrate different approaches under a single framework. For example, *Adversarial Autoencoders* [122] substitute the KL-divergence penalty in the VAE objective with an adversarial component that determines whether $\boldsymbol{z}$ originates from a specific prior distribution, enabling the network to learn arbitrary $p(\boldsymbol{z})$. Similarly, *Adversarial Variational Bayes* [123] refines the variational inference process in VAEs by employing adversarial training to distinguish between true and variational posteriors. *Bidirectional Generative Adversarial Networks* (BiGANs) [124, 125] simultaneously learn a GAN and an encoder that provides an inverse mapping, thereby eliminating the need for post-hoc inversion. [126] connect GANs and VAEs formally by interpreting GAN synthesis as performing posterior inference. Further, [127] enable unpaired image-to-image translation with DMs by using a reconstructable encoder model in the stochastic diffusion process. *Diffusion Autoencoders* [128] utilize a semantic encoder paired with a diffusion decoder. Although not closely related to latent spaces, the integration of GANs into DMs remains important. Typically, proposed methods addition-

ally evaluate each timestep in the diffusion process with a discriminator [129, 14]. On top of that, adversarial methods have proven beneficial for distilling large DMs [130].

Contributions

The latent space plays a central role in the contributions employing VAEs, GANs, and DMs. Contribution *C2* analyzes the impact of an additional supervised survival network attached to the latent bottleneck in a VAE. In addition to pure visual reconstruction, the latent space now incorporates survival information, leading to a reordered and restructured representation. In contribution *C3* this added survival head serves to manipulate data. The approximation of the posterior $q(\boldsymbol{z} \mid \boldsymbol{x})$ is guided by the gradients of the survival network to move into regions of increased or decreased hazard. Paired with the decoding network, this gradient-based walk aids in visualizing sources associated with hazards in unstructured data spaces. Further, contribution *C4* applies GAN inversion to obtain access to the latent space of a CXR generator. The paper analyzes the obtained CXR embeddings to find clusters reflecting data characteristics and applies latent space traversals in an attempt to model pathology progression. Lastly, contribution *C5* utilizes the synthesis stack to perform image in- and outpainting on CXR via a masking procedure, demonstrating the capacity of the proposed model.

2.4.2 Neural Embeddings

The previous section examined latent spaces, which are essentially neural embeddings. While latent spaces typically emerge as a useful byproduct in generative models, other methods aim specifically to produce embeddings. A neural embedding, as the name suggests, is typically produced by a highly non-linear neural network. Moreover, an embedding is generally a condensed vector representation of complex data, where semantically meaningful information is contained in disentangled form. A desired property within these representations is *invariance* [35]. In an invariant embedding, changes within the same abstract concept in $\mathcal{X}$ result in a similar location in $\mathcal{Z}$. For example, a batch of dog images should map to rather similar embeddings, whereas images of cats and dogs should not. The term *Self-Supervised Learning* (SSL) often occurs in the context of learning such invariant and disentangled embeddings. This approach leverages the inherent structure of the data without requiring explicit labels.

Language Embeddings. Embeddings are a necessity in *Natural Language Processing* (NLP), as the foundational data is not numeric but present in the forms of characters, words, or sentences. Also, language can be quite ambiguous, where two completely different words can have the same meaning. This aspect benefits substantially from the invariance property. A popular method to provide word embeddings is *Word2Vec* [131], which encodes words by predicting

word-context pairs via a shallow neural network. *GloVe* [132], on the other hand, generates embeddings by constructing a global word-word co-occurrence matrix and then factorizing it to produce vector representations. While these methods have advanced the NLP community, current attention focuses on *vector databases* [133]. Instead of single words, vector databases encode complete documents, storing them in a queryable structure. This approach saves storage and enables semantic document search as well as information retrieval, often assisted by LLMs.

Contrastive Learning. Methods in SSL can be further divided into generative and contrastive approaches [134]. For example, the previously mentioned method Word2Vec with its encoding and decoding structure belongs to the generative class. Conversely, contrastive methods explicitly promote contrasting positive and negative pairs of instances. Aligned with the general idea of invariance, similar (positive) samples should be close together, whereas different (negative) samples are pushed apart. To produce corresponding pairs, contrastive learning heavily relies on data augmentation. The simple framework *SimCLR* [135] operates by maximizing the agreement between two randomly augmented versions of the same sample. Similarly, *MoCo* [136, 137] extends this concept by introducing a dynamic dictionary, serving as a queue of negative data samples, and a momentum-updated encoder. In contrast, *BYOL* [138] does not utilize negative data samples but instead relies on predicting the representation of the augmented sample itself as the objective. To enforce consistency and disentanglement, *Barlow Twins* [139] aim to minimize the distance between the cross-correlation matrix of two augmentations and the identity matrix. Breaking with the concept of SSL, *SupContrast* [140] uses supervision to construct positive and negative data pairs based on labels. Likewise, *JEPA* [141] operates within the realm of supervision by jointly learning embeddings of data and their corresponding labels, aligning the embeddings of both. Joint embeddings are particularly useful in a multi-modality setting, as demonstrated by *CLIP* [142], which constructs an embedding space for both images and text.

Contributions

Embeddings and SSL are powerful toolboxes to extract information from data and create tailored representations. However, in an unregularized setting, the applied algorithm learns to extract any kind of information without considering its ethical applicability. This results in unintended side effects, like networks being able to predict a patient's race in CXR with high confidence [143, 144]. The prediction of sensitive attributes in CXR is feasible not only from raw images but also from neural embeddings obtained through classifiers and SSL methods, which introduces an inherent bias in downstream tasks [145, 146]. Contribution *C6* tackles this challenge using an orthogonalization procedure. In the first step, the existing bias in CXR embeddings is confirmed. The second step con-

cerns removing such a bias by projecting the embedding into a space orthogonal to the sensible features. The analysis concludes that orthogonalization makes the prediction of the targeted attributes infeasible but does not resolve subgroup disparities, highlighting the need for further investigation into sources of bias. Contribution *C6* elaborates on a highly significant and active domain—bias in clinical decision-making—and raises concerns about the practical application of CXR classifiers. It provides a platform for further research in the field of bias in SSL embeddings and showcases novel applications of post-hoc orthogonalization.

2.4.3 Dimensionality Reduction & Tensor Decomposition

Dimensionality reduction and RL describe nearly the same concept but under different umbrellas. While RL emphasizes the automatic extraction of features, the term dimensionality reduction focuses on the compression aspect. Both approaches aim to achieve a unified goal: finding a low-dimensional representation of often complex data that captures its underlying information. For instance, *Principal Component Analysis* (PCA) is a common method for reducing dimensionality. It works by rotating the original data into a new coordinate system, where the axes, or principal components, match the directions of greatest variance. PCA effectively reduces the data's dimensionality by keeping only the principal components with the highest variance. The method thus has found a compact representation of the data. Hereby, performing PCA is equivalent to obtaining the embedding from a linear autoencoder. Although both methods operate unsupervised, a key difference is in their inner workings: Autoencoders require an auxiliary network to generate embeddings, whereas PCA decomposes the data matrix directly. The subsequent sections serve as an introduction to the field of tensor decomposition.

Singular Value Decomposition. *Singular Value Decomposition* (SVD) [147, 148, 149] is one of the foundational algorithms to perform matrix decomposition. The SVD states that given a matrix $\boldsymbol{A} \in \mathbb{R}^{r \times s}$ it can be decomposed as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\top}. \tag{2.18}$$

$\boldsymbol{U} \in \mathbb{R}^{r \times r}$ and $\boldsymbol{V} \in \mathbb{R}^{s \times s}$ are orthogonal, i.e. its columns or rows have a norm of 1 and their pairwise inner product corresponds to 0. Here, the columns of $\boldsymbol{U}$ are referred to as the left singular vectors, while the columns of $\boldsymbol{V}$ are referred to as the right singular vectors. $\boldsymbol{S} \in \mathbb{R}^{r \times s}$ is a diagonal matrix with non-negative entries representing the singular values of $\boldsymbol{A}$, sorted in descending order. It can be proven that an SVD exists for every matrix [149]. In a geometrical interpretation, $\boldsymbol{A}$ is a linear mapping $\mathbb{R}^r \to \mathbb{R}^t$. The SVD decomposes this operation into a rotation in $\mathbb{R}^r$ with $\boldsymbol{U}$, a scaling operation by $\boldsymbol{S}$ with a change in dimensions and a final rotation in $\mathbb{R}^s$ by $\boldsymbol{V}$. In its current form, the SVD does not gain an advantage in the sense of dimensionality reduction but serves to discover concepts hidden in $\boldsymbol{A}$.

