Self-supervised Representation Learning for Trustworthy Ultrasound Video Analysis

Kangning Zhang

Wolfson College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity Term 2024

Abstract

Obstetric ultrasound (US) is a routinely used imaging modality for monitoring fetal development during pregnancy. With the rise of deep learning (DL)-based techniques, there is growing potential to reduce reliance on highly trained sonographers. However, the clinical adoption of DL models remains limited due to concerns around robustness, explainability, and reliability—key components of trustworthiness. This thesis addresses these challenges to advance trustworthy DL-based analysis of fetal US.

Given that manual annotation of US videos by expert sonographers is costly and time-consuming, we first investigate self-supervised learning as a scalable alternative. We propose a novel framework for feature-level local self-supervised learning, combining contrastive learning with Rubik's cube recovery, cube reconstruction, and a new strategy for generating local contrastive pairs. This method enhances the robustness of learned representations and demonstrates improved performance in two key downstream clinical tasks.

To improve model transparency, we incorporate clinician-driven insights by integrating clinical-driven insights into the explanation process. We define explainability as the model's ability to capture anatomy-aware knowledge. To evaluate this, we introduce a novel set of quantitative metrics that measure the alignment between learned representations and expert gaze-tracking data. These gaze-guided explanations provide a more clinically meaningful assessment of model behaviour, hence enhancing explainability.

While explainability provides valuable insights, reliable deployment is equally critical in clinical applications. To this end, we introduce the concept of out-of-clinical-distribution (OCD), defined as US frames lacking diagnostically meaningful content. We propose the first method for OCD detection in fetal US, enabling models to filter out irrelevant frames from large, heterogeneous real-time scans. This method enhances the safe and reliable deployment of DL models by ensuring that decisions are based only on clinically significant inputs.

In summary, this thesis presents novel methods to enhance the robustness, explainability, and reliability of DL models for fetal US analysis. All approaches are validated on retrospective clinical data, laying the groundwork for more trustworthy integration of DL technologies in prenatal care and inspiring future research in this domain.

Self-supervised Representation Learning for Trustworthy Ultrasound Video Analysis

Kangning Zhang
Wolfson College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity Term 2024

Acknowledgements

I would like to express my sincere gratitude to my supervisor and mentor, Professor J. Alison Noble, for her invaluable support and guidance throughout my PhD. Her wisdom and encouragement have shaped me both academically and personally.

I am also deeply thankful to my second supervisor, Professor Jianbo Jiao, for his consistent support, patience, and kindness. Their collective guidance has been instrumental in making this thesis possible.

My thanks also go to my collaborators, colleagues at IBME, and peers at the Health Data Science CDT for fostering a supportive and collaborative environment. Their feedback and suggestions have been invaluable to my research.

A special thanks goes to my friends who have stood by me during my PhD journey. In particular, I want to thank my wonderful housemate, Hang Yuan; my thesis writing buddy, Lin Qiu; my warmest girl, Yechuan Zhang; and my long-time friends, Yi Lin and Jiaqi Chen, for always being there for me.

Finally, to my family—thank you for your endless love and encouragement. To my mom, your positivity and resilience inspire me. To my dad, although you are no longer with us, I feel your presence every day. This thesis is dedicated to you.

Abstract

Obstetric ultrasound (US) is a routinely used imaging modality for monitoring fetal development during pregnancy. With the rise of deep learning (DL)-based techniques, there is growing potential to reduce reliance on highly trained sonographers. However, the clinical adoption of DL models remains limited due to concerns around robustness, explainability, and reliability—key components of trustworthiness. This thesis addresses these challenges to advance trustworthy DL-based analysis of fetal US.

Given that manual annotation of US videos by expert sonographers is costly and time-consuming, we first investigate self-supervised learning as a scalable alternative. We propose a novel framework for feature-level local self-supervised learning, combining contrastive learning with Rubik's cube recovery, cube reconstruction, and a new strategy for generating local contrastive pairs. This method enhances the robustness of learned representations and demonstrates improved performance in two key downstream clinical tasks.

To improve model transparency, we incorporate clinician-driven insights by integrating clinical-driven insights into the explanation process. We define explainability as the model's ability to capture anatomy-aware knowledge. To evaluate this, we introduce a novel set of quantitative metrics that measure the alignment between learned representations and expert gaze-tracking data. These gaze-guided explanations provide a more clinically meaningful assessment of model behaviour, hence enhancing explainability.

While explainability provides valuable insights, reliable deployment is equally critical in clinical applications. To this end, we introduce the concept of out-of-clinical-distribution (OCD), defined as US frames lacking diagnostically meaningful content. We propose the first method for OCD detection in fetal US, enabling models to filter out irrelevant frames from large, heterogeneous real-time scans. This method enhances the safe and reliable deployment of DL models by ensuring that decisions are based only on clinically significant inputs.

In summary, this thesis presents novel methods to enhance the robustness, explainability, and reliability of DL models for fetal US analysis. All approaches are validated on retrospective clinical data, laying the groundwork for more trustworthy integration of DL technologies in prenatal care and inspiring future research in this domain.

Contents

List of Figures

List of Abbreviations

1-D, 2-D . . . One- or two-dimensional, referring in this thesis to spatial dimensions in an image.

3D . . . . . 3D-mode in scanning

3VT . . . . . Three-vessel and trachea view

4CH . . . . . Four-chamber view

ACL . . . . . Anterior cruciate ligament

AI . . . . . Artificial intelligence

BPD . . . . . Biparietal diameter

BrainTc. . . . Brain view at the level of the cerebellum

BrainTv. . . . Brain view at the posterior horn of the ventricle

BYOL . . . . . Bootstrap your own latent

CIR . . . . . Clinically interesting region

CNN . . . . . Convolutional neural network

CPC . . . . . Contrastive predictive coding

CPU . . . . . Central processing unit

CRCR . . . . . Contrastive Rubik's Cube Recovery

CRL . . . . . Crown rump length

CT . . . . . Computed tomography

CVRL . . . . . Contrastive voxel-wise representation learning

CXR . . . . . Chest X-rays

DC . . . . . Decompressing craniectomy

DDPM . . . . Denoising diffusion probabilistic model

DeepSMILE . Deep multiple instance learning

DINO . . . . . Distillation with NO labels

DiRA . . . . . Discriminative, restorative, and adversarial learning

DL . . . . . Deep learning

ELBO . . . . . Evidence of lower bound

FASP . . . . . Fetal anomaly screening program

GAN . . . . . Generative adversarial network

GE . . . . . General Electric

GPU . . . . . Graphical processing unit

Grad-CAM . . Gradient-weighted Class Activation Mapping

H&E . . . . . Hematoxylin and eosin

HCI . . . . . Human-computer interaction

HiCo . . . . . Hierarchical contrastive

HRD . . . . . Homologous recombination deficiency

ICD . . . . . In-clinical distribution

KL . . . . . Kullback–Leibler

LDM . . . . . Latent diffusion models

LVOT . . . . . Left ventricular outflow tract

MAP . . . . . Maximum a posteriori

MICLe . . . . Multi-instance contrastive learning

MIL . . . . . Multiple instance learning

ML . . . . . Machine learning

MoCo . . . . . Momentum contrast

MRI . . . . . Magnetic resonance imaging

MSE . . . . . Mean square error

MSI . . . . . Microsatellite instability

MVG . . . . . Medical vision generalist

N2V Noise2Void

NF . . . . . Normalising flows

NHS . . . . . National Health Service

NIST . . . . . National Institute of Standards and Technology

NT . . . . . Nuchal translucency

OCD . . . . . Out-of-clinical distribution

OOD . . . . . Out-of-distribution

PDF . . . . . Probability density function

PGL . . . . . Prior-guided local

PULSE . . . . Perception ultrasound by learning sonographer experience

RCR . . . . . Rubik's cube recovery

RINCE . . . . Robust InfoNCE

RVOT . . . . . Right ventricular outflow tract

SAM . . . . . Self-supervised anatomical embedding

SLIC . . . . . Simple linear iterative clustering

SimCLR . . . Simple framework for contrastive learning of visual representations

SimSiam . . . Simple siamese

SpineCor. . . . Coronal spine view

SpineSag. . . . Sagittal spine view

SSL . . . . . Self-supervised representation learning

SW-MSA . . . Shifted window-based multihead self-attention

Swin UNETR Swin U-net transformer

TBI . . . . . Traumatic brain injury

TCPC . . . . . Task-related CPC

t-SNE . . . . . t-distributed stochastic neighbour embedding

U1D . . . . . Union of 1-dimension

US . . . . . Ultrasound

USCL . . . . . US semi-supervised contrastive learning

VarMIL . . . . Variability-aware deep multiple instance learning

ViT . . . . . Vision transformer

WCE . . . . . White cell enzymes

WHO . . . . . World health organisation

W-MAS . . . . Window-based multihead self-attention

WSI . . . . . Whole slide image

XGrad-CAM . Axiom-based Grad-CAM

The sound of chopping wood echoes clearly, as the birds chirp melodiously. From deep within the valley, they fly up to the lofty trees.

— *The Book of Songs, Minor Odes of the Kingdom, Cutting Wood.*

1 Introduction

Contents

1.1 Clinical Motivation

Obstetric ultrasound (US), in which the sound waves are used to create a real-time visual image of the fetus in the uterus, is a routinely used imaging modality for monitoring and diagnosing the health of the fetus in prenatal care [1, 2].

In general, obstetric US screening involves manual scanning, standard plane detection, biometric measurement, and diagnosis [3]. In practice, US screenings are typically performed by sonographers with careful but often labour-intensive probe movements. The quality of these screenings heavily depends on the sonographer's skill level. However, even for experienced sonographers, guiding the transducer to the correct plane through variable anatomy, performing biometric measurements, and making diagnoses based on those measurements is a complex task due to human bias and low reproducibility [4]. Additionally, the long training period, insufficient recruitment, and poor retention of sonographers bring challenges to the widespread

use of US scans [5], limiting pregnant women's access to prenatal care. As a result, there is growing interest in applying deep learning-based technologies to reduce the dependency on highly trained sonographers. Leveraging deep learning on large-scale clinical US data could provide sonographers with valuable insights, transform US video analysis, and possibly simplify the use of US for non-experts.

Despite the recent success of applying deep learning models to fetal US analysis, several challenges arises and are becoming more prominent. Firstly, most prior work in this field has been conducted in a supervised manner [6], relying heavily on the availability of large, annotated datasets. However, the labelling process for fetal US data requires expert clinical knowledge, making it both time-consuming and costly. In addition, due to the length of the scans, it is often impractical to label entire videos [7]. This highlights a practical need for novel DL approaches that can learn anatomically meaningful representations directly from raw data without extensive manual labelling—namely, self-supervised representation learning (SSL).

Secondly, to enable safe and effective deployment of DL models in clinical practice—particularly in fetal US analysis—models must satisfy three key dimensions of trustworthiness: robustness, explainability, and reliability. Robustness is vital, as fetal US images are often degraded by speckle noise and acoustic shadows [8], which can negatively impact both image quality and model performance. A robust model should be capable of extracting features that are invariant to such noise and artefacts. Explainability is equally critical, as prenatal diagnostic decisions carry significant consequences, directly impacting both maternal and fetal health. Clinicians and patients must be able to understand and trust model outputs, especially when these outputs are used to support real-time decisions or navigation. Without explanations that align with clinical reasoning, even accurate predictions may be disregarded. Finally, reliability is essential for analysing fetal US videos, where a substantial proportion of frames may be clinically irrelevant—such as when only amniotic fluid is visible. A reliable model must be able to detect and abstain from making predictions on such out-of-clinical-distribution (OCD) inputs and ensuring that predictions are

grounded in clinically meaningful content. Together, these three dimensions form the foundation for building safe and trustworthy DL tools for real-time US analysis.

Therefore, this thesis aims to address these two critical challenges that deep learning models face in real-world clinical applications—limited annotations and lack of model trustworthiness—by exploring SSL and trustworthy learning, applied to real-time fetal US video data. I propose an SSL framework tailored to the unique properties of US videos and designed to be robust to US-specific noise, enabling the learning of semantically meaningful representations from raw data. I introduce a set of quantitative explainability metrics in the context of video US and gaze tracking information to enhance model explainability. In addition, I incorporate OCD detection to evaluate model reliability in distinguishing clinically relevant from irrelevant inputs. It is our hope that the models developed, the concepts proposed, and the evaluation metrics introduced will contribute meaningfully to future research and clinical applications in this field.

1.2 Contributions and Publications

- I proposed a feature-level local-contrastive self-supervised representation learning approach for fetal ultrasound video analysis. Content associated with this contribution has been published in ASMUS, MICCAI 2024 (oral presentation) [9]. This work has been recognised with **best paper award** and **best presentation (runner-up) award**.

- I defined model explainability within the context of video US and gaze tracking information and introduced a novel set of quantitative metrics for model explainability. Content associated with this contribution has been accepted as extended abstract in Medical Imaging meets NeurIPS, 2021(poster session).

- I introduced the concept of "out-of-clinical distribution" (OCD) detection, which distinguishes the clinical relevance of data, and developed an OCD detection method to fetal US application. Content associated with this

contribution has been published in MICAD 2024 (oral presentation) [10]. This work has been recognised with **best paper award**.

1.3 Thesis Structure

Here in **Chapter 1**, I discuss the motivation for the doctoral research conducted and described in this thesis. Our main contributions are summarised and will be discussed in later chapters.

Chapter 2 is the literature review chapter. In this chapter, I begin by discussing how deep learning has advanced medical imaging analysis, along with the challenges it faces. I then introduce self-supervised learning as a method to address these challenges and identify the common pre-training tasks used in SSL for medical imaging analysis. Following this, I provide a brief review of trustworthiness in deep learning—another critical challenge—from three key dimensions: robustness, explainability, and reliability. For each dimension, I discuss its definition, significance in the context of clinical applications, and commonly used methods to address it. These topics are directly related to the work presented in this thesis.

Chapter 3 describes the datasets used in this thesis. In this chapter, I introduce the data acquisition process for fetal ultrasound videos, outline the preprocessing protocols, and explain why the dataset is well-suited for self-supervised learning. This chapter provides a comprehensive overview of each dataset utilised in the research works discussed in the following chapters in this thesis.

Chapter 4 focuses on developing a novel self-supervised representation learning approach for robust local representation learning of fetal ultrasound video. A feature-level Contrastive Rubik's Cube Recovery (CRCR) approach is proposed, which combines contrastive learning with two different pre-text tasks, Rubik's cube recovery and cube reconstruction. A local contrastive pairs generation strategy is introduced to provide strong local feature discrimination. This chapter provides detailed insights into specific approach I have built, the evaluation results I have achieved and the extensive ablation studies I have carried on.