Alternatively, the SVD can be expressed as the sum of $\min\{r, s\}$ rank-1 matrices, where one rank-1 matrix results from the outer product of the i-th left and right singular vectors ($\boldsymbol{u}_i$ and $\boldsymbol{v}_i$) scaled by the corresponding singular value s_i:

$$\boldsymbol{A} = \sum_{i=1}^{\min\{r,s\}} s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \,. \tag{2.19}$$

A low-rank approximation of $\boldsymbol{A}$ can be produced by taking only a fraction of all rank-1 matrices in the previous summation. As the singular values are in descending order and reflect the importance of the transformation along the respective axes, a rank-k approximation is created by taking the first k components of the full decomposition:

$$\hat{\boldsymbol{A}} = \sum_{i=1}^{k} s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \,. \tag{2.20}$$

Figure 2.22 illustrates how this truncated SVD relates to the full decomposition. Eventually, $\hat{\boldsymbol{A}} = \boldsymbol{U}_k \boldsymbol{S}_k \boldsymbol{V}_k^\top$, where $\boldsymbol{U}_k \in \mathbb{R}^{r \times k}$, $\boldsymbol{S}_k \in \mathbb{R}^{k \times k}$ and $\boldsymbol{V}_k \in \mathbb{R}^{s \times k}$.

A major benefit of low-rank approximations is the aspect of compression. Instead of having to store the $r \cdot s$ elements of $\boldsymbol{A}$, the truncated SVD amounts to $m \cdot k + k + k \cdot n = k \cdot (1 + m + n)$ entries, which requires substantially less space for small k. In practical scenarios, SVD can be used as a form of naive image compression. Further use cases involve denoising and efficient training of huge weight matrices in neural networks (see *LoRA* [103]). Truncated SVD is a form of lossy compression, so the choice of k is crucial. A basic indicator is given by the magnitude of the s_i, but selecting an appropriate threshold is challenging and depends on the intended application. Simply, one could evaluate the difference of $\boldsymbol{A}$ and $\hat{\boldsymbol{A}}$ and increase k until a certain quality criterion is reached. A rule of thumb could be formulated as choosing k such that the sum of the top k singular values is minimally a pre-defined domain-specific constant times the sum of the remaining singular values [148].

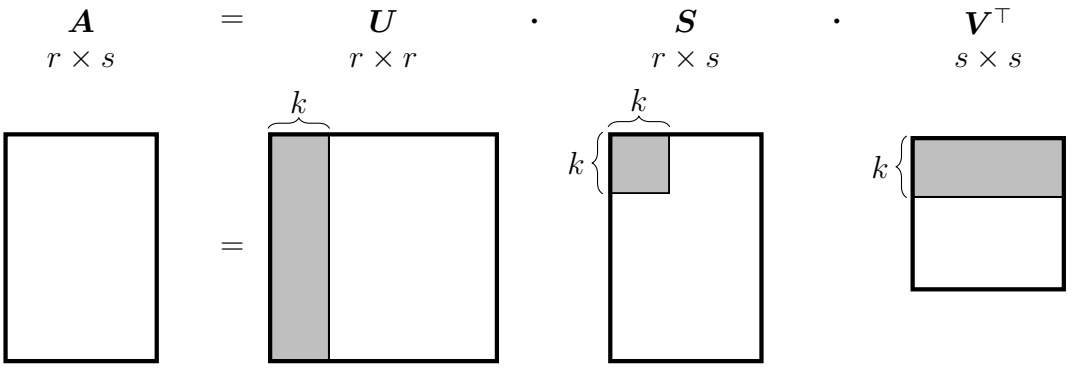


Figure 2.22. Truncated SVD. Instead of conducting a full SVD, the grey boxes demonstrate how a rank-k approximation of $\boldsymbol{A}$ is achieved by considering only the first k left and right singular vectors, along with their corresponding singular values.

Connection to PCA. PCA aims to project a matrix onto its principal components, capturing the primary sources of variance within the data. The first principal component corresponds to the direction that maximizes variance in its projection. Subsequent components are orthogonal to the preceding ones and continue to maximize the explained variance. Geometrically, this is equivalent to rotating the coordinate system, and dimensionality reduction can effectively be performed by disregarding components that do not substantially contribute to the total variance. Finding these directions is the same as computing the eigenvectors. Given that the matrix of interest $\boldsymbol{A}$ is scaled and centered, PCA computes the eigenvectors and eigenvalues of the covariance matrix $\boldsymbol{A}^\top \boldsymbol{A}$ via the eigendecomposition

$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{Q} \boldsymbol{\Lambda} \boldsymbol{Q}^\top , \tag{2.21}$$

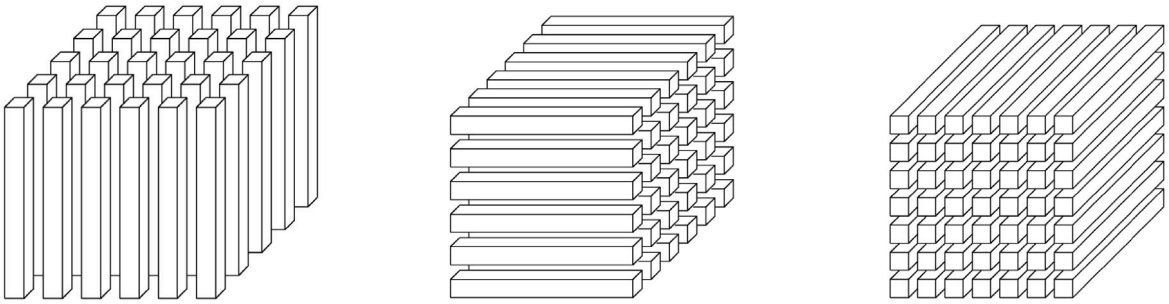
where $\boldsymbol{Q} \in \mathbb{R}^{r \times r}$ contains the eigenvectors as columns and $\boldsymbol{\Lambda} \in \mathbb{R}^{r \times r}$ is a diagonal matrix containing the eigenvalues [149]. By substituting the SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ into the covariance matrix, one can see that

$$\boldsymbol{A}^\top \boldsymbol{A} = (\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top)^\top (\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top) = \boldsymbol{V}\boldsymbol{S}^\top \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{S}^2 \boldsymbol{V}^\top , \tag{2.22}$$

with $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$ due to orthogonality. This implies that the right singular vectors of $\boldsymbol{A}$ correspond to the eigenvectors of the covariance matrix. Thus, eigendecomposition and PCA can be conducted without computing $\boldsymbol{A}^\top \boldsymbol{A}$. Dimensionality reduction can then be performed by selecting a subset of k components via $\boldsymbol{V}_k$ and using the projection $\boldsymbol{A}\boldsymbol{V}_k$. Ultimately, with SVD as a backbone, a data representation was found that focuses on capturing the explained variance.

Tucker decomposition. The methods of SVD and PCA cover only matrices (2D arrays). A possible generalization of SVD to higher-order tensors or multi-dimensional arrays can be formulated using the Tucker decomposition [150, 151, 152]. The following introduction demonstrates the method using a third-order tensor $\boldsymbol{A} \in \mathbb{R}^{r \times s \times t}$ for visualization. Nevertheless, it generalizes to an arbitrary number of dimensions.