Chapter 5 explores deep learning explainability in the context of fetal US and gaze tracking information. I introduce a new definition of deep learning explainability in the clinical setting, which focus on the capturing anatomy-aware knowledge. I propose a novel set of quantitative metrics to evaluate explainability, based on visually salient landmarks—features that naturally draw clinicians' attention—thereby incorporating clinical domain knowledge into model interpretation. These metrics assess the alignment between the model's focus and the sonographer's gaze patterns during US interpretation. Comprehensive experiments are presented in this chapter to demonstrate the efficacy of the proposed explainability metrics on various SSL approaches.

Chapter 6 introduces out-of-clinical distribution detection as a method for assessing the reliability of deep learning models in medical imaging analysis. Firstly, I introduce the concept of "out-of-clinical distribution", as a new definition of OOD under the clinical setting, to distinguish medical data that contains clinically relevant regions from those that lack clinically meaningful information. I then propose a novel detection method that incorporates a softmax-conditioned variational autoencoder. Both quantitative and qualitative results demonstrate the effectiveness of the proposed method on fetal ultrasound data.

Chapter 7 is the conclusion chapter, where I summarise the contributions of this thesis and discuss how it may open up potential directions for future research.

The wild geese are flying, their wings rustling as they soar. That man is on a journey, toiling laboriously in the fields.

– *The Book of Songs, Minor Odes of the Kingdom, The Wild Geese.*

2 Literature Review

Contents

2.1 Introduction

In the first part of this literature review, I explore the relevant literature on SSL, with a particular emphasis on its applications in medical imaging analysis. I provide supporting evidence for why current deep learning trends are moving towards SSL in medical imaging analysis. Additionally, I outline common SSL techniques applied to both natural and medical images/videos.

The second part of this review focuses on trustworthy deep learning, an

increasingly important concern, particularly in critical areas such as healthcare. This section explores the concept of trustworthiness by examining its theoretical foundation and delving into three key dimensions—robustness, explainability, and reliability. For each, I discuss its definition, clinical significance, and relevant methodologies. Together, these perspectives offer comprehensive insights that inform and guide the design of my subsequent research.

2.2 Self-supervised Learning in Medical Imaging Analysis

This section is split into three subsections. Subsection 2.2.1 introduces some preliminary knowledge on deep learning in medical imaging analysis. Subsection 2.2.2 discusses the importance of introducing self-supervised learning in medical imaging analysis and its benefits over the existing deep learning methods. Subsection 2.2.3 introduces some current widely-used self-supervised learning methods that have been applied to medical imaging analysis.

2.2.1 Deep Learning (DL) in Medical Imaging Analysis

Machine learning (ML), is an interdisciplinary field that targets to construct algorithms that could learn from and make predictions based on data [11, 12]. Deep learning, a subset of machine learning, has emerged as a powerful tool by integrating the feature extraction step directly into the learning algorithm [13]. It enables models to automatically process and learn low-level to high-level abstract features from data, including images, videos, and other inputs. This approach has demonstrated state-of-art performance in a variety of applications, such as natural language processing and computer vision [14–16]. In addition, recent remarkable technical progress in (1) data size: the global rise of "big data," (2) computing power: advancements in central processing units (CPUs) and graphical processing units (GPUs), and (3) new deep learning algorithms [17] has also enabled individual researchers to access the necessary resources, which further advancing the field.

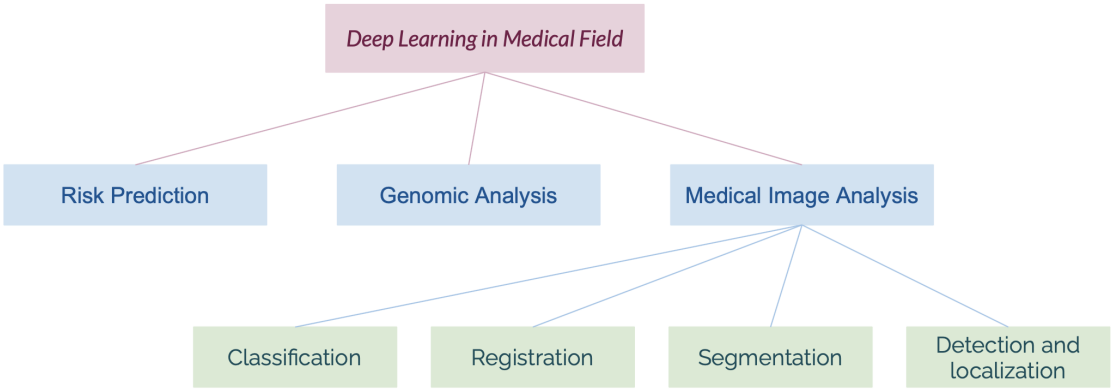


Figure 2.1: Overview of key DL application categories in the medical field and medical imaging analysis, with representative categories selected from the broader field based on [18] and [19].

According to [18], deep learning (DL) in the field of medical imaging analysis offers promising results in three main areas: (1) risk prediction; (2) genomic analysis and phenotype-genotype association studies; and (3) medical imaging analysis, as illustrated in Fig. 2.1. Here, I focus on medical imaging analysis, which primarily involves processing and analysing medical images from various modalities (e.g., computed tomography (CT), magnetic resonance imaging (MRI), and US) to extract valuable information that can aid in precise decision-making and support diagnosis [20]. Medical imaging analysis, adapted from computer vision tasks, includes several categories. Here, I highlight four key tasks according to the categories in [19]: classification, segmentation, detection and localisation, and registration, as shown in Fig. 2.1. The success of computer vision in natural image analysis has introduced effective methodologies and yielded promising results for medical applications, but it also presents new challenges. For instance, medical images often have extremely high resolution, contain fewer and smaller regions of interest, and exhibit subtle differences between object classes [21], making analysis more complex. US, as one of the core diagnostic imaging modalities, also offers unique challenges, including sonographer dependency, noise, limited imaging structure behind bone and air, artefacts, and fan-shape images [17]. Hence, the applications of DL to US imaging still lag behind other modalities like CT and MRI, however, it has rapidly progressed.

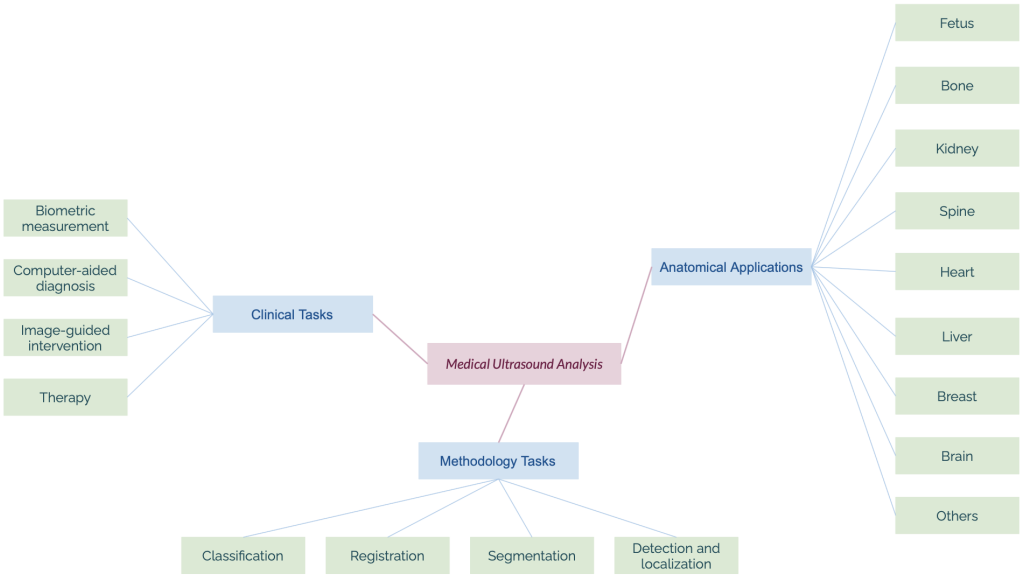


Figure 2.2: Demonstration of selected key DL applications in Medical Ultrasound Analysis. This figure is motivated by [38].

Fig. 2.2 illustrates selected recent applications of deep learning in medical US analysis, which have involved both basic methodology tasks including classification [22], segmentation [23], detection [24] and registration [25], as well as emerging clinical tasks including biometric measurements [26], computer-aided diagnosis [27], image-guided interventions [28] and therapy [29]. These tasks have been applied to various anatomical structures, including the fetus [30], bone [31], kidney [32], spine [33], heart [34], liver [35], breast [36], brain[37], and more. In this thesis, I focus on fetal US, which is considered the most important tool for assessing pregnancy health.

2.2.2 Steps to Self-supervised Learning

Despite the remarkable success that deep learning has achieved in medical imaging analysis, the majority of the literature focuses on the supervised learning paradigm [39], where models are trained on datasets that include a large number of medical images and their corresponding annotated labels to learn relevant patterns in the data. However, the heavy reliance on annotated datasets has become a major concern, limiting further advancements in the field. Building large, high-quality annotated medical imaging datasets is both expensive and time-consuming, and in

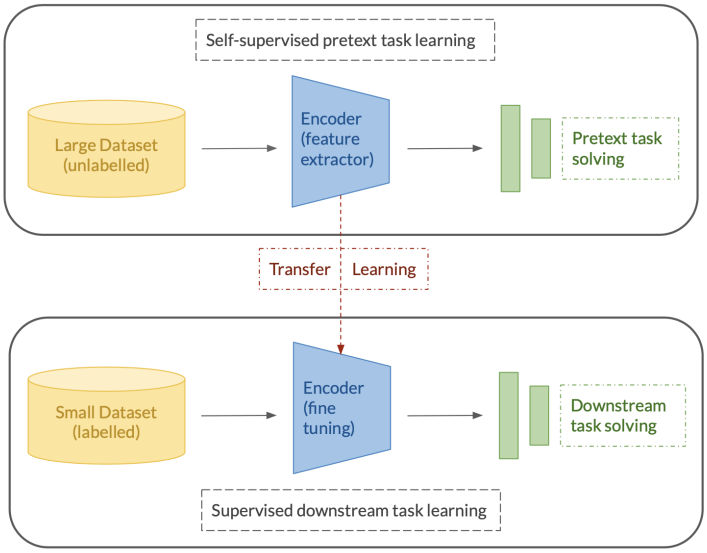


Figure 2.3: Self-supervised learning main workflow. (top): Self-supervised learning scheme is applied by training an auxiliary task from a large unlabelled dataset. (bottom): The learned representations are transferred from the pretext task to the down-stream task to accomplish the training on a small amount of data with ground truth labels

some cases, not feasible for two main reasons. Firstly, unlike natural image datasets, which can be annotated by individuals with minimal expertise, medical datasets require experts with domain-specific knowledge for accurate annotation. Secondly, the annotation process is complicated by patient privacy concerns, particularly when dealing with specific medical conditions [40].

Self-supervised learning, in which the learning process is supervised based on raw data itself, is an emerging solution to address challenges caused by difficulties in curating large-scale annotations [41]. Fig. 2.3 illustrates the general pipeline of self-supervised learning, which usually consists of two steps: (1) pre-train a CNN with a pre-defined task on a large set of unlabelled data, and (2) fine-tune the pre-trained network for the high-level downstream task with a small set of annotated data. The core of the approach is based on the principle of providing "labels for free" from unlabelled raw data. The informative representations could be manually extracted by learning the objective function of certain "pretext tasks". Such a "pretext task" is pre-defined and trained in a supervised manner. The representation learned by performing "pretext task" is transferred to downstream task training and serves as a starting point for fine-tuning. It is assumed that if a

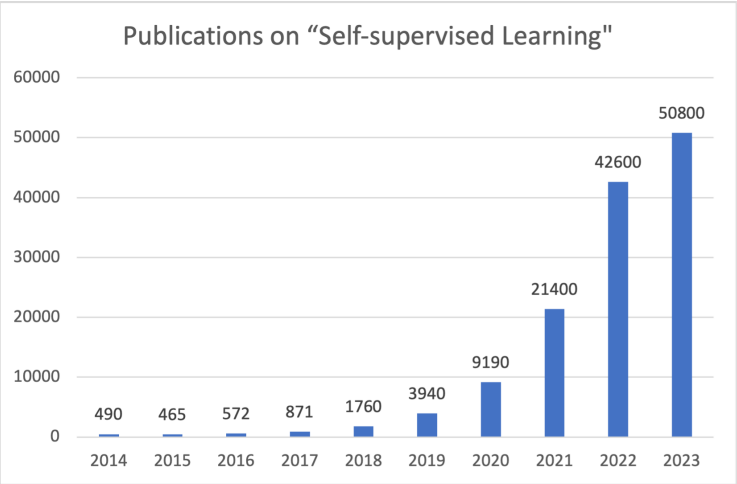


Figure 2.4: Timeline showing the number of publications on “self-supervised learning” based on the results from Google Scholar search. The vertical and horizontal axes denote the number of publications and the year, respectively

“pretext task” is well-designed, the model could learn meaningful and transferable features, leading to improved performance on the downstream task [41].

SSL was initially popularised in the field of natural language processing (NLP) [42] and later extended to a variety of computer vision tasks [43], where it has achieved state-of-the-art performance. The development of SSL has seen rapid expansion, attracting significant attention from the research community. Fig. 2.4 shows the number of publications on “self-supervised learning” per year over the past decade, based on the inclusion criterion that the keyword “self-supervised learning” appears in the title or abstract. As depicted, Google Scholar reports a substantial increase in SSL-related publications, with an average of 140 papers published per day in 2023 — a remarkable indicator of the field’s growing momentum. Fig. 2.5 presents the number of publications on “deep learning for medical imaging classification”, alongside the number of publications specifically focused on “self-supervised learning for medical imaging classification”. The inclusion criterion is similar to the one mentioned earlier, requiring the keyword to appear in the title or abstract. The figure highlights that both topics have experienced significant growth over the past five years, with self-supervised learning emerging as a rapidly expanding subset within the broader deep-learning literature in the area of medical imaging classification.

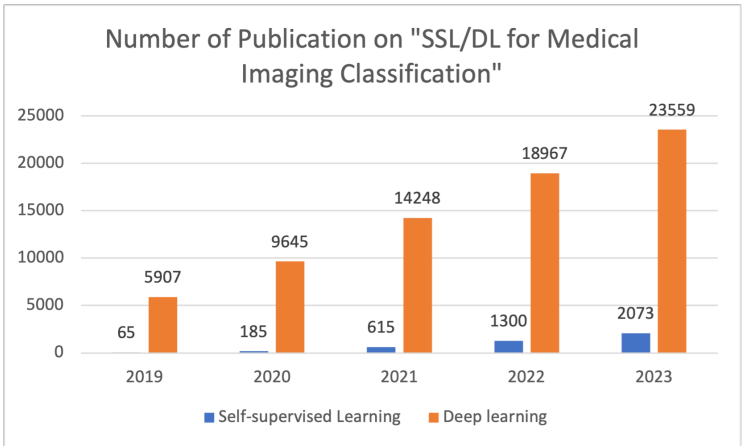


Figure 2.5: Timeline illustrating the number of publications per year on both deep learning and self-supervised learning for medical image classification respectively, using the same search criteria across PubMed, Scopus, and arXiv. The vertical axis represents the number of publications, while the horizontal axis denotes the corresponding year.