A necessary prerequisite is the concept of tensor *fibers*. This refers to a selection process in which all indices except one are fixed [152]. The mode specifies the dimension along which the selection occurs. For example, in a matrix, a mode-1 fiber is a column, whereas a mode-2 fiber is a row. Third-order tensors, such as $\boldsymbol{A}$, also have mode-3 fibers, known as tubes (see Figure 2.23). A tensor can be rearranged into a matrix by matricization, also called unfolding or flattening. With mode-n matricization, the fiber vectors of the n-th mode are stacked as columns of a new matrix. In this demonstration $n \in \{1, 2, 3\}$. Specifically, the mode-2 matricization of $\boldsymbol{A}$ would result in the matrix $\boldsymbol{A}_{(2)} \in \mathbb{R}^{s \times rt}$. Building on the unfolded tensors, a mode-n product denoted as $\times_n$ can be defined, which multiplies every fiber vector of the n-th mode with an arbitrary matrix. Returning to the example of $\boldsymbol{A}$, the mode-2 product with a matrix $\boldsymbol{B} \in \mathbb{R}^{b \times s}$ would be defined as $\boldsymbol{A} \times_2 \boldsymbol{B} \in \mathbb{R}^{r \times b \times t}$ where the resulting mode-2 fibers are computed as $\boldsymbol{B}\boldsymbol{A}_{(2)} \in \mathbb{R}^{b \times rt}$.



(a) Mode-1 column fibers. **(b)** Mode-2 row fibers. **(c)** Mode-3 tube fibers.

Figure 2.23. Fibers of a third-order tensor. Image from [152].

The Tucker decomposition utilizes the mode-n product and decomposes an n-th order tensor into a core tensor $\boldsymbol{C} \in \mathbb{R}^{k_1 \times \cdots \times k_n}$ and n factor matrices. For the third-order tensor $\boldsymbol{A}$ this corresponds to the decomposition

$$\boldsymbol{A} = \boldsymbol{C} \times_1 \boldsymbol{F}_1 \times_2 \boldsymbol{F}_2 \times_3 \boldsymbol{F}_3 \,, \tag{2.23}$$

with factor matrices $\boldsymbol{F}_1 \in \mathbb{R}^{r \times k_1}$, $\boldsymbol{F}_2 \in \mathbb{R}^{s \times k_2}$ and $\boldsymbol{F}_3 \in \mathbb{R}^{t \times k_3}$. The decomposition of $\boldsymbol{A}$ is visualized in Figure 2.24. Hereby, the Tucker decomposition can be thought of as some higher-order PCA as the factor matrices contain the principal components in each mode. The core tensor $\boldsymbol{C}$ captures the interactions between those components. The choice of the core tensor size is thus crucial, as this determines the number of retained components. A small core tensor will result in a high degree of compression and a compact representation coupled with a presumably lossy reconstruction. The optimal choice of these parameters is not trivial and denotes the problem of automatic rank selection [153, 154, 155]. Notably, the tucker decomposition does not have a unique solution.

Higher-Order Singular Value Decomposition. A special case of the Tucker decomposition is the *Higher-Order Singular Value Decomposition* (HOSVD) [156], a generalized version

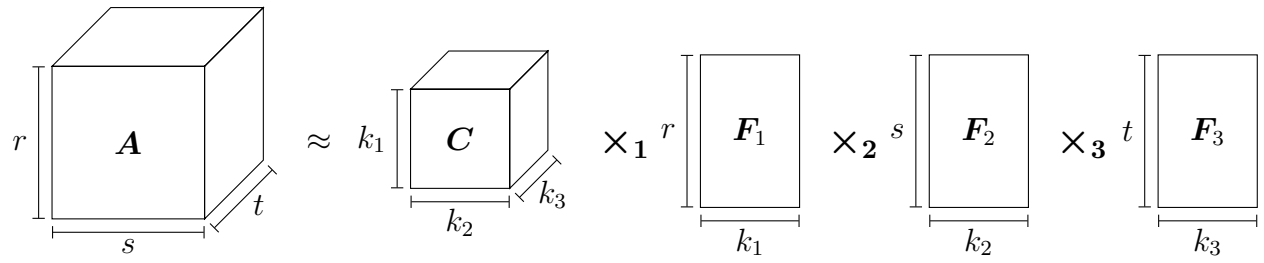


Figure 2.24. Tucker decomposition of a third-order tensor $\boldsymbol{A}$ into a core tensor $\boldsymbol{C}$ and three factor matrices F_1, F_2, F_3.

of the SVD for higher-order tensors. In fact, the decomposition follows Equation (2.23) as well. However, there are some constraints on the core tensor and the factor matrices. While there are no restrictions in the full Tucker decomposition, the HOSVD has an all-orthogonal core, i.e., each mode's fibers are orthogonal to each other, and factor matrices are orthogonal as well. The HOSVD is computed by performing an SVD on each of the mode-n fibers separately, expressed as $\boldsymbol{A}_{(n)} = \boldsymbol{U}_n \boldsymbol{S}_n \boldsymbol{V}_n^\top$. The factor matrix F_n is formed by the left singular vectors, $\boldsymbol{U}_n$. Eventually, the core tensor is derived by the mode-n product of $\boldsymbol{A}$ with the Moore-Penrose inverse of each $\boldsymbol{U}_n$:

$$\boldsymbol{C} = \boldsymbol{A} \times_1 \boldsymbol{U}_1^+ \times_2 \boldsymbol{U}_2^+ \times_3 \boldsymbol{U}_3^+ . \tag{2.24}$$

Note that $\boldsymbol{U}_n^+ = \boldsymbol{U}_n^\top$ due to orthogonality. The HOSVD plays a central role as an initialization for further refining the approximation via, e.g., *Alternating Least Squares* (ALS) algorithms [157, 158] or the *Higher-Order Orthogonal Iteration* (HOOI) [159].

An additional approach for generalizing a SVD to higher-order tensors is per canonical polyadic decomposition, also known as CANDECOMP [160], PARAFAC [161] or CP [162] decomposition. This method decomposes a tensor into a sum of rank-1 tensors, which are expressed as the outer product of vectors. Ultimately, the presented algorithms share a unified goal: to extract patterns and condense information into a compact form in an automated manner, aligning closely with the principles of representation learning.

Contributions

The utilization and development of foundational segmentation models enjoy increasing popularity within the medical image community. For example, the *TotalSegmentator* (TS) [163] is able to segment over 100 anatomical structure in CTs. However, these models are typically quite large and demand substantial computational resources, resulting in inference times that can extend to several minutes. High-performance GPUs that can alleviate the burden are often scarce in everyday clinical practice. A substantial contributor to the computational demands is the use of cost-intensive 3D convolutions, which power models such as the TS. Contribution *C7* tackles this problem by simplifying the network behind the TS. To reduce the number of *Floating Point Operations* (FLOPs) the Tucker decomposition is adapted to decompose the weights of 3D convolutional kernels. Instead of one complex 3D convolution, the factorization yields a series of three simple 3D convolutions. This post-hoc procedure results in a lightweight network that performs nearly on par with the full original model but contains only a small fraction of its parameters and FLOPs. Further, contribution *C7* represents pioneering work in investigating the impact of tensor decomposition on the performance of 3D segmentation networks and provides the community with open-sourced lightweight alternatives to TS.

CHAPTER 3

CONTRIBUTIONS

C1 Constrained Probabilistic Mask Learning for Task-specific Undersampled MRI Reconstruction

Contributing Article:

Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. "Constrained Probabilistic Mask Learning for Task-specific Undersampled MRI Reconstruction". In: *Proceedings of the IEEE/ CVF Winter Conference on Applications of Computer Vision, WACV*. 2024, pp. 7665–7674

Author Contributions:

TW is the corresponding author of this paper and was responsible for its conception, design, methodology, implementation, experiments, and data analysis, all under the supervision of DR, MI, and BB. TW conceived the manuscript, with DR, MI, and BB providing critical feedback and revisions. This paper was presented by TW as a poster and video at WACV in January 2024.

Code Repository: `https://github.com/saiboxx/bernoulli-mri`
arXiv Preprint: `https://arxiv.org/abs/2305.16376`
DOI: 10.1109/WACV57701.2024.00749
Video: `https://www.youtube.com/watch?v=jlSdX4hkIxI`

Copyright Information:

© Reprinted, with permission, from Tobias Weber, Michael Ingrisch, Bernd Bischl and David Rügamer.
In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Ludwig-Maximilians University Munich products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to `http://www.ieee.org/publications_standards/publications/rights/rights_link.html` to learn how to obtain a License from RightsLink. If applicable, University Microfilms and/or Pro-Quest Library, or the Archives of Canada may supply single copies of the dissertation.

C2 Survival-Oriented Embeddings for Improving Accessibility to Complex Data Structures

Contributing article:

Tobias Weber, Michael Ingrisch, Matthias Fabritius, Bernd Bischl, and David Rügamer. "Survival-Oriented Embeddings for Improving Accessibility to Complex Data Structures". In: *Bridging the Gap: From Machine Learning Research to Clinical Practice, NeurIPS Workshops*. 2021

Author Contributions:

TW is the corresponding author of this paper and was responsible for its conception, design, methodology, implementation, experiments, and data analysis, all under the supervision of DR, MI, and BB. TW conceived the manuscript, with DR, MI, MF, and BB providing critical feedback and revisions. This paper was presented by TW as a poster at NeurIPS Workshops in December 2021.
Note: This paper is based on the master's thesis written by TW with the title "Survival-oriented Embeddings with Application to CT Scans of Colorectal Carcinoma Patients with Liver Metastases" at the LMU Munich submitted on May 15, 2021. The master's thesis was supervised by DR, MI, and BB. In this thesis, the primary focus was on enhancing survival prediction performance by proposing a VAE with an additional Cox loss. In contrast, the paper version focuses on the future potential and implications of this Cox-infused VAE with a focus on increased accessibility of deep learning models in clinical settings. The paper was written from scratch to align with the concise workshop format, incorporating an additional literature review. Additionally, new figures were added. Existing figures were re-created as well as refined.

arXiv Preprint: `https://arxiv.org/abs/2110.11303`
Workshop Webpage: `https://neurips.cc/virtual/2021/workshop/21832`

Copyright Information:

This article is licensed under a Creative Commons Attribution 4.0 International license (`https://creativecommons.org/licenses/by/4.0/`).

C3 Towards Modelling Hazard Factors in Unstructured Data Spaces Using Gradient-Based Latent Interpolation

Contributing article:

Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. "Towards Modelling Hazard Factors in Unstructured Data Spaces Using Gradient-Based Latent Interpolation". In: *Deep Generative Models and Downstream Applications, NeurIPS Workshops*. 2021

Author Contributions:

TW is the corresponding author of this paper and was responsible for its conception, design, methodology, implementation, experiments, and data analysis, all under the supervision of DR, MI, and BB. TW conceived the manuscript, with DR, MI, MF, and BB providing critical feedback and revisions. This paper was presented by TW as a poster at NeurIPS Workshops in December 2021.
Note: This paper is based on the master's thesis written by TW with the title "Survival-oriented Embeddings with Application to CT Scans of Colorectal Carcinoma Patients with Liver Metastases" at the LMU Munich submitted on May 15, 2021. The master's thesis was supervised by DR, MI, and BB. In this thesis, the primary focus was on enhancing survival prediction performance by proposing a VAE with an additional Cox loss. As a byproduct, TW discovered that latent samples can be manipulated via gradients. The paper version builds on this discovery and proposes a new framework for gradient-based latent walks in VAEs. The paper was written from scratch to align with the concise workshop format, incorporating an additional literature review and additional justifications on the proposed methodology. New experiments were conducted using the MNIST dataset. Moreover, new figures were added. Existing figures were re-created as well as refined.

arXiv Preprint: `https://arxiv.org/abs/2110.11312`
Workshop Webpage: `https://neurips.cc/virtual/2021/workshop/21878`

Copyright Information:

This article is licensed under a Creative Commons Attribution 4.0 International license (`https://creativecommons.org/licenses/by/4.0/`).

C4 Implicit Embeddings via GAN Inversion for High Resolution Chest Radiographs

Contributing article:

Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. "Implicit Embeddings via GAN Inversion for High Resolution Chest Radiographs". In: *Medical Applications with Disentanglements, MICCAI Workshops*. 2022, pp. 22–32

Author Contributions:

TW is the corresponding author of this paper and was responsible for its conception, design, methodology, implementation, experiments, and data analysis, all under the supervision of DR, MI, and BB. TW conceived the manuscript, with DR, MI, and BB providing critical feedback and revisions. This paper was presented by TW as a poster at MICCAI in September 2022.

Code Repository: `https://github.com/saiboxx/chexray-inversion`
DOI: 10.1007/978-3-031-25046-0_3

Copyright Information:

© 2023 The Author(s), under exclusive license to Springer Nature Switzerland AG

C5 Cascaded Latent Diffusion Models for High-Resolution Chest X-ray Synthesis

Contributing article:

Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. "Cascaded Latent Diffusion Models for High-Resolution Chest X-ray Synthesis". In: *Advances in Knowledge Discovery and Data Mining: 27th Pacific-Asia Conference, PAKDD*. 2023

Author Contributions:

TW is the corresponding author of this paper and was responsible for its conception, design, methodology, implementation, experiments, and data analysis, all under the supervision of DR, MI, and BB. TW conceived the manuscript, with DR, MI, and BB providing critical feedback and revisions. This paper was presented by TW as an oral at PAKDD in May 2023.

Code Repository: `https://github.com/saiboxx/chexray-diffusion`
Dataset Repository: `https://github.com/saiboxx/machex`
arXiv Preprint: `https://arxiv.org/abs/2303.11224`
DOI: 10.1007/978-3-031-33380-4_14

Copyright Information:

© 2023 The Author(s), under exclusive license to Springer Nature Switzerland AG

C6 Post-hoc Orthogonalization for Mitigation of Protected Feature Bias in CXR Embeddings

Contributing article:

Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. "Post-hoc Orthogonalization for Mitigation of Protected Feature Bias in CXR Embeddings". In: *arXiv preprint arXiv:2311.01349*. 2023

Author Contributions:

The idea of removing biases from CXR embeddings was conceived by TW. TW and DR conceptualized the application of orthogonalization to address this problem. TW conducted all experiments, analyzed the data, and wrote the manuscript, with critical feedback and revisions provided by DR, MI, and BB. TW and DR collaboratively wrote the methods section of the manuscript.

Code Repository: `https://github.com/saiboxx/chexray-ortho`
arXiv Preprint: `https://arxiv.org/abs/2311.01349`

Copyright information:

This article is licensed under a Creative Commons Attribution 4.0 International license (`https://creativecommons.org/licenses/by/4.0/`).

C7 Post-Training Network Compression for 3D Medical Image Segmentation: Reducing Computational Efforts via Tucker Decomposition

Contributing article:

Tobias Weber, Jakob Dexl, David Rügamer, and Michael Ingrisch. "Post-Training Network Compression for 3D Medical Image Segmentation: Reducing Computational Efforts via Tucker Decomposition". In: *Radiology: Artificial Intelligence*. Vol. 7. 2. Radiological Society of North America, 2025

Author Contributions:

TW and JD are the joint first authors of this paper. JD identified the need to accelerate the TotalSegmentator model inference. TW conceived the idea of applying and extending Tucker decomposition to the 3D convolution kernels in the model. JD was responsible for data cleaning, preparation, and handling for model training as well as evaluation. TW implemented the mechanics of dynamic layer replacements in neural networks and conducted the experiments and benchmarks. The mixed model for evaluating TotalSegmentator's performance was conceptualized by DR and TW. TW wrote the draft of the manuscript's introduction, methods, and discussion sections with feedback from JD. The results section was composed equally by TW and JD, with TW focusing on quantitative and JD on qualitative evaluation. The manuscript underwent critical feedback and revisions by MI and DR.