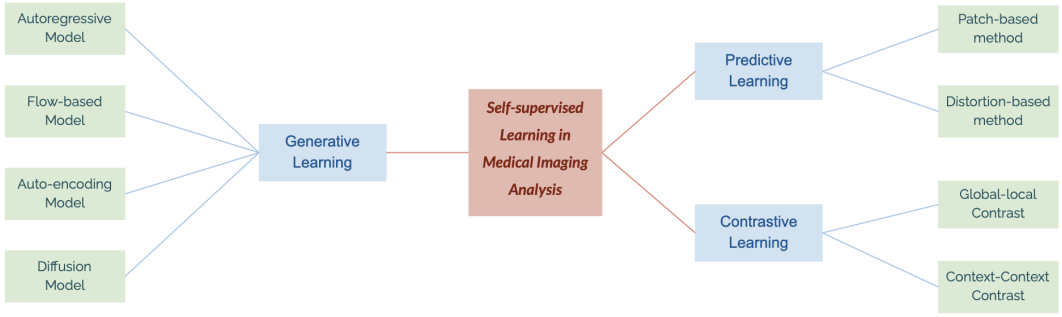


Figure 2.6: Illustration of different self-supervised pre-training methods

2.2.3 Self-supervised Learning Approaches

SSL is a form of representation learning where the objective function is derived from a pretext task, which can be solved using large amounts of unlabelled data. As such, pretext tasks play a critical role in SSL approaches and serve as their backbone. The high-level representations learned from these pretext tasks are then applied to subsequent downstream learning tasks, where only a small amount of labelled data is available. While downstream tasks may vary according to specific research needs, pretext tasks could be common across different tasks.

As shown in Fig. 2.6, three main categories of SSL approaches are reviewed based on the nature of the pretext tasks: predictive, contrastive, and generative

Category	Goal	Output	Examples
Predictive Learning	Infer or classify a specific structured outcome	Fixed label or value	Rotation prediction, order prediction
Contrastive Learning	Distinguish between similar and dissimilar pairs	Similarity score	SimCLR, MoCo
Generative Learning	Generate or reconstruct plausible data	High-dimensional sample	Autoencoders, GANs, MIM

Table 2.1: Summary of predictive, contrastive, and generative learning

self-supervised learning. To clarify these distinctions, Table 2.1 summarises how these three categories are defined based on their goal, output, and focus, which is motived by [44]. As indicated in the table, predictive learning focuses on structured output prediction, contrastive learning aims to distinguish between positive and negative pairs to learn structured embeddings, and generative learning focuses on producing new high-dimensional samples or reconstructing missing content that is consistent with the underlying data distribution.

There are mainly two paths to follow when employing self-supervised learning approaches in medical imaging analysis [44]. Firstly, to adapt approaches trained and tested on natural images. Given the significant differences between medical and natural images, challenges remain in effectively adapting existing SSL frameworks to address tasks in medical imaging analysis. Secondly, designing novel approaches exploits knowledge from the medical domain and unique anatomical structures from the medical image. In this section, I review SSL approaches that have been applied in the field of medical imaging analysis in both manner.

Predictive Self-supervised Learning

The predictive self-supervised learning approach aims to learn robust representations by assigning pseudo labels to each unlabelled image and framing the pretext task as a classification or regression problem [45]. The output is typically a well-defined label or a measurable value, such as predicting a class label, rotation angle. The key

assumption is that by solving the structured prediction problem that leverages the internal patterns or correlations in the data, the model is encouraged to learning meaningful representations. It is noteworthy that pretext tasks and output must be carefully designed and generated to ensure effective representation learning. Additionally, these labels are typically only effective for the specific pretext task they were generated for, offering limited utility when directly applied to downstream tasks.

Many predictive pretext tasks have been designed or adapted in the field of medical imaging analysis. As demonstrated in Table 2.2, I provide a detailed illustration of several of these methods, which could be categorised into two categories: patch-based method and distortion-method. Here, based on the defined categorisation, predictive tasks always involve a fixed output (e.g., a label or a score). Therefore, tasks like image reconstruction and masked image modelling are not considered predictive, as the output is not a fixed structured label but a generated high-dimensional sample that matches the learnt data distribution.

In a patch-based predictive pretext task, the model is tasked with predicting relationships, dependencies, or missing information at the level of image patches or local regions. The focus is on capturing spatial correlations and learning fine-grained local patterns within the image.

1. **Patch-based method:** In a patch-based pretext task, the model is designed to predict the relationships, dependencies, or missing information between image patches. This relies on spatial context semantics, assuming that the model needs to understand the spatial structure of the object to accurately determine the position or identity of each patch. Examples include relative locations prediction [46], which is defined as the problem of predicting the relative location of random image patches from the chosen anchor patch, and "jigsaw puzzle" solving [47], which is defined as the problem of recognizing the order of image patches after random shuffling.

   Inspired by relative position prediction task [46], Zhang et al. [48] introduced slice ordering as a pretext task for body part recognition using MRI scans.

By slicing the 3D MRI scans into groups of successive 2D slices, the task is designed to predict the spatial order (i.e., whether one slice is below or above) of two given successive slices. A new Siamese convolutional architecture, named Paired-CNN is designed as a pre-training network and a fine-grained body part recognition is used as a downstream task. Bai et al. [49] proposed a novel pretext task, anatomical position prediction, for cardiac MR scan segmentation by leveraging the rich information encoded in the various cardiac MR scan view planes. A standard cardiac MR scan provides images acquired at various angulated planes relative to the heart, including short-axis, long-axis 2-chamber, 4-chamber, and 3-chamber views. A set of anatomical positions is defined with respect to these cardiac views, and the pretext task is designed to predict these anatomical positions, utilising the spatial and structural information present in the different views.

Adapting the same logic as the original Jigsaw puzzle solving task [47], Zhuang et al. [50] proposed a novel pre-text task for 3D medical data, named Rubik cube recovery. The 3D input volume is partitioned into a grid (i.e. $2 \times 2 \times 2$) of sub-cubes, which are then randomly shuffled and rotated. The pretext task is designed to recover the original configuration, involving two operations: cube rearrangement and cube rotation. As a result, the task enables the model to learn both translational and rotational invariant features, which are tested on two downstream tasks including brain hemorrhage classification and brain tumour segmentation. Building on the previous work, Zhu et al. [51] introduced the Rubik cube+ pretext task by adding a cube masking operation to increase the difficulty, in which the content of the cube is randomly blocked. Manna et al. [52] designed a pretext task where the model predicts the correct ordering of jumbled image patches into which MRI video frames are divided. A novel CNN architecture is designed to efficiently solve jigsaw puzzles and learn explainable representational features. These features are then used to detect anterior cruciate ligament (ACL) tear injuries sustained in the knee of a human in the downstream task. Taleb et al. [53] made significant advancements in

multi-modal settings by introducing a pretext task called the multi-modal Jigsaw puzzle. In this approach, T1 and T2 scans are simultaneously used as multiple imaging modalities, and a Sinkhorn network [54] is used as the backbone for solving the Jigsaw puzzle pretext task and for the downstream segmentation task.

2. **Distortion-based method:** In a distortion-based pretext task, the model is designed to predict a specific transformation or distortion applied to the input data, assuming that global consistency and semantic-level understanding are learned in the process. Geometric transformation serves as a common kind of modification. Geometric transformations serve as a common type of modification. These transformations, such as rotation, translation, and scaling, are used to help models learn spatial relationships and invariant features within images.

   Long et al. [55] used image rotation prediction as a pretext task to recognise the types of rotations applied to input images. COVID-19 classification was selected as the downstream task. The assumption is that to accurately predict image rotations, the network must learn as many semantic features of the CT images of COVID-19 as possible. Vats et al. [56] proposed adapting both jigsaw puzzle-solving and rotation prediction as pretext tasks for feature learning for pathology images. These methods were chosen to learn representations of anatomical concepts, such as pathology colour and texture, from the original and relatively undistorted views of images. The effectiveness of these approaches was tested on downstream pathology classification tasks (white cell enzymes (WCE) diagnosis). Jiao et al. [57] designed a joint reasoning task for pre-training to learn anatomy-aware representations: correct the order of a reshuffled US video clip and predict the geometric transformation applied to the US video clip simultaneously. This task is tested to be effective on two downstream tasks, standard plane detection and saliency prediction.

Year	Authors	Imaging Modalities	Pre-text task	Down-stream task
2017	Zhang et al. [48]	3D MRI scanning	Slices ordering	Body parts recognition
2019	Bai et al. [49]	Cardiac MRI scan	Anatomical position prediction	Short-axis cardiac MRI segmentation
2019	Zhuang et al. [50]	3D MRI scanning	Rubik cube	Prostate segmentation
2020	Zhu et al. [51]	3D MRI scanning	Rubik cube+	Brain hemorrhage classification
2022	Manna et al. [52]	3D MRI scanning	Jigsaw puzzle	ACL tear injuries classification
2021	Taleb et al. [53]	FMRI scanning	Jigsaw puzzle	Skin lesions segmentation & Brain tumor segmentation
2021	Long et al. [55]	CT images	Rotation prediction	COVID-19 classification
2021	Vats et al. [56]	Pathology images	Rotation prediction, Jigsaw puzzle	pathology classification tasks
2020	Jiao et al. [57]	Ultrasound video	Geometric transformation prediction & Slice order prediction	Standard plane detection & Saliency prediction

Table 2.2: Predictive self-supervised learning applications in medical imaging analysis

Contrastive Self-supervised Learning

Contrastive learning is a self-supervised learning approach that learns useful and transferable representations by training the model to differentiate between similar and dissimilar data pairs. The fundamental assumption behind this method is that variations caused by transformations of the input image should not alter its semantic meaning. As a result, different augmentations of the same image should have similar representations, referred to as "positive pairs," while representations for different images and their augmentations should differ, known as "negative pairs."

The pre-training model is optimised to pull together the representations of positive pairs and push away the representations of negative pairs. This process minimises the distance between positive pairs and maximises the distance between negative pairs in the latent space. An arbitrary distance function is selected for the contrastive loss function to achieve this optimisation.

As demonstrated in Table 2.3, this section introduces applications of contrastive-based approaches in the field of medical imaging analysis, which can generally be divided into two categories: context-context contrast and global-local contrast.

1. **Context-context contrast:** Context-context contrastive learning aims to model the relationship of different samples. One pioneering context-context contrast method is SimCLR, a simple framework for contrastive learning of visual representations [58], which outperformed supervised models on the ImageNet benchmark using 100 times fewer labels. To address the large batch size required to generate negative samples, Momentum Contrast (MoCo) [59] introduced a momentum-encoded queue to keep negative samples. Methods such as DINO (Distillation with NO labels)[60], BYOL (Bootstrap Your Own Latent) [61], and SimSiam (Simple Siamese) [62], further eliminates the need for negative samples. Contrastive predictive coding (CPC) [63], on the other hand, learns the representation of spatial information by predicting the future in latent space using powerful autoregressive models. These methods

have achieved impressive performance in the natural image domain and have inspired further research in the medical imaging field.

- *SimCLR:* Inspired by SimCLR, Azizi et al. [64] proposed Multi-Instance Contrastive Learning (MICLe). The main adaption is the positive pairs construction. It not only includes augmented views of the same image, but also leverages the availability of multiple views of the same medical pathology per patient. Such corrected views of the same patient are also considered positive pairs. The author tested MICLe on a Dermatology dataset with twenty-seven classes as a downstream task. Ciga et al. [65] employed SimCLR framework for histopathological image analysis. The authors conducted different downstream tasks and discovered that the combination of multiple multiorgan with several types of staining and resolution properties enhanced the quality of the learned features. Schirris et al. [66] proposed a Deep learning-based method to analyse whole slide images (WSIs) of Hematoxylin and Eosin (H&E) stained tumour tissue using Self-supervised pre-training and heterogeneity-aware deep Multiple Instance LEarning (DeepSMILE). DeepSMILE compresses a SimCLR-based contrastive strategy is used to pre-train a feature extractor on histopathology tiles and variability-aware deep multiple instance learning (VarMIL), which is extended from DeepMI [67] is used to model tumor heterogeneity. The proposed method is applied to the task of Homologous recombination deficiency (HRD) and microsatellite instability (MSI) classification. Inglese et al. [68] modified the optimisation algorithm of SimCLR to better fit 3D structural MRI data and applied it to distinguish two important diagnostically different Systemic lupus erythematosus (SLE) patient groups: NPSLE and non-NPSLE patients.
- *MOCO:* Sriram et al. [69] adapted MoCo to predict COVID patient deterioration based on chest X-rays (CXR). Non-COVID CXR images

are used in the pre-training stage due to the relative scarcity of COVID-19 patient data. Three tasks are defined to indicate COVID patient deterioration, including oxygen requirements prediction, deterioration prediction from a single image, and multiple images. The third task requires multiple time-indexed radiographs, where a new transformer-based architecture is proposed to process sequences of multiple images. Building on the advantages of MoCo, Sowrirajan et al. [70] proposed a MoCo-CXR (CXR) model for CXR classification problem. Since not all augmentation strategies implemented in the original MoCo paper are suitable for grayscale X-ray images, they adjusted data augmentation used in MoCo to obtain high-quality representation, which were then used to detect pathologies across different CXR datasets. Vu et al [71] developed a method, MedAug, which leverages patient metadata to select positive pairs for contrastive learning. They applied this approach to CXRs for the downstream task of pleural effusion classification. By using patient metadata, positive pairs could be selected from views of possibly different images that come from the same patient, imaging study, or laterality, which medical knowledge to the learning algorithm and are expected to be rich in pathological features.

- *BYOL:* Based on the framework of BYOL, Xie et al. [72] introduced the Prior-Guided Local (PGL) algorithm for 3D medical images' segmentation, which focuses on minimising local consistency loss based on the spatial and regional location relationship of the two augmented views of the same region. A prior-guided aligner was proposed and added on top of the projection head for both online and target networks. This aligner leverages the spatial relationship between different views based on the prior information of augmentation transformation and aligns the features during the learning process.

- *CPC:* Extended from the early work on CPC, Zhu et al. [73] developed a new method called Task-related CPC (TCPC) to handle 3D medical

data. The method firstly uses Simple Linear Iterative Clustering (SLIC) to detect the potential lesion areas, then crops and uses the sub-volumes surrounding the generated supervoxels as input to the TCPC encoder. A U-shape path is employed with a larger cube size, considering the fewer differences in content between medical images compared to natural images. Brain hemorrhage classification and lung nodule classification tasks were utilized as downstream tasks. Lu et al. [74] combined CPC with multiple instance learning (MIL) [67] for the classification of breast cancer histology images. CPC was first used to learn semantic features from raw breast cancer histopathological images, and then MIL was applied for classification without the need for patch-level annotations, allowing for efficient learning.