Note: This article was accepted for publication at *Radiology: Artificial Intelligence*, but was not yet released at the time of writing. The following pages contain the corresponding arXiv preprint.

Code Repository: `https://github.com/ClinicalDataScience/tucker-cnn`
arXiv Preprint: `https://arxiv.org/abs/2404.09683`
DOI: 10.1148/ryai.240353

Copyright information:

© 2025 RSNA.
This manuscript has been accepted for publication in *Radiology: Artificial Intelligence* (`https://pubs.rsna.org/journal/ai`), which is published by the Radiological Society of North America (RSNA).

CHAPTER 4

OUTLOOK AND CONCLUSION

After presenting the contributions, this thesis ends with an outlook toward further research and concluding remarks.

4.1 Outlook

The applied generative algorithms in this thesis‘ contributions, mainly VAEs, GANs, and DMs, while all having their pros and cons, can be depicted as some stage in the evolution of generative models. Undoubtedly the latest stage and the one with the most recent attention is the era of DMs. With the development of accelerating the diffusion process substantially, the concept of the generative model trilemma begins to crumble. DMs exhibit superior synthesis quality with high diversity and a steadily increasing inference speed. DMs are already dominating the medical generative model community and it is most likely that this trend will not change in the near future.

Foundation Models. An observable movement in the generative community is the shift from dedicated task-specific specialist models toward large foundational models that can handle a variety of tasks in a zero-shot fashion or with minimal adjustments. This development is most noticeable in the success of LLMs but also vision models, such as Stable Diffusion [39]. Moreover, the boundaries between data modalities begin to blur as multi-modal synthesis models are on the rise, which can generate not only images but also text data. This evolution proposes a new challenge but also an opportunity for the medical image community. A foundation model for the medical domain is particularly challenging as the model needs to account for a myriad of different attributes. This begins with the diversity of medical imaging modalities (e.g., X-ray, CT, MRI), where images from the same modality and anatomical region can exhibit entirely different properties just by changing the scanning device or sequence.

A medical foundational generative model should thus be able to provide a fine-grained level of control.

Another challenge is posed by the volumetric nature of medical scans. Synthesizing 3D is not only subject to the curse of dimensionality but also requires exponentially more computational effort. Finding methods that allow the efficient synthesis of these complex data spaces could be of great value. An initial approach to mitigate this issue involves the use of LDMs, which has also been applied in the context of 3D medical data. Lastly, integrating text in the synthesis process can be done in the form of, e.g., radiological reports and enables automatic report generation or new ways of detailed control over the synthesis process. This is in principle possible with already existing methods but it is still far from being a holistic and matured field of research.

Medical Data. When entering the realm of foundational models, a further requirement is an enormous amount of training data, which opens another set of challenges. Notably, this aspect is not as present in the CXR community, as multiple open-source datasets are accessible nowadays. In contrast, large volumetric datasets consume terabytes or even petabytes in storage, straining the physical capacity of hardware drives. Medical image formats, such as DICOM or NIfTI are not exactly suited for efficient neural network training. For example, the decoding process causes a substantial CPU overhead, whereas the sheer amount of data maximizes the I/O load of the system. Further development towards a data standard that is tailored for fast loading and resource-saving storage of 3D data poses a significant factor in advancing medical neural network training and fully utilizing the otherwise undersupplied high-capacity GPUs.

Apart from processing data, the data itself must be available at first. In theory, most hospitals have an immense stock of patient data in their archive. Nevertheless, human data is subject to strict data protection laws in many countries. While this protects privacy to some degree, this can lead to decelerating the progress in DL research, hungry for huge amounts of any kind of data. Optimally, a middle ground would be found, where sharing anonymized data across multiple institutions and nations is encouraged and not riddled by intense bureaucracy. New generation datasets with the scale of those in natural computer vision may only exist in a realm of open-source and collaboration. Further research could thus be concerned with building partnerships and establishing ways to provide a common and easy-to-use interface for sharing and composing large medical datasets. This open-source character not only facilitates the training of more capable and robust models but also enables the community to directly engage with the data, mitigating the occurrence of biases or leakages.

Evaluation. Another point is the translation from conceptual research into clinical practice. Here, generative models or representation learning are merely a toolbox that requires proper use to gain an actual practical benefit. In all cases, rigorous evaluation of the proposed method poses a necessity. For generative models, this naturally requires the attention of experienced medical practitioners that examine the plausibility and validity of the synthesized content. In

the case of representation learning, the data representation should encode relevant information and should less be prone to proxy or protected features. This transition to clinical practice is an important step in future research, facilitating care to prevent harmful actions in real-world scenarios.

Deployment. Lastly, the deployment phase of such models gives opportunities for further improvement. As currently seen in research surrounding LLMs, fast inference speed and model compression promote overall adoption and democratize large neural networks. In particular, creating an efficient clinical workflow for integrating DL into everyday practice is of importance. Its implementation can take many forms, ranging from cloud solutions, where neural network hooks enhance stored patient data, to real-time interventional measures that require reaction times in milliseconds. Efficient neural network inference can be done with various methods, e.g., powered by model compression via tensor decomposition or weight quantization for fast CPU inference. However, the application of advanced inference methods in the medical domain paired with its specific requirements, is still in its infancy, paving the way for more elaborate methods and implementation.

4.2 Conclusion

This thesis focuses on advancing DL for medical imaging with a focus on generative modeling and representation learning. The presented contributions introduce several innovations and novel methods all with open-sourced code repositories. Among other things, this thesis established a new perspective on MRI undersampling by direct optimization of the undersampling pattern enabling data- and task-specific masking. Moreover, the VAE framework was successfully extended to incorporate survival information in the synthesis process, building a basis for interpretable survival analysis of complex data. The research area of CXR synthesis was improved by proposing GAN inversion for obtaining image embeddings and by open-sourcing a foundational CXR synthesis diffusion model including a unified collection of CXR datasets. Also within the field of CXR analysis, the first application of post-hoc orthogonalization continues the ongoing discussion about inherent bias in CXR classifiers. Lastly, this thesis explored the compressibility of foundational 3D segmentation models by developing a pioneering approach to factorizing convolutional kernels with Tucker decomposition.

These contributions provide various experimental perspectives, ideas, and techniques but are limited in their evaluation of actual clinical applicability. Aligning with the previous section on research outlook, further work is needed to strengthen the position of the proposed contributions and ultimately enable their application in real-world scenarios beyond the confines of the research environment.

Status Quo. This is also the point to reflect on the contrastive quotes from Geoffrey Hinton and Curtis Langlotz (see Chapter 1). As Langlotz himself later elaborates in an editorial [164],

the need for radiologists will persist. An imminent transformation in the field is the changing role of radiologists, similar to the evolution of bank tellers' tasks with the advent of ATMs. While a radiologist might not directly be required to segment anatomical structures or create a radiological report in the near future, a medical expert is still a necessity. For instance, although direct communication between radiologists and patients is not common, radiologists play an essential indirect role in patient care. They ensure that the diagnostic information they provide is accurate and interpretable, acting as a vital bridge between complex imaging results and the physicians treating the patients. This connection is crucial for informed patient care and is less likely to be replaced by AI. Additionally, DL algorithms are not entirely infallible, and generative models, in particular, are prone to hallucinations. A radiologist is thus an extremely important instance for controlling the outputs of DL models, as mistakes can result in costing the life of a patient. Overall, it appears that Hinton, the godfather of AI, might have been somewhat overenthusiastic about the hype at that time. As the demand for radiologists still exists in clinical practice, their roles are shifting, whereas AI is a steady companion in alleviating typical tasks and eventually becomes a standard everyday tool.