2. **Global-local contrast** Global-local contrastive learning aims to model the relationship between both local semantic representations and global representations. In medical imaging, data from different subjects may exhibit inherent consistency due to the structural uniformity of organs. For instance, in MRI and CT scans, the same anatomical region across different subjects tends to have similar content. In such cases, contrastive strategies based solely on data augmentation may not adequately capture the similarities present across different images.

   To address this, researchers have begun incorporating local contrast into contrastive strategies, operating under the assumption that encoding the same anatomical region in different medical images should produce similar embeddings. Chaitanya et al. [75] proposed a domain-specific contrastive strategy for volumetric medical images, that combines the proposed global and local strategies. The global contrastive strategy uses similarity across corresponding slices in different volumes as an additional cue and the local contrastive strategy encourages representations of local regions in an image to be similar under different transformations, and dissimilar to those of

other local regions in the same image. By integrating both global and local representations, this combined approach leads to further improvements in representation learning for MRI. Yan et al. [76] introduced a novel contrastive learning approach called Self-supervised Anatomical eMbedding (SAM). As human organs are intrinsically structured, the authors aimed to develop an algorithm capable of learning from unlabelled medical images to detect the semantic location of a pixel, referred to as body part. To capture both global and local information, a coarse-to-fine architecture is designed with two-level embeddings. The global embedding is trained to distinguish every body part at a coarse scale. The local embedding is at the pixel level, encoding anatomical context information. A pixel-level contrastive learning framework is proposed to differentiate pixels corresponding to various body parts, focusing on smaller regions to capture finer features. The learned embeddings were extensively tested in various downstream applications, including landmark detection and lesion matching, across different imaging modalities such as 3D CT and 2D X-ray, and for various body parts like the chest and hand. You et al. [77] introduced a novel Contrastive Voxel-wise Representation Learning (CVRL) method, which consists of two contrastive learning strategies tailored for volumetric medical imaging segmentation tasks. The author argued that not all voxels within the same image are equally positive, given the presence of dissimilar anatomical structures within a single image. By combining a voxel-to-volume contrastive algorithm to obtain global information from 3D images, and a local voxel-to-voxel distillation approach to better utilize local signals in the embedding space, CVRL effectively learns both low-level and high-level features.

Generative Self-supervised Learning

The generative self-supervised learning approach refers to a pretext task in which a model is trained to generate new high-dimensional data or reconstruct missing content by learning the input data distribution [44]. The goal is to model the

Year	Authors	Imaging Modalities	Pre-text task	Down-stream task
2021	Azizi et al. [64]	Dermatology Images	MICLe	Skin lesion classification
2022	Ciga et al. [65]	Histopathological whole slide images (WSI)	SimCLR	Cancer cellularity score regression & Cancer/tissue classification & Cancer/tumor segmentation
2021	Schirris et al. [66]	H&E whole-slide images	DeepSMILE	HRD and MSI prediction
2022	Inglese et al. [68]	3D structural MRI	SimCLR	NPSLE and non-NPSLE patients classification
2021	Sowrirajan et al. [70]	CXR	MoCo-CXR	Pleural effusion Identification
2021	Vu et al [71]	CXR	MoCo	Pleural effusion classification
2021	Sriram et al. [69]	CXR	MoCo	Adverse event prediction from single and multiple images & oxygen requirements prediction
2020	Xie et al. [72]	3D CT image	PGL	Segmentation
2020	Zhu et al. [73]	3D Brain hemorrhage dataset	TCPC	Brain hemorrhage classification & Lung nodule classification
2020	Lu et al. [74]	Breast cancer histopathological images	CPC, MIL	Breast cancer histology classification
2020	Chaitanya et al. [75]	3D MRI scanning	Global and local contrastive learning	Segmentation
2022	Yan et al. [76]	Chest CT & X-rays	SAM	Landmark detection & Lesion matching
2022	You et al. [77]	Left Atrium (LA) CT	CVRL	Segmentation

Table 2.3: Contrastive self-supervised learning applications in medical imaging analysis

underlying data distribution so that the model can produce coherent outputs resembling the input data. Unlike predictive learning, generative learning is open-ended, meaning that multiple plausible solutions may exist for a given input. This approach has gained popularity with the advancement of traditional autoencoders [78], variational autoencoders [79], and generative adversarial networks (GANs) [80].

As demonstrated in Table 2.4, this section briefly reviews the applications of generative self-supervised learning in the field of medical imaging analysis, which has been divided into four major categories: autoregressive model, flow-based model, auto-encoding model and diffusion model. This categorisation is followed by [81] and based on the underlying probabilistic frameworks that have been used to learn and sample from the data distribution. Each of these task involves model the underlying data distribution and generate new samples coherent from the distribution, aligning with the training objective of generative learning, which will be discussed further in the following sections.

1. **Autoregressive (AR) Model:** AR models are a type of machine learning model that predict the next value in a sequence based on previous values. These models assume that future values depend on past ones and use this dependency to make predictions. As a result, the joint distribution of the data is explicitly factorised into an ordered product of conditional distributions. Originally developed for natural language processing (NLP)—where, for example, the next word in a sentence is predicted based on preceding words—AR models have also been adapted for computer vision tasks. Models such as PixelRNN [82] and PixelCNN [83] model images pixel by pixel, where each pixel is generated by conditioning on the previously generated pixel. For instance, lower pixels in an image can be generated based on the information from upper pixels. Luo et al. [84] modelled the MRI reconstruction problem using an autoregressive model and Bayesian theory. In their approach, a PixelCNN++ model was employed as a data-driven MRI prior, providing a computationally tractable solution for reconstruction. PixelCNN++ was first trained on a dataset of MRI images to learn the pixel-wise joint probability

distribution. During reconstruction, this generative prior was incorporated within a Bayesian framework. The study demonstrated that the PixelCNN++ prior enhanced reconstruction quality by effectively capturing complex image intensity distributions. Ren et al. [85] proposed the Medical Vision Generalist (MVG), the first foundation model capable of handling various 2D medical imaging tasks such as cross-modal synthesis, segmentation, denoising, and inpainting. MVG unifies these tasks within a single image-to-image generation framework by standardising inputs and outputs as images and using a single-channel colouring scheme. A task prompt specifies the desired task, enabling consistent handling across tasks. To support both local and global context learning, MVG combines masked image modelling with autoregressive modelling—applying random masking to a concatenated square image and training the model to predict the next image in the sequence based on the prompt image-label pair. This hybrid approach delivers robust performance across all evaluated tasks and generalises well across four imaging modalities: CT, MRI, X-ray, and micro-ultrasound.

2. **Flow-based Models:** The goal of flow-based generative models is to estimate complex high-dimensional data distributions that are difficult to model directly. The core idea is to transform a simple distribution (e.g., a standard Gaussian) into the target data distribution using a sequence of invertible and differentiable transformations. Since the mapping is invertible, the transformed and original distributions must have the same dimensionality, and the transformations must be carefully designed to ensure that the Jacobian determinant is easy to compute. Models such as RealNVP [86] and Glow [87] achieve this by stacking invertible convolutional layers and affine coupling layers, progressively transforming the base distribution into the image distribution with data split into blocks. Van et al. [88] proposed an unsupervised US image despeckling and denoising method using a normalising flow prior trained on high-quality MRI images, eliminating the need for paired noisy and speckle-free US data. They formulated the US despeckling task as a maximum a posteriori (MAP)

optimisation problem under a deep generative image prior, learned from high-quality MRI data using normalising flows (NFs). Rather than focusing on high-level anatomical semantics, the model was trained on patches capturing local tissue structures from speckle-free MRI images. This cross-modality approach leveraged the tractable density estimation of flow-based models as a prior and successfully improved the image quality of US scans. Hajij et al. [89] explored the use of NFs as an alternative method for generating synthetic medical images, addressing limitations commonly associated with VAEs and GANs, such as complex hyperparameter tuning and the lack of an explicitly learned data probability distribution. Specifically, RealNVP was employed for medical image synthesis, and the quality of the generated images was evaluated using the Wasserstein distance and permutation tests. The results provide promising evidence of the potential of NFs for medical image generation.

3. **Auto-encoding (AE) Models:** AE models aim to reconstruct (part of) the input data from corrupted or incomplete inputs. As one of the most widely used generative models, AEs have numerous variants —such as variational autoencoders (VAEs) — due to their flexibility and adaptability. In the medical imaging field, AEs have been applied to tasks including image denoising, restoration, and reconstruction.

    - **Image Denoising:** Image denoising refers to the task of removing noise from a corrupted image to recover a clearer version. The model is trained to learn the underlying data distribution in order to fill into the corrupted pixels, producing a noise-free version of the original input image. Prakash et al. [90] adapted the image denoising approach as a pretext task for the segmentation of cell/nuclei images. A Noise2Void (N2V) network [91] was trained to denoise the entire dataset and then fine-tuned for segmentation. This method demonstrated superior performance in downstream task, even in scenarios with extreme levels of noise and limited training data.

- **Image Restoration:** Image restoration refers to the task of enhancing the quality of a degraded image by reversing transformations. The model is trained to learn the underlying data distribution in order to synthesise missing high-frequency details that are absent from the input, and to produce a restored version of the degraded image. While the task may involve elements of prediction—such as "predicting" a clearer version of the input—the model is in fact generating missing or enhanced visual information that remains consistent with the semantic content of the image. Unlike predictive tasks, which typically have a fixed and single correct output (e.g., predicting a rotation angle), image restoration can have multiple plausible outcomes. Therefore, based on the nature of its output and learning objective, image restoration is regarded as a generative task. Chen et al. [92] proposed a novel generative pretext task called context restoration. For a given image, two randomly selected isolated patches are swapped. By repeating this process, a corrupted version of the image is produced, preserving the intensity distribution but altering the spatial information. The generative model is then trained to restore the corrupted image back to its original version. This simple yet effective strategy learns semantic features useful for various subsequent image analysis tasks. The approach was tested on fetal standard plane classification, abdominal multi-organ localization, and brain tumor segmentation, demonstrating its effectiveness. Zhou et al. [93] developed a framework that trains generic, source models for 3D imaging. The model trained with the framework is called Generic Autodidactic Models, nicknamed Models Genesis. The framework integrates four kinds of transformations and shares the same autoencoder to restore the input. This learning process enables the model to obtain latent representations from multiple perspectives, such as learning appearance through non-linear transformations, texture through local pixel shuffling, and context through out-painting and in-painting. These learned representations are

then transferred to specific downstream tasks, including both 3D and 2D image classification and segmentation tasks.

- **Image Reconstruction:** Image reconstruction refers to the task of recreating an image or part of an image from the input data. The model is trained to generate content that is consistent with the surrounding context, producing a reconstructed version of the input image. A simple example is reconstruction using an autoencoder, where the encoder projects the input into a latent space, and the decoder reconstructs this representation back into the original image. The model is optimised by minimising the difference between the reconstructed output and the input image. While this task may seem predictive—since the model is "predicting" a reconstruction from the input—the primary goal is to generate new content that is coherent with the input data. This requires learning the global semantics of the image and modelling the underlying data distribution. Such emphasis on data generation, along with the nature of the output, aligns the task more closely with generative learning rather than predictive learning. Matzkin et al. [94] proposed a pre-text task to reconstruct the skull defect removed during decompressing craniectomy (DC) performed after traumatic brain injury (TBI) from post-operative CT images. A virtual craniectomy procedure is designed to simulate the effect of DC on full skulls and generate DC post-operative CT scans with bone flaps removed from different parts of the upper head. Two architectures, including U-Net [95] and denoising auto-encoder [96] were employed. Hervella et al. [97] introduced a multi-modal reconstruction task for retinal anatomy learning. The main assumption is that, in order to generate one image modality from another, the model must learn relevant patterns from the images to successfully perform the task. These learned patterns can then be transferred to solve downstream tasks in the same domain, such as the localization and segmentation of key anatomical structures in eye fundus retinography.

4. **Diffusion Models:** Diffusion models have emerged as one of the latest powerful classes of deep generative models, demonstrating state-of-the-art performance in image synthesis and video generation [98]. They work by defining a forward process that gradually adds Gaussian noise to an image until it becomes pure noise, and a learned reverse process that removes the noise step by step to reconstruct the original image. Once trained, the model can generate new data by applying the reverse process to randomly sampled noise. One of the most widely used formulations is the Denoising Diffusion Probabilistic Model (DDPM) [99], proposed by Ho et al., which uses two Markov chains: a forward chain that perturbs data into noise, and a backward chain that reconstructs data from noise, both conditioned on discrete time steps. Diffusion models are increasingly used in medical imaging for tasks such as synthetic data generation and image reconstruction. Pinaya et al. [100] explored the use of latent diffusion models to generate synthetic high-resolution 3D MRI images of the adult human brain. To improve efficiency, they employed compression models inspired by the architecture of Latent Diffusion Models (LDMs) [101], enabling generation in a lower-dimensional latent space. Image synthesis was further conditioned on variables such as age, gender, ventricular volume, and brain volume relative to intracranial volume to produce realistic brain scans. The results outperformed alternative GAN-based methods in the unconditioned setting. Domínguez et al. [102] proposed a physics-inspired diffusion model specifically designed for US image generation. The model incorporates a US-specific scheduler scheme inspired by the physical behaviour of sound wave propagation in ultrasound imaging. This tailored approach enables the diffusion model to generate more realistic US images, capturing characteristic features such as speckle patterns and depth-dependent textures. The results demonstrate that the proposed method effectively models the dynamics of US imaging during the image synthesis process.

Year	Authors	Imaging Modalities	Pre-text task	Down-stream task
2020	Luo et al. [84]	MRI images	PixelCNN++ model	MRI reconstruction
2024	Ren et al. [85]	CT images	Medical Vision	Cross-modal synthesis
		MRI images	Generalist (MVG)	& Segmentation
		X-ray images		& Denoising
		micro-US images		& Inpainting
2021	Van et al. [88]	US images	MAP optimisation on MRI data	US despeckling and denoising
2022	Hajij et al. [89]	MRI brain scan	RealNVP	Image synthesis
2020	Prakash et al. [90]	Microscopy data	Imaging denoising	Nuclei images' segmentation
2019	Chen et al. [92]	2D fetal US images	Context restoration	Standard plane classification
		& Abdominal CT images		& Organ localisation
		& Brain MRI		& Brain tumour segmentation
2019	Zhou et al. [93]	MRI	Models Genesis	Brain tumour/ liver/ lung nodule segmentation
		& X-ray images		& pulmonary diseases classification
		& US images		& RoI, bulb, and background classification
		& CT images		& FPR for pulmonary embolism
2020	Matzkin et al. [94]	CT images	Skull reconstruction	Bone flap volume estimation
2020	Hervella et al. [97]	Retinography image	Multi-modal	Fovea localization
		& Angiography image	reconstruction	& Optic disc localization
				& Vasculature segmentation
2022	Pinaya et al. [100]	3D MRI image	Latent diffusion models	3D image synthesis
2024	Domínguez et al. [102]	US images	Physics-inspired diffusion models	Image synthesis

Table 2.4: Generative self-supervised learning applications in medical imaging analysis

2.2.4 Conclusion

In this section, I have reviewed applications of different SSL methods in medical imaging, which has achieved significant success when performing on certain medical imaging tasks [103]. After self-supervised training, clinical downstream tasks require fewer labeled data, which is particularly valuable for rare diseases, where building large datasets is inherently more challenging [104]. Among the reviewed applications, the majority of SSL methods are directly adapted from natural image or video analysis, without accounting for the unique characteristics of medical images. For example, in fetal US, the extensive presence of speckle noise and the property that US frames typically share a global pattern with local variations can significantly influence the quality of learned representations and downstream clinical task performance. This highlights a key area for further investigation to fully leverage the potential of SSL in medical imaging.