Concluding Remarks. DL is a highly capable enabler, demonstrating promising results in the medical imaging domain. However, the rapid pace of advancements in this field raises questions about the rigor and quality of the exponentially increasing scientific contributions. While the surge in publications is often seen as evidence of a thriving field, it is essential to ensure that these contributions stem from sincere scientific exploration rather than from a *publish or perish* mentality. The stakes are undeniably high, given that innovation can have a direct impact on human lives. Therefore this thesis ends with an appeal to the scientific community to prioritize research quality and integrity over mere quantity. Senior researchers, in particular, bear the responsibility of using their influence to advance medical research that benefits the public, especially considering the substantial public funding that supports most open research. Failing to adhere to these principles could lead to a scenario where proprietary industry solutions overshadow open collaboration, ultimately slowing overall progress. This is similar to the challenges faced by the LLM community, which involves major industrial players like OpenAI, Google, and Meta that now have a substantial advantage over public research with work originally built from public funding and effort. To ensure continued meaningful scientific progress, the academic community must remain committed to transparency, collaboration, and the ethical use of DL.

While DL is a powerful toolbox with vast potential, its application in the medical domain requires special caution and responsibility. This thesis provided a brief overview of its current and future possibilities with focus on generative modeling and representation learning, emphasizing the importance of collaboration and transparency in achieving long-term success. Fostering these values is a key component for continuous progress in open medical science, and, ultimately, improved patient outcomes.

REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems, NeurIPS* 25 (2012).

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2009, pp. 248–255.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI*. Springer. 2015, pp. 234–241.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2016, pp. 770–778.

[5] Jerrold T Bushberg and John M Boone. *The essential physics of medical imaging*. Lippincott Williams & Wilkins, 2011.

[6] Thorsten M Buzug. "Computed tomography". In: *Springer handbook of medical technology*. Springer, 2011, pp. 311–342.

[7] Andrew Hamilton. *Black Holes*. Lecture in Astrophysics. University of Colorado. 2019.

[8] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. 4th ed. Pearson, 2017.

[9] J Grunert. *Strahlenschutz für Röntgendiagnostik und Computertomografie*. Springer, 2019.

[10] David Bernard. "History of X-rays – 125 years in the making". In: *excillum.com* (2020).

[11] Nadine Barrie Smith and Andrew Webb. *Introduction to medical imaging: physics, engineering and clinical applications*. Cambridge university press, 2010.

[12] Lee W Goldman. "Principles of CT and CT technology". In: *Journal of Nuclear Medicine Technology* 35.3 (2007), pp. 115–128.

[13] Robert A Pooley. "Fundamental physics of MR imaging". In: *Radiographics* 25.4 (2005), pp. 1087–1099.

[14] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. "Tackling the Generative Learning Trilemma with Denoising Diffusion GANs". In: *International Conference on Learning Representations, ICLR*. 2022.

[15] David Foster. *Generative Deep Learning: Teaching Machines to Paint, Write, Compose and Play*. O'Reilly Media, Inc., 2022.

[16] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[17] Jakub Tomczak. *Deep Generative Modeling*. Springer, 2021, pp. 1–12.

[18] Andrew Ng and Michael Jordan. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes". In: *Advances in Neural Information Processing Systems, NeurIPS* 14 (2001).

[19] Kevin P Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

[20] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[21] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. "Gemini: a family of highly capable multimodal models". In: *arXiv preprint arXiv:2312.11805* (2023).

[22] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *International Conference on Learning Representations, ICLR*. 2014.

[23] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic backpropagation and approximate inference in deep generative models". In: *International Conference on Machine Learning, ICML*. 2014, pp. 1278–1286.

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[25] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International Conference on Machine Learning, ICML*. PMLR. 2015, pp. 2256–2265.

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems, NeurIPS* 33 (2020), pp. 6840–6851.

[27] Danilo Rezende and Shakir Mohamed. "Variational inference with normalizing flows". In: *International Conference on Machine Learning, ICML*. PMLR. 2015, pp. 1530–1538.

[28] Chang Gao, Shu-Fu Shih, J Paul Finn, and Xiaodong Zhong. "A Projection-Based K-space Transformer Network for Undersampled Radial MRI Reconstruction with Limited Training Subjects". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI*. 2022, pp. 726–736.

[29] Aleksandr Belov, Joël Stadelmann, Sergey Kastryulin, and Dmitry V Dylov. "Towards ultrafast MRI via extreme k-space undersampling and superresolution". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI*. 2021, pp. 254–264.

[30] Hyungjin Chung and Jong Chul Ye. "Score-based diffusion models for accelerated MRI". In: *Medical Image Analysis* 80 (2022), p. 102479.

[31] Salman UH Dar, Şaban Öztürk, Yilmaz Korkmaz, Gokberk Elmas, Muzaffer Özbey, et al. "Adaptive diffusion priors for accelerated mri reconstruction". In: *arXiv:2207.05876 [cs, eess]* (2022).

[32] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. "fastMRI: An open dataset and benchmarks for accelerated MRI". In: *arXiv:1811.08839 [physics, stat]* (2019).

[33] Cagla Deniz Bahadir, Adrian V Dalca, and Mert R Sabuncu. "Learning-based Optimization of the Under-sampling Pattern in MRI". In: *Information Processing in Medical Imaging, IPMI*. 2019, pp. 780–792.

[34] Artem Razumov, Oleg Y Rogov, and Dmitry V Dylov. "Optimal MRI Undersampling Patterns for Pathology Localization". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI*. 2022, pp. 768–779.

[35] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), pp. 1798–1828.

[36] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. "Designing an encoder for stylegan image manipulation". In: *ACM Transactions on Graphics (TOG)* 40.4 (2021), pp. 1–14.

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2018, pp. 586–595.

[38] Patrick Esser, Robin Rombach, and Bjorn Ommer. "Taming transformers for high-resolution image synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2021, pp. 12873–12883.

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models. 2022 IEEE". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2021, pp. 10674–10685.

[40] Richard Bellman and Robert Kalaba. "A mathematical theory of adaptive control processes". In: *Proceedings of the National Academy of Sciences* 45.8 (1959), pp. 1288–1290.

[41] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877.

[42] Lei Cai, Hongyang Gao, and Shuiwang Ji. "Multi-stage variational auto-encoders for coarse-to-fine image generation". In: *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM. 2019, pp. 630–638.

[43] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. "Understanding disentangling in *beta*-VAE". In: *arXiv preprint arXiv:1804.03599* (2018).

[44] Christopher M Bishop and Hugh Bishop. *Deep learning: Foundations and concepts*. Springer Nature, 2023.

[45] Mehdi Mirza and Simon Osindero. "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784* (2014).

[46] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. "Least Squares Generative Adversarial Networks". In: *Proceedings of the IEEE International Conference on Computer Vision, ICCV*. IEEE Computer Society, 2017, pp. 2813–2821.

[47] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks". In: *International Conference on Machine Learning, ICML*. PMLR. 2017, pp. 214–223.

[48] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. "Improved training of wasserstein gans". In: *Advances in Neural Information Processing Systems, NeurIPS* 30 (2017).

[49] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. "Spectral Normalization for Generative Adversarial Networks". In: *International Conference on Learning Representations, ICLR*. 2018.

[50] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. "Training generative adversarial networks with limited data". In: *Advances in Neural Information Processing Systems, NeurIPS* 33 (2020), pp. 12104–12114.

[51] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *International Conference on Learning Representations, ICLR*. 2016.

[52] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. "Progressive Growing of GANs for Improved Quality, Stability, and Variation". In: *International Conference on Learning Representations, ICLR*. 2018.

[53] Tero Karras, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2019, pp. 4401–4410.

[54] Xun Huang and Serge Belongie. "Arbitrary style transfer in real-time with adaptive instance normalization". In: *Proceedings of the IEEE International Conference on Computer Vision, ICCV*. 2017, pp. 1501–1510.

[55] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and improving the image quality of stylegan". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2020, pp. 8110–8119.

[56] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Alias-free generative adversarial networks". In: *Advances in Neural Information Processing Systems, NeurIPS* 34 (2021), pp. 852–863.