Despite its effectiveness, SSL in medical imaging still faces challenges related to quality control, ethics, patient confidentiality, and reimbursement [105]. Therefore, trustworthy learning becomes a crucial topic in this safety-critical domain and will be discussed in more detail in the following section. For instance, "explainable AI" [106] is increasingly important in clinical settings, where the black-box nature of deep learning models, which may be acceptable in other fields, is insufficient for clinical implementation. In the case of fetal US, the diagnostic decisions made by the model can directly impact both the fetus and the mother's life. As a result, patients and clinicians are highly motivated to understand how the model works in order to build trust in its use.

Currently, multi-modal self-supervised learning has gained increasing attention in medical imaging analysis, based on the assumption that different types of data can contain distinct but diagnostically valuable information about the patient [53]. Utilising multi-modal data makes these approaches more relevant in clinical settings, as in real-world practice, doctors rely on a combination of expensive medical tests to make accurate diagnoses [107].

2.3 Trustworthy Deep Learning

This section is split into five subsections. Subsection 2.3.1 introduces conceptual foundations of trust and trustworthiness and the key criterion that needed to be examined. In the following sections—2.3.2, 2.3.3, and 2.3.4—we address each dimension of trustworthiness: robustness, explainability, and reliability, respectively. For each, I discuss its definition, significance in the clinical context, and review relevant state-of-the-art methodologies.

2.3.1 Conceptual Foundations of Trust and Trustworthiness

Trust plays a critical role in contexts involving uncertainty, risk, or dependence on others. Despite its importance, the concept of trust is broad and context-dependent. A comparative analysis of major English dictionaries—Webster's (9 definitions), Random House (24), and Oxford (18)—reveals an average of 17 meanings, highlighting its definitional ambiguity [108]. Different disciplines, from philosophy and psychology to systems engineering and AI ethics, offer divergent but complementary interpretations of trust, each emphasising different aspects such as belief, reliability, or expectation [109].

In the context of AI and DL, trustworthiness has a more practical and evaluative focus. For AI-assisted clinical decision systems, it is essential to clarify who or what is actually being "trusted" [110]. O'Neill's framework addresses this by situating trust within relationships between agents—individuals or institutions capable of making commitments and acting upon them [111]. As Baier notes, objects like a chair or timetable cannot truly be trusted or distrusted; rather, I trust the people responsible for their design and implementation [112]. Similarly, in AI, rather than asking whether users should "trust AI," the more important question is whether AI systems—and those behind them—are worthy of trust [113]. O'Neill argues that trust is only valuable when directed at agents or things that are worthy of it: that is, those that are 'trustworthy', misplaced trust can lead to harm, particularly in safety critical applications such as healthcare and justice [111].

Public discourse often treats trust as a generic attitude, such as polling data or media report about "trust in science" or "trust in AI" [114]. However, such generalisations can be superficial, ignoring the normative dimensions that distinguish trust from mere belief or hope. In practice, people tend to trust based on past experience, perceived reliability, and consistent behaviour. For instance, patients may trust clinicians who have demonstrated competence and ethical conduct, but not those with a history of negligence or misinformation.

To better understand trust, O'Neill proposes a tripartite framework consisting of three interrelated dimensions, which reflect different "directions of fit" between belief, action, and reality [111]:

1. **Trust in truth claims:** belief that an agent provides accurate, honest information.

2. **Trust in commitments:** believing that the agent will follow through on what they have promised or declared.

3. **Trust in competence:** believing that the agent has the necessary skills and knowledge to fulfil those commitments.

These dimensions reflect both empirical and normative expectations: trust in truth claims corresponds to an empirical "fit" with reality, while trust in commitments and competence reflect normative obligations. O'Neill argued that one must be trustworthy both in empirical truth claims and in normative action [111]. For example, a clinician who is honest and punctual but frequently misdiagnoses patients lacks competence, while one who is skilled but violates confidentiality fails in commitment. This tripartite model offers a conceptual foundation for evaluating trustworthiness in AI systems. Although DL models are not moral agents, their deployment in critical settings requires a careful evaluation of the systems and the people behind them [115]. As such, the goal should not simply be to promote "trust in AI," but to ensure that AI systems are demonstrably trustworthy—competent, reliable, transparent, and accountable.

In this thesis, I adopt O'Neill's framework as a conceptual foundation for exploring trustworthiness in deep learning, particularly in the domain of medical image analysis in fetal US. I argue that trustworthy DL systems must satisfy a set of technical and ethical criteria, which I explore through three key dimensions in the following sections:

- **Robustness (Section 2.3.2):** the model's ability to maintain accuracy and stability across varied input conditions and imaging settings. This reflects trust in competence, as it ensures consistent performance under realistic conditions.
- **Explainability (Section 2.3.3):** the degree to which a model's decisions can be interpreted and understood. This supports trust in truth claims, particularly in clinical contexts where transparent reasoning could support safe decision-making.
- **Reliability (Section 2.3.4):** the model's consistency in behaviour, especially its ability to distinguish between clinically relevant and irrelevant inputs. This aligns with trust in commitments and competence, ensuring the model acts within its intended scope.

Together, these dimensions form a comprehensive lens for assessing trustworthy AI systems. By grounding technical implementation in a clear conceptual model of trustworthiness, this work aims to support the development of deep learning systems that clinicians can confidently and justifiably rely on in practice.

2.3.2 Robustness of Deep Learning Models

Definition

Merriam-Webster defines robustness as *the quality or state of being robust,* with robust meaning *having or exhibiting strength or vigorous health* [116]. The Oxford English Dictionary describes robustness as *the quality of being robust; strength;*

sturdiness. In statistics, robustness is defined as *the insensitivity of the conclusions of an analysis to violations of the assumptions* [117].

In the context of AI and DL, robustness had been widely used and defined as following. Ben et al. defines model robustness as *the capacity of a model to sustain stable predictive performance in the face of variations and changes in the input data* [118]. Freiesleben et al. describes it as *the relative stability of a robustness target with respect to specific interventions on a modifier* [119]. Furthermore, Zheng et al. defines adversarial robustness as *ability of models to maintain their performance under potential adversarial attacks and perturbations* [120].

In my proposed research, I address the challenge of robustness in DL model by introducing a feature-level local self-supervised learning framework. This approach is designed to enhance the model's capacity to learn invariant and stable feature representations from unlabelled data, therefore improving its resilience to variations and noises commonly encountered in medical imaging. Specifically, I formulate pretext tasks at the feature level to mitigate the influence of pervasive speckle noise in fetal US, which often degrades image quality and hinders representation learning. Furthermore, by emphasising local representation learning, the model is encouraged to focus on strong local variations—such as anatomical boundaries or region-specific structures—rather than global patterns that are often redundant or visually similar across fetal ultrasound frames. This design aims to improve the quality of learned features and support consistent model performance across diverse imaging scenarios, including cases with noise, or limited global context. In doing so, my method directly aligns with established definitions of robustness, ensuring stable and generalisable behaviour when applied to real-world medical imaging inputs—which are often affected by practical challenges such as noise, artefacts, and acquisition variability.

Significance

Robustness is playing an critical role in DL, referring to a model's ability to maintain consistent performance in the presence of variations, noise, or adversarial perturbations in input data. In practice, input data often exhibit inconsistencies due

to factors such as sensor noise, environmental conditions, or variation in user inputs. Real-world data rarely resemble the ideal conditions, inputs may be corrupted, incomplete, poorly captured. In such scenarios, a lack of robustness can result in substantial performance degradation, or even harmful errors—particularly in safety-critical domains such as healthcare and autonomous driving.

Robustness is closely linked to the safety and trustworthiness of AI systems. A robust DL model should be capable of managing such variability effectively [121]. However, DL models are known to be vulnerable to adversarial examples—inputs that have been deliberately perturbed to mislead the model making incorrect predictions. Improving robustness strengthens a model's resistance to such manipulations. Beyond adversarial attacks, even benign sources of variation, such as imaging artefacts, lighting changes, or background noise, can compromise the performance of non-robust models.

In healthcare, robustness is especially critical due to the wide variability of clinical data. A robust model must handle this variability while continuing to produce clinically reliable outputs. As noted by [122], even small perturbations—such as increased noise, minor artefacts, or shifts in data distribution—can lead to significant performance drops. In clinical settings, such issues can result in misclassification, missed pathologies, or erroneous measurements, potentially affecting patient safety. Furthermore, [123] argue that medical imaging is particularly susceptible to adversarial attacks, and that substantial incentives may exist for bad actors to carry on these vulnerabilities. In complex healthcare systems, these threats may be difficult to detect or mitigate, further highlighting the importance of robustness in AI-based diagnostics.

In my research, which focuses on fetal US—a modality characterised by high variability—robustness is especially important. US images are affected by speckle noise, acoustic shadowing, and operator-dependent variation in probe positioning. Additionally, differences across ultrasound machines or clinical sites may produce images with distinct characteristics. A model that lacks robustness to these factors may

misidentify anatomical structures, overlook abnormalities, or produce inconsistent outputs, ultimately undermining clinical decision-making and patient trust.

In summary, robustness is foundational to the safe, effective, and trustworthy deployment of deep learning systems in real-world clinical environments. It ensures that models do not simply rely on memorising training data but instead learn stable, generalisable representations that remain robust under imperfect and dynamic conditions. As deep learning continues to transition from research to real-world applications like medical imaging analysis, robustness must be prioritised as a core design objective.

Methodology

Enhancing model robustness has become an increasingly important objective in the development of DL systems. In the context of medical imaging, robustness refers to a model's ability to maintain consistent performance despite natural variations in input data or the presence of noise and perturbations. Among various strategies, contrastive learning has emerged as one of the most effective methodologies for improving robustness, particularly in SSL frameworks. Contrastive learning enhance a model's ability to handle input variability by learning discriminative feature representations.

As discussed previously, contrastive learning is a self-supervised approach in which models are trained to distinguish between similar and dissimilar instances. This is typically achieved by bringing the representations of similar instances closer in the embedding space while pushing apart the representations of dissimilar ones. The similar instances (positive pairs) are usually constructed through different augmentations of the same input, while dissimilar instances (negative pairs) are drawn from other samples in the mini-batch. Through this process, the model learns to extract invariant features that are resilient to changes in appearance, viewpoint, and noise, thereby improving robustness.

For example, Fu et al. [124] proposed Anatomy-Aware Contrastive Representation Learning (AWCL), which integrates anatomical information into the positive

and negative pair construction process. By embedding clinical prior knowledge into contrastive sampling, AWCL enables the model to focus on clinically meaningful differences, thus enhancing robustness in medical image analysis. Ghosh et al. [125] demonstrated that contrastive learning can significantly improve robustness under label noise. Their work showed that supervised models initialised with representations learned through contrastive learning performed substantially better, even when trained on datasets with incorrect or noisy labels. Chuang et al. [126] introduced Robust InfoNCE (RINCE), a contrastive loss function designed to enhance robustness against noisy views. RINCE dynamically adjusts the importance of training samples based on their estimated noise level, allowing the model to prioritise more reliable inputs during training and mitigate the influence of corrupted or irrelevant data. This highlights contrastive learning's potential to provide noise-invariant features.

In addition to contrastive learning, adversarial training has also been widely studied as a methodology for improving robustness. This technique involves deliberately exposing the model to adversarial examples—inputs that have been intentionally perturbed to mislead the model during training. By learning from these examples, the model becomes more capable of recognising and resisting such manipulations, therefore improving its robustness to both natural and adversarial perturbations. While adversarial training has shown significant promise in various domains, it is not a primary focus of this thesis. A comprehensive review of adversarial robustness methods could be found at [127].

In my research, I design a contrastive learning framework tailored specifically to the characteristics of fetal US data to enhance model robustness. Given the presence of speckle noise in US and the frequent appearance of similar global background patterns across frames, I incorporate two key methodological adaptations. First, I introduce feature-level pretext tasks to discard irrelevant low-level noise and focus on more stable mid-to-high-level representations. Second, I emphasise local representation learning to capture fine-grained anatomical differences rather than overfitting to global similarities. These design choices allow the model to learn

invariant and discriminative features that are resilient to input noise, poor image quality, and local variations common in real-world ultrasound data. This approach illustrates that achieving robustness in medical imaging requires not only a strong learning objective, but also careful alignment between model design and the specific properties of the dataset and clinical task.

2.3.3 Explainability of Deep Learning Models

Definition

Merriam-Webster dictionary defines explain, interpret as *to make something clear or understandable* [116]. Josephson [128] defines an explanation as *an assignment of causal responsibility.* Lewis [129] defines that *explain an event is to provide some information about its causal history. In an act of explaining, someone who is in possession of some information about the causal history of some event — explanatory information,I shall call it — tries to convey it to someone else.*

In the context of AI and DL, explainability has been clearly defined in several foundational works. ISO/IEC [130] defines explainability as *level of understanding how the underlying (AI) technology works* and interpretability as *level of understanding how the AI-based system came up with a given result.* Ali et al. [131] describes explainability as *providing insight into the DNN's decision to the end-user in order to build trust that the AI is making correct and non-biased decisions based on facts* and interpretability as *enabling developers to delve into the model's decision-making process, boosting their confidence in understanding where the model gets its results.* Das et al. [132] describes explanations as *extra metadata information from the model that offers insight into a specific AI decision or the internal functionality of the AI model as a whole.*

In my proposed research, which will be further discussed in Chapter 4, I define explainability in the clinical context as the model's ability to capture anatomy-aware knowledge. I incorporate clinician-driven insights—specifically, sonographers' gaze information—into the explanation process and quantify explainability by measuring the alignment between the model's learned features and the regions where expert

sonographers focus their attention (e.g., through gaze data). In this setting, expert knowledge, such as gaze patterns and anatomical cues, guides the explanation. The model is thus encouraged to "think" more like a sonographer, and its decisions can be explained by referencing the same anatomical features that a human expert would naturally focus and rely on during interpretation.