[57] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. "Scaling up gans for text-to-image synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2023, pp. 10124–10134.

[58] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. "Stylegan-T: Unlocking the power of gans for fast large-scale text-to-image synthesis". In: *International Conference on Machine Learning, ICML*. PMLR. 2023, pp. 30105–30118.

[59] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. "Gan inversion: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.3 (2022), pp. 3121–3138.

[60] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. "Ganalyze: Toward visual definitions of cognitive image properties". In: *Proceedings of the IEEE International Conference on Computer Vision, ICCV*. 2019, pp. 5744–5753.

[61] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. "Ganspace: Discovering interpretable gan controls". In: *Advances in Neural Information Processing Systems, NeurIPS* 33 (2020), pp. 9841–9850.

[62] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. "Encoding in style: a stylegan encoder for image-to-image translation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2021, pp. 2287–2296.

[63] Fiete Lüer and Christian Bohm. "Anomaly Detection using Generative Adversarial Networks Reviewing methodological progress and challenges". In: *ACM SIGKDD Explorations Newsletter* 25.2 (2024), pp. 29–41.

[64] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. "Autoencoding beyond pixels using a learned similarity metric". In: *International Conference on Machine Learning, ICML*. PMLR. 2016, pp. 1558–1566.

[65] Fiete Lüer, **Tobias Weber**, Maxim Dolgich, and Christian Böhm. "Adversarial anomaly detection using gaussian priors and nonlinear anomaly scores". In: *IEEE International Conference on Data Mining Workshops, ICDMW*. IEEE. 2023, pp. 550–559.

[66] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery". In: *International Conference on Information Processing in Medical Imaging, IPMI*. Springer. 2017, pp. 146–157.

[67] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks". In: *Medical Image Analysis* 54 (2019), pp. 30–44.

[68] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. "Interpreting the latent space of gans for semantic face editing". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2020, pp. 9243–9252.

[69] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. "Restyle: A residual-based stylegan encoder via iterative refinement". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*. 2021, pp. 6711–6720.

[70] Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. "Hyperinverter: Improving stylegan inversion via hypernetwork". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2022, pp. 11389–11398.

[71] Ken Shoemake. "Animating rotation with quaternion curves". In: *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*. 1985, pp. 245–254.

[72] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. "Deep Learning Face Attributes in the Wild". In: *Proceedings of the IEEE International Conference on Computer Vision, ICCV*. 2015.

[73] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. "Support vector machines". In: *IEEE Intelligent Systems and Their Applications* 13.4 (1998), pp. 18–28.

[74] Bradley Segal, David M Rubin, Grace Rubin, and Adam Pantanowitz. "Evaluating the clinical realism of synthetic chest x-rays generated using progressively growing gans". In: *SN Computer Science* 2.4 (2021), p. 321.

[75] Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat gans on image synthesis". In: *Advances in Neural Information Processing Systems, NeurIPS* 34 (2021), pp. 8780–8794.

[76] Lilian Weng. "What are diffusion models?" In: *lilianweng.github.io* (2021).

[77] Yang Song and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution". In: *Advances in Neural Information Processing Systems, NeurIPS* 32 (2019).

[78] Yang Song and Stefano Ermon. "Improved techniques for training score-based generative models". In: *Advances in Neural Information Processing Systems, NeurIPS* 33 (2020), pp. 12438–12448.

[79] Calvin Luo. "Understanding diffusion models: A unified perspective". In: *arXiv preprint arXiv:2208.11970* (2022).

[80] Alexander Quinn Nichol and Prafulla Dhariwal. "Improved denoising diffusion probabilistic models". In: *International Conference on Machine Learning, ICML*. PMLR. 2021, pp. 8162–8171.

[81] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising Diffusion Implicit Models". In: *International Conference on Learning Representations, ICLR*. 2021.

[82] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. "Pseudo Numerical Methods for Diffusion Models on Manifolds". In: *International Conference on Learning Representations, ICLR*. 2022.

[83] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. "Elucidating the design space of diffusion-based generative models". In: *Advances in Neural Information Processing Systems, NeurIPS* 35 (2022), pp. 26565–26577.

[84] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. "Consistency Models". In: *International Conference on Machine Learning, ICML*. Vol. 202. PMLR, 2023, pp. 32211–32252.

[85] Andy Shih, Suneel Belkhale, Stefano Ermon, Dorsa Sadigh, and Nima Anari. "Parallel sampling of diffusion models". In: *Advances in Neural Information Processing Systems, NeurIPS* 36 (2024).

[86] Tim Salimans and Jonathan Ho. "Progressive distillation for fast sampling of diffusion models". In: *International Conference on Learning Representations, ICLR*. 2022.

[87] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. "Zero-shot text-to-image generation". In: *International Conference on Machine Learning, ICML*. PMLR. 2021, pp. 8821–8831.

[88] Jonathan Ho and Tim Salimans. "Classifier-free diffusion guidance". In: *arXiv preprint arXiv:2207.12598* (2022).

[89] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in Neural Information Processing Systems, NeurIPS* 30 (2017).

[90] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models". In: *International Conference on Machine Learning, ICML*. Vol. 162. PMLR, 2022, pp. 16784–16804.

[91] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* 1.2 (2022), p. 3.

[92] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. "Improving image generation with better captions". In: *OpenAI* (2023).

[93] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. "Photorealistic text-to-image diffusion models with deep language understanding". In: *Advances in Neural Information Processing Systems, NeurIPS* 35 (2022), pp. 36479–36494.

[94] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. "Image Super-Resolution via Iterative Refinement". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (2023), pp. 4713–4726.

[95] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. "Palette: Image-to-image diffusion models". In: *ACM SIGGRAPH Conference Proceedings*. 2022, pp. 1–10.

[96] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. "Sdxl: Improving latent diffusion models for high-resolution image synthesis". In: *arXiv preprint arXiv:2307.01952* (2023).

[97] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. "Scaling rectified flow transformers for high-resolution image synthesis". In: *International Conference on Machine Learning, ICML*. 2024.

[98] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. "Fast high-resolution image synthesis with latent adversarial diffusion distillation". In: *arXiv preprint arXiv:2403.12015* (2024).

[99] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion". In: *International Conference on Learning Representations, ICLR*. 2023.

[100] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2023, pp. 22500–22510.

[101] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. "Multi-concept customization of text-to-image diffusion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2023, pp. 1931–1941.

[102] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. "DiffEdit: Diffusion-based semantic image editing with mask guidance". In: *International Conference on Learning Representations, ICLR*. 2023.

[103] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "LoRA: Low-Rank Adaptation of Large Language Models". In: *International Conference on Learning Representations, ICLR*. 2022.

[104] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*. 2023, pp. 3836–3847.

[105] Dimitri Von Rütte, Elisabetta Fedele, Jonathan Thomm, and Lukas Wolf. "Fabric: Personalizing diffusion models with iterative feedback". In: *arXiv preprint arXiv:2307.10159* (2023).

[106] William Peebles and Saining Xie. "Scalable diffusion models with transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*. 2023, pp. 4195–4205.

[107] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. "ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers". In: *arXiv preprint arXiv:2211.01324* (2022).

[108] Xingchao Liu, Chengyue Gong, and Qiang Liu. "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow". In: *The Eleventh International Conference on Learning Representations, ICLR*. 2023.

[109] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. "Flashattention: Fast and memory-efficient exact attention with io-awareness". In: *Advances in Neural Information Processing Systems, NeurIPS* 35 (2022), pp. 16344–16359.

[110] Tri Dao. "Flashattention-2: Faster attention with better parallelism and work partitioning". In: *arXiv preprint arXiv:2307.08691* (2023).

[111] Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. "Post-training quantization on diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2023, pp. 1972–1981.

[112] Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. "Q-diffusion: Quantizing diffusion models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*. 2023, pp. 17535–17545.

[113] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. "Stable video diffusion: Scaling latent video diffusion models to large datasets". In: *arXiv preprint arXiv:2311.15127* (2023).