Significance

With the development of DL, the autonomous machine and the black-box algorithm started to make decision in place of humans with domain knowledge. The black-box algorithm leads to a lack of transparency or opaqueness [133]. This means that the users, and regulators are blinded to the inner logic and inner workings, which becomes a serious disadvantage due to the inability of humans to verify, interpret and understand the reasoning of the system machine [134]. This problem becomes even more urgent as DL models start to be applied to some critical areas, such as the criminal justice system and medical imaging analysis [135, 136]. Automatic decisions generated by DL models can have profound effects in these fields, potentially impacting people's lives. As a result, both users and regulators demand an understanding of how these systems work, how they might be misused, and how they could lead to false or harmful outcomes. As a result, model explainability has garnered significant attention from the public, industry, and scientific communities, with the goal of increasing trust in assistive and fully automated decisions [137].

While explainability is increasingly prioritised in DL research and practice, it is important to acknowledge that not all DL applications require explanations. As noted by Carvalho et al. [138], in scenarios where the cost of incorrect predictions is low or where systems have already been thoroughly validated, high predictive performance may be sufficient. Doshi-Velez and Kim state two kinds of such situation [139]: (1) when there are no significant consequences for incorrect results, and (2) when the problem is significantly well-studied and validated in real applications that the system's decision could be trusted, even if the system is not perfect. Examples of former situations could be advertisement servers and product recommend system,

while the latter situations could be referred to the postal code sorting and aircraft collision systems [140]. Otherwise, explainability could provide its added value

The history of explainability study could be dated back to the 1970s. Researchers started to work with attention on expert system [141–143], and a decade later, to neural network [144], then to recommendation system in another decade [145, 146]. Lent et al. introduced Explainable Artificial Intelligence (XAI) to describe the ability to explain the behaviour of AI in simulation game applications [147]. XAI has since emerged as a growing field, aiming to develop methods for model explanation or create more explainable models, while maintaining task performance [148].

The progress of XAI has slowed down for a decade, until the recent achievements of DL call out another wave towards explainability study [149], due to the mistrust on the DL. In 2016, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a criminal risk assessment tool, has found to be unreliable and racially biased [150]. In addition, deep neural networks (DNNs) have been shown to be highly vulnerable to adversarial perturbations, in which imperceptible modifications to input data that lead to incorrect classifications [151, 152]. Even as defences are proposed [153–155], new attack methods emerge, highlighting the fragility of these systems and the urgent need for transparent decision processes. In parallel, legal frameworks such as the EU General Data Protection Regulation (GDPR) have amplified the call for explainability by granting individuals the right to obtain meaningful information about algorithmic decisions that affect them [156].

In healthcare, the need for explainability is particularly acute. It is widely argued that XAI can foster trust among clinicians, enhance transparency, and reduce the risk of harmful or biased outcomes. However, recent work has cautioned against overly optimistic interpretations of explainability in clinical settings. As highlighted by Ghassemi et al. [157], explainability tools may create false confidence in DL systems—especially in patient-level decision support—by masking model limitations under superficially plausible outputs.

One key limitation is that most explainability approaches rely on humans to interpret what a given explanation actually means. For instance, in medical

imaging analysis, post hoc visual explanations such as heatmaps are among the most widely used techniques. These maps highlight the regions that contributed most to the model's decision. However, this approach can be problematic. Heatmaps often provide coarse information that is difficult to interpret—especially in clinical settings. Even the most salient region in the heatmap may include both relevant and irrelevant anatomical features, and simply localising a region does not clarify what specific information the model used from that area. Moreover, humans have a natural tendency to interpret these explanations in a favourable light, potentially attributing meaningful reasoning to features that may not be clinically relevant. As Rudin noted [158], "You could have many explanations for what a complex model is doing. Do you just pick the one you 'want' to be correct?".

Furthermore, explanations are not always guaranteed to be reliable, and their performance is rarely evaluated from a human-centred perspective. For example, the study in [159] on human-AI collaboration in chest X-ray analysis showed that text-based explanations led to significant over-reliance on the AI system. This effect was mitigated when text-based explanations were combined with saliency maps, suggesting that explanation quality—and its alignment with the model's actual correctness—can significantly affect user trust and performance. These observations underscore the importance of carefully selecting explainability methods and maintaining a critical perspective when applying them in clinical practice. Explainability should not be pursued for its own sake, but rather developed and evaluated with a clear focus on clinical utility, factual accuracy, and alignment with expert reasoning.

Methodology

Here,I review several state-of-the-art methodologies for DL explainability that were available at the time of this research, and which may be applicable to or inform my proposed approach. During that period, most explainability methods focused on visualisation techniques that highlight image regions contributing to a model's prediction. While these methods offer intuitive insights into model behaviour, they

are often coarse and difficult to interpret—particularly in medical imaging, where specialised domain knowledge is required to understand the content. As a result, such visual explanations are not always easily comparable or accessible to lay audiences.

In response to these limitations, several quantitative approaches have been proposed to provide more objective and comparable metrics for explainability. However, in the context of medical image analysis—where models are trained to mimic expert clinical reasoning—explainability methods must go beyond general-purpose techniques. Specifically, there is a need to integrate clinical domain knowledge to ensure that explanations are both relevant and meaningful to healthcare professionals. This highlights the importance of my work on developing explainability frameworks that are grounded in the clinical context, aligning with experts' gaze information.

- **Qualitative Visualization** CNN visualization includes visualizing filters, activation [160, 161], hidden neurons [162, 163] and CNN decision [163, 164]. Here I focus on techniques that are proposed to visualize a CNN decision, that highlight the "important" regions for decision-making. Techniques mainly fall into one of three categories [165]: perturbation-based, propagation-based methods, and attention-based methods.

  - **Permutation-based methods:** One of the earliest contributions to CNN interpretation was made by Zeiler and Fergus [163], who blocked random patches of images with grey patches and observed how the classifier degrades when presented with these examples. A heat map was then generated where each pixel represents the class score change if the occlusion is in place.

  - **Propagation-based methods:** Simonyan [164] proposed to produce a saliency map by computing partial derivatives of the predicted class score with respect to input pixel intensities. To reduce noise in the saliency map, Springenberg [166] extend [164] to guided-backpropagation, where a modified backpropagation step is introduced to get better visualization. These works are compared in [167], though the visualizations produced are

of high resolution, they are not class-discriminative. Direct comparison and analysis are also hard to conduct based on visualization solely. Other subsequent methods, for instance, Layerwise Relevance propagation [168] and integrated gradients [169] are further examples of this class of method.

  - **Attention-based methods:** Differing from propagation-based methods, attention-based methods highlight the map by resorting the activation of feature maps. One of the important branch is the Class Activation Mapping(CAM) methods [170–173]. The original CAM method is proposed by Zhou et al. [174]. It linearly combined the feature maps at the penultimate layer, where the weights of each feature map are determined by the corresponding class weights from the last fully-connected layer. However, the method is restricted as it only applies models with global average pooling (GAP) layer. Grad-CAM is then proposed by Selvaraju et al. [173], it weighted each feature map using gradients instead. Aditya et al. [170] further developed Grad-CAM++ via introducing higher-order derivatives from Grad-CAM and getting a closed-form solution for the weights of each feature map. The followed Smooth Grad-CAM++ [172] method inherited the framework of Grad-CAM++, while combing SmoothGrad [175] methods in gradient calculation. Fu et al. proposed Axiom-based Grad-CAM (XGrad-CAM) to satisfy Sensitivity and Conservation axioms for clear and sufficient theoretical support [165]. More recently, Desai et al. [171] proposed Ablation-CAM to remove the dependence on gradients by running forward propagation, while stays time consuming.

- **Semantic Interpretability** Kim et al. introduced the "concept activation vector" (CAV) that automatically extracts visual concepts that are user-defined, human meaningful and important for the CNN prediction [176]. Testing with CAV (TCAV) quantitatively analyse the relative importance of a user-provided concept, which could be easily used and understood by non-experts and turn the high-dimensional representations into low-dimensional

interpretations. TCAV pioneered dataset-level analysis, however, not supported for image-level concepts. Pfau et al. proposed Robust CAV (RCAV) to quantify concept sensitivity on individual model predictions and on model behaviour as a whole [177].

In addition to methods built on TCAV, Adel et al. apply normalizing flows to semantic interpretability, with the cost of a large number of inputs and high computational burden [178]. Other approaches include model modification, for instance, using random forest to replace certain convolutional layers [179, 180], but not widespread yet.

2.3.4 Reliability of Deep Learning Models

Definition

Merriam-Webster defines reliability as *the quality or state of being reliable* and *the extent to which an experiment, test, or measuring procedure yields the same results on repeated trials* [116]. The Cambridge Dictionary describes it as *the quality of being able to be trusted or believed because of working or behaving well* [181]. The Oxford English Dictionary defines it as *the quality of being reliable; the extent to which a person or thing may be relied upon; trustworthiness, dependability* [182].

In context of AI and DL, the National Institute of Standards and Technology (NIST) refers AI system reliability as *the ability of an item to perform as required, without failure, for a given time interval, under given conditions* [183]. ISO/IEC [130] defines reliability as *degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.*

In my proposed research, I address the challenge of reliability in AI systems by introducing the concept of out-of-clinical-distribution (OCD) detection, tailored specifically for US video analysis. I define reliability not only as the model's ability to make accurate predictions on clinically relevant data but also as its capacity to recognise and reject non-diagnostic frames—those lacking anatomical content essential for clinical interpretation. By detecting and filtering out these background frames during both training and inference, my method ensures that the model

focuses on semantically rich, clinically meaningful inputs. This targeted approach to reliability improves representation learning, reduces noise in decision-making, and supports safer and more trustworthy deployment of deep learning models in clinical practice.

Significance

Reliability is a fundamental property for any system intended for real-world deployment. In the context of DL, reliability refers to a model's ability to produce consistent, dependable, and accurate outputs across a range of conditions, including under uncertainty and data variability. A reliable model should not only perform well on average but also maintain stable behaviour when exposed to diverse inputs, shifting data distributions, or unfamiliar deployment environments [184]. It should avoid erratic or unpredictable outputs when encountering unfamiliar or out-of-distribution (OOD) data. Reliability becomes particularly important in safety-critical domains, where incorrect predictions can have serious consequences. In healthcare, for example, an unreliable model could lead to missed diagnoses, inappropriate treatments, or misleading guidance for clinicians—potentially influence patient safety [185].

A key aspect of reliability is a model's ability to recognise the limits of its competence. This involves identifying when an input lies outside the distribution of data seen during training—a task known as OOD detection [186]. Most deep learning models are trained under closed-world assumptions, where the training and test distributions are assumed to match [187]. However, real-world applications often violate this assumption, as models are frequently exposed to data from distributions that differ from those seen during training [188]. Without the ability to detect OOD inputs, models may produce incorrect and overconfident predictions when encountering unseen cases [189]. OOD detection is thus essential to ensuring the safety and reliability of DL systems [190].

Without OOD detection, a DL model risks extrapolating into unknown regions of the input space, potentially generating confidently incorrect predictions. This

risk is further magnified if the model fails silently—producing erroneous outputs with no signal of uncertainty. In medical imaging, this could occur when, for instance, a model trained on chest X-rays is tested on breast X-rays, or when rare anatomical variations or imaging artefacts are present. A reliable system should instead flag such cases as uncertain or defer to human judgment, rather than make unsupported predictions.

Furthermore, reliability is essential for trust, accountability, and regulatory compliance in AI deployment. For the regulators, models that fail in unpredictable or unexplainable ways are unlikely to meet clinical safety and performance standards [191]. For the end-users, clinicians are less likely to use AI tools that show inconsistent or unreliable behaviour [192]. Ensuring reliability, therefore, is not only essential for safe model operation but also crucial for increasing the acceptance and integration of real-world AI-driven clinical applications.

Methodology

Ensuring model reliability has become a key area of research, particularly with OOD detection and uncertainty quantification. In this section,I review the main state-of-the-art methods for OOD detection, which serve as the foundation for the reliability assessment in my subsequent research. These methods can be broadly categorised into four groups: classification-based, density-based, distance-based, and reconstruction-based approaches.

While each category has demonstrated effectiveness for specific tasks, they also come with notable limitations. For example, [193] highlights the challenge faced by classification-based methods in balancing the dual objectives of accurate classification and effective OOD detection within a shared network. Furthermore, different methods tend to perform better under different OOD settings—such as far-OOD versus near-OOD—indicating that no single approach is universally optimal.

Given these considerations, my research aims to develop a simple yet effective method for reliable OOD detection, tailored to the needs of my specific clinical application (namely OCD detection).

- **Classification-Based Methods**

  The majority of early works derive OOD scores from the network's output, while more recent approaches utilize information from the feature space or gradient space, and some even incorporate foundation models as additional tools. In this section, I focus on the literature related to output-based methods and gradient-based methods for OOD detection.

  - **Output-based Methods**

    * **Post-Hoc Detection** Post-hoc detection methods do not requires modification of training procedures and objectives, which is easier to use and apply in real-world environments. Hendrycks et al. [194] presented a very early work on OOD detection, which provided a simple baseline for OOD detection that use the maximum softmax probability as the ODD score. ID samples that are correctly classified are tend to have higher maximum softmax probability than those OOD samples. Liang et al. [195] proposed a method named ODIN (OOD detector for neural networks), which builds on the baseline method to detect OOD samples in neural networks. ODIN is based on two key components: temperature scaling and input perturbation. These techniques are used to amplify the gap between ID and OOD samples since the authors observed that applying temperature scaling in the softmax function and adding small perturbations to the inputs increased the difference in softmax scores between ID and OOD samples. Therefore, combination of these techniques leads to improved detection performance.

      Sun et al. [196] identified that a key challenge in existing OOD detection methods is their tendency to be overconfident in OOD predictions. To address this, they proposed ReAct (Rectified Activations), which is based on the observation that OOD samples often trigger internal activation patterns in neural networks that

are significantly different from those of ID samples. By rectifying higher activations, ReAct improves detection performance, reducing overconfidence on OOD predictions. Dong et al. [197] proposed a novel OOD metric called NMD (Neural Mean Discrepancy), which measures the discrepancy between the mean of the input sample and the mean of the training data. This metric can be computed directly from model activations without requiring a backward pass and is independent of the batch normalization layer. Sun et al. [198] proposed a sparsification-based OOD detection framework called DICE (Directed Sparsification), which is based on the observation that predictions for ID classes depend only on a subset of the neural network's weights. DICE selects the most relevant or contributing weights for OOD detection by ranking them according to their contribution. This weight sparsification technique results in better separation of ID and OOD samples, leading to improved OOD detection performance.

* **Training-based Detection**
  - **Confidence Estimation** DeVries et al. [199] trained a neural network to output confidence score estimations for each input, which are then used to distinguish between ID and OOD samples. This confidence estimation is produced by a separate confidence estimation branch, added to the feedforward architecture alongs with the original class prediction branch. Such branch is usually added after the penultimate layer, ensuring both confidence branch and prediction branch have the same information for their respective tasks of confidence estimation and class prediction. Wang et al. [200] introduced EOW-Softmax (Energy-based Open-World Softmax), a novel $K+1$-way softmax formulation that incorporates open-world uncertainty as an additional output dimension. The neural network generates $K+1$ logits, where

the first K dimensions represent the prediction scores for the classification task, and the $K + 1$-th dimension encodes the model's open-world uncertainty. The authors also proposed a new energy-based objective function that seamlessly integrates classification and uncertainty modelling.