[114] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. "Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models". In: *arXiv preprint arXiv:2402.17177* (2024).

[115] Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. "Fast Timing-Conditioned Latent Audio Diffusion". In: *arXiv preprint arXiv:2402.04825* (2024).

[116] Lequan Lin, Zhengkun Li, Ruikun Li, Xuliang Li, and Junbin Gao. "Diffusion models for time-series applications: a survey". In: *Frontiers of Information Technology & Electronic Engineering* (2023), pp. 1–23.

[117] Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. "Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2022, pp. 650–656.

[118] Sarthak Mittal, Korbinian Abstreiter, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. "Diffusion Based Representation Learning". In: *International Conference on Machine Learning, ICML*. Vol. 202. PMLR, 2023, pp. 24963–24982.

[119] Zehui Li, Yuhao Ni, Tim August B Huygelen, Akashaditya Das, Guoxuan Xia, Guy-Bart Stan, and Yiren Zhao. "Latent Diffusion Model for DNA Sequence Generation". In: *arXiv preprint arXiv:2310.06150* (2023).

[120] Shaoyan Pan, Elham Abouei, Jacob Wynne, Chih-Wei Chang, Tonghe Wang, Richard LJ Qiu, Yuheng Li, Junbo Peng, Justin Roper, Pretesh Patel, et al. "Synthetic CT generation from MRI using 3D transformer-based denoising diffusion model". In: *Medical Physics* 51.4 (2024), pp. 2538–2548.

[121] Yunqian Ma and Yun Fu. *Manifold learning theory and applications*. Vol. 434. CRC press Boca Raton, 2012.

[122] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. "Adversarial autoencoders". In: *arXiv preprint arXiv:1511.05644* (2015).

[123] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. "Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks". In: *International Conference on Machine Learning, ICML*. PMLR. 2017, pp. 2391–2400.

[124] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. "Adversarial Feature Learning". In: *International Conference on Learning Representations, ICLR*. 2017.

[125] Jeff Donahue and Karen Simonyan. "Large scale adversarial representation learning". In: *Advances in Neural Information Processing Systems, NeurIPS* 32 (2019).

[126] Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P. Xing. "On Unifying Deep Generative Models". In: *International Conference on Learning Representations, ICLR*. 2018.

[127] Chen Henry Wu and Fernando De la Torre. "A latent space of stochastic diffusion models for zero-shot image editing and guidance". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*. 2023, pp. 7378–7387.

[128] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. "Diffusion autoencoders: Toward a meaningful and decodable representation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2022, pp. 10619–10629.

[129] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. "Diffusion-GAN: Training GANs with Diffusion". In: *International Conference on Learning Representations, ICLR*. 2023.

[130] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. "Adversarial diffusion distillation". In: *arXiv preprint arXiv:2311.17042* (2023).

[131] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space". In: *International Conference on Learning Representations, ICLR*. 2013.

[132] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*. 2014, pp. 1532–1543.

[133] James Jie Pan, Jianguo Wang, and Guoliang Li. "Survey of vector database management systems". In: *arXiv preprint arXiv:2310.14021* (2023).

[134] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. "Self-supervised learning: Generative or contrastive". In: *IEEE Transactions on Knowledge and Data Engineering* 35.1 (2021), pp. 857–876.

[135] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations". In: *International Conference on Machine Learning, ICML*. PMLR. 2020, pp. 1597–1607.

[136] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 2020, pp. 9729–9738.

[137] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. "Improved baselines with momentum contrastive learning". In: *arXiv preprint arXiv:2003.04297* (2020).

[138] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. "Bootstrap your own latent-a new approach to self-supervised learning". In: *Advances in Neural Information Processing Systems, NeurIPS* 33 (2020), pp. 21271–21284.

[139] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. "Barlow twins: Self-supervised learning via redundancy reduction". In: *International Conference on Machine Learning, ICML*. PMLR. 2021, pp. 12310–12320.

[140] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. "Supervised contrastive learning". In: *Advances in Neural Information Processing Systems, NeurIPS* 33 (2020), pp. 18661–18673.

[141] Yann LeCun. "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27". In: *Open Review* (2022).

[142] 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". In: *International Conference on Machine Learning, ICML*. PMLR. 2021, pp. 8748–8763.

[143] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. "AI recognition of patient race in medical imaging: a modelling study". In: *The Lancet Digital Health* 4.6 (2022), e406–e414.

[144] Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, et al. "Reading race: AI recognises patient's racial identity in medical images". In: *arXiv preprint arXiv:2107.10356* (2021).

[145] Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck. "Algorithmic encoding of protected characteristics in chest X-ray disease detection models". In: *EBioMedicine* 89 (2023).

[146] Ben Glocker, Charles Jones, Mélanie Roschewitz, and Stefan Winzeck. "Risk of bias in chest radiography deep learning foundation models". In: *Radiology: Artificial Intelligence* 5.6 (2023), e230060.

[147] Tim Roughgarden and Gregory Valiant. *The Singular Value Decomposition (SVD) and Low-Rank Matrix Approximations*. Lecture in CS168: The Modern Algorithmic Toolbox. Stanford University. 2024.

[148] Jonathon Shlens. "A tutorial on principal component analysis". In: *arXiv:1404.1100* (2014).

[149] Gilbert Strang. *Introduction to linear algebra, Fifth Edition*. Wellesley-Cambridge Press, 2016.

[150] Ledyard R Tucker. "Some mathematical notes on three-mode factor analysis". In: *Psychometrika* 31.3 (1966), pp. 279–311.

[151] Wolfgang Hackbusch. *Tensor spaces and numerical tensor calculus*. Vol. 42. Springer, 2012.

[152] Tamara G Kolda and Brett W Bader. "Tensor decompositions and applications". In: *SIAM review* 51.3 (2009), pp. 455–500.

[153] Farnaz Sedighin, Andrzej Cichocki, and Anh-Huy Phan. "Adaptive rank selection for tensor ring decomposition". In: *IEEE Journal of Selected Topics in Signal Processing* 15.3 (2021), pp. 454–463.

[154] Tatsuya Yokota and Andrzej Cichocki. "Multilinear tensor rank estimation via sparse tucker decomposition". In: *2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS)*. IEEE. 2014, pp. 478–483.

[155] Maxim Kodryan, Dmitry Kropotov, and Dmitry Vetrov. "Mars: Masked automatic ranks selection in tensor decompositions". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 3718–3732.

[156] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. "A multilinear singular value decomposition". In: *SIAM Journal on Matrix Analysis and Applications* 21.4 (2000), pp. 1253–1278.

[157] Pieter M Kroonenberg and Jan De Leeuw. "Principal component analysis of three-mode data by means of alternating least squares algorithms". In: *Psychometrika* 45 (1980), pp. 69–97.

[158] Arie Kapteyn, Heinz Neudecker, and Tom Wansbeek. "An approach to n-mode components analysis". In: *Psychometrika* 51 (1986), pp. 269–275.

[159] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. "On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors". In: *SIAM Journal on Matrix Analysis and Applications* 21.4 (2000), pp. 1324–1342.

[160] J Douglas Carroll and Jih-Jie Chang. "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition". In: *Psychometrika* 35.3 (1970), pp. 283–319.

[161] Richard A Harshman et al. "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis". In: *UCLA Working Papers in Phonetics* 16.1 (1970), p. 84.

[162] Henk AL Kiers. "Towards a standardized notation and terminology in multiway analysis". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 14.3 (2000), pp. 105–122.

[163] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. "TotalSegmentator: robust segmentation of 104 anatomic structures in CT images". In: *Radiology: Artificial Intelligence* 5.5 (2023).

[164] Curtis P Langlotz. "Will artificial intelligence replace radiologists?" In: *Radiology: Artificial Intelligence* 1.3 (2019).

LIST OF FIGURES

EIDESSTATTLICHE VERSICHERUNG

(Siehe Promotionsordnung vom 12. Juli 2011, §8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 10.01.2025 Tobias Weber