- **Adversarial Training** Bitterwolf et al. [201] introduced a training scheme called GOOD (Guaranteed Out-Of-Distribution Detection) for certifiable OOD detection. The goal of this method is to provide worst-case guarantees by not only enforcing low confidence at OOD points but also within the l_∞-ball surrounding those points. The approach leverages techniques from interval bound propagation (IBP) [202] to compute an upper bound on the maximum confidence within the l_∞-ball. This method is effective even for closely related OOD samples, such as those from EMNIST and MNIST datasets. Chen et al. [203] were the first to address the vulnerability of existing OOD detection methods to adversarial ID samples, and proposed a novel algorithm called ALOE to enhance the robustness of OOD detection. Their approach involves training the model using two types of perturbed adversarial samples: adversarially crafted ID samples and OOD samples. Both the softmax confidence score and Mahalanobis distance-based confidence score are utilized for OOD detection. Recently, chen et al. [204] proposed a novel framework called ATOM (Adversarial Training with Informative Outlier Mining) for robust OOD detection, based on the observation that most auxiliary OOD data may be uninformative or even harmful for OOD detection. To address this, they introduced informative outlier mining to focus on more relevant outliers and enable the model to estimate a tighter decision boundary between ID and

OOD samples. Hence, this approach improves the robustness and accuracy of OOD detection.

- **Stronger Data Augmentation** Thulasidasan et al. [205] incorporated Mixup [206] to improve calibration and uncertainty prediction in deep learning models. During Mixup training, additional samples are generated by convexly combining random pairs of images and their corresponding labels. The resulting Mixup-trained models is less prone to overconfident predictions on OOD samples. Hendrycks et al. [207] proposed a technique for robust OOD detection called AUGMIX. The method employs diverse augmentation generation to induce robustness, utilizes a mixing technique to combine multiple augmented images, and introduces a Jensen-Shannon Divergence consistency loss to enforce smoothness in the neural network's response. This approach is both simple and effective in enhancing the robustness of OOD detection. As strong data augmentation techniques often enhance the robustness of OOD detection but can negatively impact accuracy, Hendrycks et al. [208] proposed PIXMIX to achieve Pareto improvements. The method consists of two main components: a set of structurally complex images ("Pix") and a pipeline for augmenting clean training images ("Mix"). Diverse patterns, including fractals and feature visualisations, are integrated into the training set, with the networks trained for standard classification. This approach balances robustness and accuracy in OOD detection.

– **Gradient-based Methods**

Liang et al. [195] proposed ODIN, one of the first methods to explore incorporating gradient information for OOD detection. ODIN implicitly uses gradients during the input perturbation process, where small perturbations are applied based on the input gradients. This technique

increases the softmax score gap between ID and OOD samples, enhancing the effectiveness of OOD detection. Lee et al. [209] were the first to propose using backpropagated gradients directly as a signal for OOD detection. The obtained gradients can serve as a distance measure in the model's space, indicating how much change is required to properly represent the given input and providing insight into how familiar and certain the model is regarding the input. Huang et al. [210] presented an novel approach, named GradNorm to detect OOD samples by using information of gradient from pre-trained neural network. GradNorm uses the vector norm of the gradient as OOD score, which is backpropagated from the KL divergence between the softmax output and a uniform probability distribution. ID samples are more likely to have a larger KL divergence than OOD samples because the softmax outputs for ID samples tend to be higher for the ground-truth class, leading to a more concentrated distribution, whereas OOD samples are more likely to produce a uniform distribution, making it a good measure to separate ID and OOD samples.

- **Density-based Methods**

  The core of density-based OOD detection methods is to explicitly model the probability distribution of in-distribution data and flag test data that fall into low-density regions—those that are unlikely under the modelled distribution—as OOD data. Recent works include techniques such as parametric density estimation, density estimation with deep generative models and energy-based models.

  - **Classic density estimation** Desforges et al. [211] proposed using probability density function (PDF) estimation through kernel methods, where OOD samples can be detected based on the estimated density. Low-density regions indicate data points that are likely to be OOD. Lee et al. [212] considered the case where the ID samples contains multiple classes

and proposed modelling ID samples using a class-conditional Gaussian distribution. This distribution is derived from features under Gaussian discriminant analysis, and OOD samples are identified by calculating the Mahalanobis distance to the closest class-conditional distribution.

– **Density estimation with deep generative models** Sabokrou et al. [213] proposed an end-to-end architecture for one-class classification, consisting of two deep learning networks. One network functions as the OOD detector, while the other supports it by enhancing inlier samples and distorting outliers. Pidhorskyi et al. [214] leveraged an encoder-decoder network to compute the likelihood that a sample was generated by the inlier distribution. To reduce computational complexity, the parametrised manifold capturing the underlying structure of the inlier distribution is linearised. Deecke et al. [215] used generative adversarial networks (GANs) for OOD detection. The goal is to assess the representation of a given input sample in the latent space, which is trained using ID samples. If a good representation cannot be found for the input, it is classified as OOD. Flow-based generative methods could also for probabilistic modelling. Zisselman et al. [216] proposed to leveraged density modelling based on deep normalising flow, for modelling the distribution of the feature space. Residual flow, which is an architecture that learns the residual distribution from a base Gaussian distribution is introduced. This method is effective for modelling distributions that are approximately Gaussian.

– **Energy-based models** Liu et al. [217] proposed using an energy score for OOD detection. In this energy-based model, each input sample is mapped to a single value, where ID samples have lower energy scores and OOD samples have higher scores. Higher energy values indicate a lower likelihood of occurrence. This method is theoretically aligned with the probability density of the input sample and is less prone to the overconfidence issue often seen in traditional models. Lin et

al. [218] proposed a novel framework called MOOD (multi-level OOD detection), which explores the relationship between different network depths and OOD inference. Based on the intrinsic complexity of the input data, MOOD can detect easy OOD samples at earlier layers, preventing unnecessary propagation to deeper layers. Additionally, an adjusted energy score is introduced as the OOD detection metric, which is shown to be more suitable for networks with multiple classifiers.

- **Distance-based Methods**

  The core idea of distance-based OOD detection method is to measure the distance between an input sample and the centriods or prototypes of ID classes. This distance metric is used for OOD detection, as OOD samples are expected to be relatively far away from the ID class representations, making it possible to distinguish them based on their spatial separation in the feature space. Here, I review some recent works categorized based on the distance function they employ for OOD detection.

  – **Cosine similarity** Techapanurak et al. [219] proposed using cosine similarity for OOD detection, where class probabilities are modeled using the softmax of scaled cosine similarity. This score is then used for OOD detection. The proposed method is hyperparameter-free and can be applied to any network by simply replacing the final output layer. Chen et al. [220] leveraged the cosine similarity for boundary based OOD detection to decompose the generalized Zero-Shot Learning problem (GZSL) problem to a conventional Zero-Shot Learning (ZSL) problem. Zaeemzadeh et al. [221] proposed to embed the ID samples onto a union of 1-dimensional sub-spaces that are spanned by the first singular vector (FSV) of features vectors from the same class. An input sample is detected as OOD if the spectral discrepancy between the input and the FSV is too large, which is measured using cosine similarity, which is shown to be more robust and effective.

– **Mahalanobis distance (MD)** Lee et al. [212] assumed that the pre-trained features follow a class-conditional Gaussian distribution and used Mahalanobis distance with respect to the closest class-conditional distribution as the score for OOD detection. This approach firstly introduces the MD to measure how far a sample is from the in-distribution classes. Ren et al. [222] argued that while MD based methods are effective for detecting far OOD samples (those semantically very different from ID samples), they may fail to identify near OOD samples (which are semantically close to ID samples). To address this issue, they proposed a simple modification called Relative Mahalanobis Distance (RMD) for near OOD detection tasks. This approach splits the original image into foreground and background and uses the ratio of Mahalanobis distances between two generative models on these two spaces as the OOD detection score.

– **Euclidean distance** Huang et al. [223] observed that OOD samples with bounded norms tend to cluster near the Feature Space Singularity (FSS) in the feature space. They proposed using the Euclidean distance between the feature of input sample and the FSS, termed Feature Space Singularity Distance (FSSD), as an effective measure for OOD detection.

– **Non-parametric nearest-neighbour distance** Prior distance-based methods impose a strong distribution assumption of the underlying feature space being class-conditional Gaussian, which might not always hold. Sun et al. [224] therefore proposed to leverage the non-parametric nearest neighbour approach for OOD detection, which does not impose any distributional assumption, hence provide stronger flexibility, simplicity and generality.

– **Geodesic distance** Gomes et al. [225] proposed a novel method, named IGEOOD, for OOD detection. This method explores the information-geometric properties of the feature space and uses the Fisher-Rao distance

on the probabilistic manifold of the learned features as an effective score for OOD detection. The Fisher-Rao distance, which represents the geodesic distance between probability distributions in a statistical manifold equipped with the Fisher metric, serves as a powerful tool for clustering and a distance metric in the context of multivariate Gaussian distributions [226].

- **Reconstruction-based Methods**

  The core idea of reconstruction-based OOD detection methods is to compress the input data into a latent space and then attempt to accurately reconstruct the input using the learned compressed representations. This encoder-decoder framework is trained and optimized using only ID data. The underlying assumption is that OOD data cannot be effectively compressed and reconstructed due to its dissimilarity to the ID data. As a result, OOD samples tend to produce significant reconstruction errors. This difference in reconstruction performance can then be used as an indicator for OOD detection. Denouden et al. [227] used a reconstruction-based autoencoder as the encoder-decoder framework to train the ID samples. The models identifies OOD samples it could not be well recovered. Mahalanobis distance is Incorporated in latent space to better capture the OOD samples. Zhou et al. [228] proposed a novel framework of layer-wise semantic reconstruction to improve the performance of autoencoder-based OOD detection. The proposed framework reserved the reconstruction power of the autoencoder, while maximumly compressing the latent space. Yang et al. [229] proposed a framework named MoodCat that masks a random portion of the input image to synthesise the masked image and identifies the OOD samples based on the classification-based reconstruction error. Semantic difference between the original image and synthesised one is used to calculate Reconstruction error. [230] introduced the idea of not reconstructing the original image itself but rather the raw pixel of image. They proposed a novel method called READ (Reconstruction Error Aggregated

Detector), which combines inconsistencies from both the classifier and the autoencoder. In this approach, the reconstruction error of raw pixels is transformed into the latent space of the classifier, allowing for a more effective aggregation of errors for OOD detection.

2.3.5 Conclusion

In this section, I have reviewed the conceptual foundations and practical implications of trustworthiness in DL, focusing on three core dimensions: robustness—whether the model performs well under diverse and imperfect conditions; explainability—whether its decision-making process can be understood; and reliability—whether it behaves consistently and safely in real-world use. Each of these aspects plays a crucial role in the safe and effective deployment of DL models in clinical settings. Robustness ensures performance stability across varying inputs; explainability provides insight into the model's reasoning process; and reliability safeguards against silent failures by ensuring appropriate outputs in the face of uncertainty or unfamiliar data.

These trustworthiness dimensions directly shape the design and direction of my research. To improve robustness, I propose a feature-level, local self-supervised learning strategy tailored to the unique characteristics of fetal US data. To address explainability, I incorporate gaze-based anatomical priors to anchor model attention in clinically meaningful features. For reliability, I introduce an OCD detection framework that enables the model to differentiate between diagnostic and non-diagnostic frames, allowing it to focus on semantically rich, clinically relevant content for meaningful representation learning.

However, several challenges remain. First, balancing the three dimensions is non-trivial, and improvements in one area may compromise another. For example, techniques that improve robustness through heavy data augmentation may reduce the interpretability of model features and hinder explainability. Second, there is a lack of standardised metrics for evaluating these dimensions. Explainability methods are rarely assessed from a human-centred perspective, and OOD detection

often relies on fixed thresholds that may not generalise across tasks or datasets. Third, clinical integration introduces additional complexity: explainability must serve diverse end-users (e.g., clinicians, regulators), domain-specific knowledge must be incorporated, and model behaviour must adapt to the unique characteristics of medical data and workflows. In my work, I address these challenges by integrating gaze-tracking data into the explanation process, accounting for US-specific imaging properties, and redefining OOD detection as OCD detection to reflect clinical relevance. These efforts aim to promote clinically meaningful, interpretable, and trustworthy DL systems suitable for real-world deployment in healthcare.

As constant as the moon in its course, as the sun in its rising. Like the longevity of the southern mountains, never failing, never falling.

— *The Book of Songs, Major Odes of the Kingdom, King Wen.*

3 Data

Contents

3.1 Introduction

This chapter provides a brief description of the dataset used in this thesis and outlines the preparation steps undertaken for both self-supervised and supervised deep learning models in our study.

The National Health Service (NHS) has standardized prenatal care in the UK through the Fetal Anomaly Screening Program (FASP) [231] since 2001. A key component of this program is the 20-week screening scan, also known as the second

trimester anomaly scan, which examines 11 physical conditions of the fetus. This scan plays a crucial role in prenatal care, offering pregnant women the opportunity to benefit from treatment before or after birth or leading to a discussion about options, including continuing or terminating the pregnancy. Therefore, is the main focus of my reserach.

In this thesis, I explore and analyse second-trimester ultrasound scans using a self-supervised approach, where raw video recordings are leveraged to learn anatomical representations for clinical downstream tasks. Self-supervised learning alleviates the burden of manual annotation of ultrasound videos, which is time-consuming, labour-intensive, and nearly impossible for entire video scans. Though previous studies that rely on human annotation in a supervised manner have achieved promising results, only a small fraction of the available information has been utilised, which is merely "the tip of the iceberg" compared to the entire video scan due to limited annotations available. By employing a self-supervised approach, I would be able to explore the dataset from a different angle, accessing the substantial portion "beneath the surface", which is the unlabelled scanning video, and making better use of the entire scanning dataset for feature extraction and representation learning.

All ultrasound video scans used in this thesis were acquired as part of the PULSE (Perception Ultrasound by Learning Sonographer Experience) project, which is detailed in Section 3.2. Several datasets, both labelled and unlabelled, are used in this thesis for the subsequent analysis chapters. Section 3.3 and 3.4 provide a detail description of each dataset used in my studies.

3.2 Overview of PULSE Data

3.2.1 Motivations

Despite US being widely used due to its low-cost, portability and safety, it requires a high degree of expertise to conduct the scan effectively [232–234]. In the United Kingdom, the specialised healthcare professional, called a sonographer, is responsible for not only performing diagnostic US, but also interpreting the real-time US movements and issuing diagnostic reports [235]. Hence, a sonographer needs to be

highly educated and well trained. In addition, the motions of a handheld probe and the capture of visual information from the screen are often prone to human bias and hard to standardise. There is no universal standard and skill assessment for sonographers [236]. This operator-dependency restricts the usage of US to the setting of hospitals and this has largely held back the wider development of US. According to WHO, world-widely most of the end-users of US equipment are likely to be individuals with little or no formal training [237, 238].

Therefore, to improve sonographers' accuracy, efficiency, and productivity, as well as to reduce the reliance on highly trained ultrasound operators, the use of artificial intelligence-based technologies has received considerable attention [239–245].

The PULSE project is an interdisciplinary research project that aims to understand the entire obstetric imaging scanning process as a data science problem [5]. It works towards "*combining machine learning, ultrasound video image analysis and human-computer interaction (HCI) research in an attempt to better understand how to build easier to use ultrasound systems of the future [246].*", with the ultimate goal of automatic interpretation of freehand ultrasound. Multi-modal HCI data has been collected, covering sonographer eye-tracking data, probe motion data and sonographer speech recording to enhance understanding of human visual search and navigation.

3.2.2 Data Acquisition

The PULSE study includes anonymized full-length obstetric ultrasound scan videos while tracking the actions of the sonographer, including transducer motion, eye-gaze attention and audio recording. In this section, I describe the data acquisition process of the PULSE study, based on the scientific report written by Lior [5], a maternal fetal medicine ultrasound expert and a senior collaborator of our lab.

Routine Care Setting

As part of the standard prenatal care, ultrasound scanning were carried out at the maternity ultrasound unit, Oxford University Hospitals NHS Foundation Trust,

Oxfordshire, United Kingdom. Three routine ultrasound scans are offered, including a first-trimester dating scan at 11–13+6 weeks to estimate the viability and gestational age, a 20-week screening scan, also known as the second-trimester anomaly scan to identify fetal anomalies and measure fetal growth, and a growth scan in the third trimester (around 36 weeks).

For quality assurance purposes, all ultrasound examinations are conducted by accredited sonographers, supervised trainees, or fetal medicine doctors using standard ultrasound equipment. In this thesis, I refer to all of them as "sonographers". The stored images and measurements are regularly reviewed by a senior accredited sonographer, following established quality criteria [247].

In addition, if any abnormality is identified or suspected during the scanning, all data from this scan would be excluded from the study and the pregnant women would be referred for further examination at the Fetal Medicine Unit, Oxford University Hospitals NHS Foundation Trust.

Participants

This project involves two groups of participants: pregnant women and sonographers.

Any pregnant woman aged 18 or older who can provide verbal and written informed consent in English is invited to participate in this study during routine obstetric scans by agreeing to have their ultrasound scan recorded. No additional follow-up research scans are required. Data from women who choose to withdraw after giving consent will be excluded from the study.

Participating sonographers include accredited sonographers, trainees supervised by accredited sonographers, or fetal medicine doctors who perform obstetric ultrasound and are willing and able to provide informed consent in English. Sonographers will receive an introduction to the project and will be asked to perform routine scans without altering their usual practices for the study.

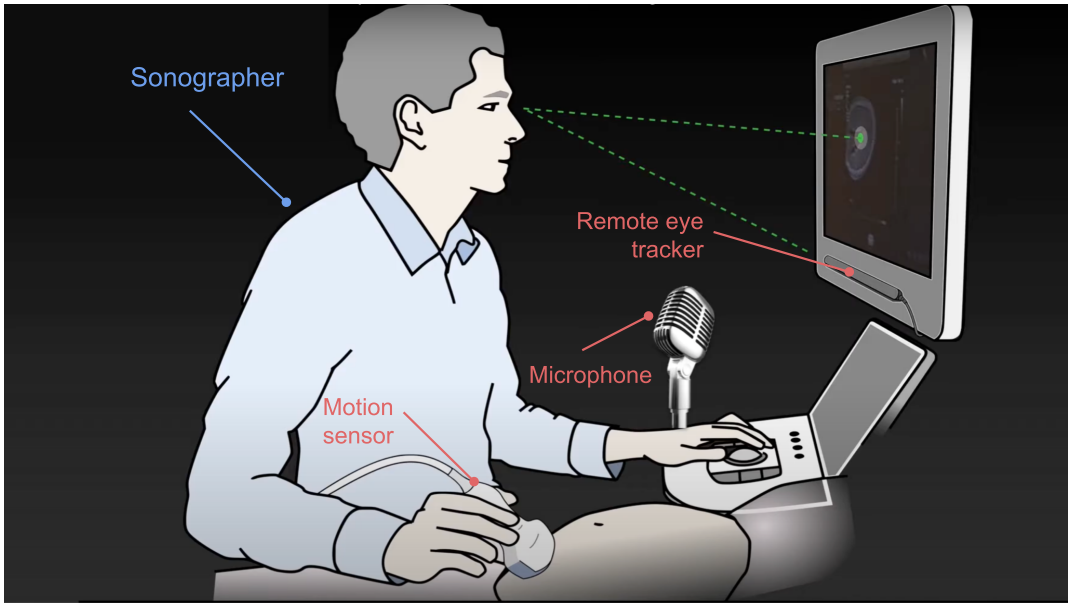


Figure 3.1: Illustration of the setup of the routine scan acquisition process for PULSE study.

Ultrasound System

All US scans are performed using a commercial Voluson E8 version BT18 (General Electric Healthcare, Zipf, Austria) ultrasound machine equipped with standard curvilinear (C2-9-D, C1-6-D, C1-5-D) and 3D/4D transducers (RAB6-D, RC6M). A customised addition was made to create a dedicated multi-modality ultrasound system that simultaneously acquires ultrasound scans and records the sonographer's actions. This system captures full-length ultrasound video scans, eye-tracking data, transducer motion data, and audio of the sonographer's commentary in real-time.

Ethics Approval

This study was approved by the West of Scotland Research Ethics Service, UK Research Ethics Committee (Reference 18/WS/0051). Written informed consent was obtained from all participants in English, and all procedures were conducted in accordance with the guidelines and regulations.

Routine Scan Acquisition

The setup of the routine scan acquisition process for this study is illustrated in Fig. 3.1. During the scan, the sonographer typically sits or stands beside the pregnant woman, who is lying on the examination bed. The sonographer manipulates the transducer and adjusts the machine settings as needed, receiving real-time visual feedback from the display screen. As they continuously move their hands to adjust the transducer's position, the displayed video updates accordingly.

During this process, data from different modalities are record and stored simultaneously:

- **Video recording:** A computer equipped with a video capture card (DVI2PCIe, Epiphany Video, Palo Alto, California) and a built-in software is used to capture real-time US video. Full-length US scans are captured using lossless compression at resolution of 1920×1080 pixels at 30 frames per second. Real-time anonymisation is conducted to ensure that no personal information is included in the recordings.

- **Eye tracking (sonographer's gaze points):** A remote eye-tracker (Tobii Eyetracking Eye Tracker 4C, Danderyd, Sweden) is mounted below the display screen. For each ultrasound frame, the precise point of the sonographer's gaze, indicating their visual attention, is recorded. No visual cues or signals are provided to the sonographers to indicate whether the eye-tracking device is active, ensuring the fairness of the recorded data.

- **Transducer tracking (probe movement):** A motion sensor (NGIMU, X-IO Technologies, Bristol, UK) attached to the US transducer is used for transducer motion tracking, with 3-axis accelerometer data recorded at 100 Hz.

- **Audio recording:** Microphones (PCC160, Crown HARMAN, Northridge, California) are used to record the sonographer's voice. Two microphones are utilised to differentiate the sonographer's voice from background noise and

other people's voices presented in the room. One microphone is placed near the ultrasound display screen to best capture the sonographer's voice, while the other is positioned next to the pregnant woman and the accompanying people, away from the sonographer, to isolate ambient sounds.

In this thesis, a large-scale of raw video recordings were used for self-supervised pre-training, while a smaller set of labelled video frames were employed as the downstream task dataset and the out-of-clinical-distribution detection dataset. Eye-tracking dataset was also utilised for the Saliency-VAM model [248].

Sonographer Workflow

The sonographer moves the probe to explore different views of the fetal US from various angles and zoom levels in real-time, aiming to find the best anatomical view. Once the appropriate structure or view is identified, the sonographer freezes the frame. If the current view is not satisfactory, the sonographer may unfreeze the image to continue scanning, possibly taking multiple freeze frames while adjusting the probe angle, zooming in, or changing angles to capture the optimal view, known as the standard plane. Once the best view is located, the sonographer freezes the image and proceeds with the necessary biometric measurements and interpretation, such as assessing head circumference, femur length, or evaluating the heart and brain of the fetus.

Sample Size

Based on logistical and clinical considerations—ensuring the study could be conducted within manageable time frames and resources while also providing a large enough sample to achieve adequate representation across different sonographers and fetal gestational ages—the PULSE study prospectively decided to recruit a total of 3,000 routine ultrasound scans, with up to 20 sonographers performing these scans, resulting in approximately 1,000 scans for each gestational age group.

Figure 3.2: An example of the original ultrasound video recording.

3.2.3 Basic Preprocessing for Ultrasound Video Frames

Using the aforementioned equipment, Fig. 3.2 shows an example of the original US video recording from the display screen. The screen displays the US image, along with text information and various control buttons. A basic preprocessing step is applied to the stored video frames, where a predefined bounding box is used to crop the clear ultrasound image to a size of 1008×784, excluding any additional information displayed on the screen. The images were then scaled down to 576×448 to reduce storage space on the local machine for subsequent analysis and augmentation. This basic preprocessing was implemented in Python 3.8.

3.2.4 Existing Researches and Future Directions

A large amount of researches have been conducted under the PULSE project. In this section, I summarise a selection of papers that utilise different combination of data modalities and are designed for different clinical tasks. As shown in Table 3.1, most of the researches have focused on the second-trimester dataset due to the optimal size and visibility of the fetus during this stage. Studies on the first-trimester dataset are also gaining momentum, reflecting a growing real-world need.

Authors	data modalities	US scan trimester	Task
Sharma et al. [249]	US video data	2nd	Automatic US video description
Gao et al. [250]	US video data	2nd	Fetal structures detection
Lee et al. [251]	US video data	2nd	Standard planes classification
Cai et al. [252]	US video and gaze data	2nd	Standard plane detection
Cai et al. [253]	US video and gaze data	2nd	Standard biometry plane-finding navigation
Wang et al. [254]	US video and probe motion data	2nd	Operator skill assessment
Alsharid et al. [255]	US video and audio data	2nd	Automatic image captioning
Yasrab et al. [256]	US video data	1st	Standard plane segmentation
Savochkina et al. [257]	US video and gaze data	1st	Video saliency prediction

Table 3.1: Summaries of selected paper published under PULSE project

Most prior researches have relied on supervised learning, which requires a large number of manual annotations for US video datasets to enable effective training. However, only a small number of US frames have been labelled by clinicians during the scanning process, typically focusing on standard planes that require clinical measurements and assessments. Lior, a maternal fetal medicine ultrasound expert and a collaborator of our lab, has also manually annotated a small portion of video clips based on anatomy, organs, and standard planes, covering less than 20% of the dataset until 2021.

While supervised learning approaches have achieved promising results, the labelled data used for learning represents are only the tip of the iceberg compared to the entire dataset, leaving much of the information in the unused data untapped. A significant amount of the unlabelled data could be leveraged to extract additional

Dataset	Trime -ster	Number of US scans	Mean scan length (minutes)	Number of Clips	Number of frames	Length of clips
Dataset I	2nd	719	34.2 ± 12.1	70,661	2,261,179	32
Dataset II	2nd	342	36.7 ± 10.8	119,328	715,968	6

Table 3.2: Summary of the unlabelled US video datasets

anatomical information, which could be beneficial for the model representations learning. Given the vast amount of unlabelled data and its potential richness, a key challenge is how to effectively incorporate these datasets for more efficient representation learning.

Therefore, self-supervised representation learning, designed to extract meaningful representations from raw image or video datasets, makes the PULSE dataset an ideal candidate for such an approach. With its large amount of unlabelled data, the PULSE dataset can greatly benefit from this approach, enhancing ultrasound video analysis by tapping into the untapped information within the dataset.

3.3 Unlabelled US Video Dataset

In this thesis, I rely exclusively on US video as the modality for self-supervised training. Two unlabeled ultrasound video datasets (Dataset I and Dataset II) have been used as self-supervised pre-training datasets for Chapter 4 and Chapter 5, respectively. These unlabeled datasets are used in the pre-training process with a designed pretext task to learn meaningful representations, which can later be transferred to various, more complex clinical downstream tasks. Chapter 4 focuses on developing a novel SSL method for more robust local representation learning, while Chapter 5 focuses on enhancing the understanding of SSL through explainability. Table. 3.2 shows the summary of the unlabelled datasets used in this thesis.

Dataset I, which has been used in chapter 4, is extracted from 719 full-length second trimester fetal US scans collected between May 2018 and March 2023, with an average duration of 34.2 ± 12.1 minutes per scan, totaling approximately 409

hours. Invalid data, such as initial frames without relevant information before the sonographer begins the anatomy search, are discarded. After temporally down-sampling at a rate of eight, a total of 2,261,179 frames are obtained, resulting in 70,661 clips, each with a length of 32 frames. A length of 32 frames is chosen to ensure that each clip contains enough information for analysis, while not being so long that multiple scanning tasks might be performed within a single clip. This balance allows for more focused and relevant representation learning.

Dataset II, which has been used in chapter 5, is a subset of Dataset I, consisting of 342 full-length second trimester scans collected between May 2018 and June 2020, due to the earlier timeline of that research work. Using similar selection criteria, 715,968 frames were selected, resulting in 119,328 clips, each consisting of 6 frames. The clip length of 6 frames is selected based on the design used in [258].

All frames were central cropped to remove the fan shape to avoid trivial solution, before any further augmentations or processing.

3.4 Labelled US Frame Dataset

Three labeled datasets (Dataset III, IV, and V) were used for supervised training in this thesis, which is summarised in Table. 3.3. Dataset III, from first-trimester scans, is used in Chapter 4 for cross-domain downstream task training. Dataset IV, from second-trimester scans, is used in Chapters 4 and 5 for in-domain downstream task training. Dataset V, also from second-trimester scans, is used as the in-clinical distribution (ICD) data for out-of-clinical distribution (OCD) detection. The following subsection provides a detailed explanation of each dataset.

3.4.1 Anatomy Recognition Dataset (Dataset III)

The first trimester fetal US scan is an essential part of prenatal care [259]. It is offered to pregnant women to assess fetal viability and determine pregnancy dating by measuring the fetal crown-rump length (CRL). Additionally, it helps to estimate the likelihood of chromosomal abnormalities by measuring the fetal nuchal translucency (NT).

Dataset	Trime-ster	Number of US scans	Mean scan length (minutes)	Number of frames	Number of classes	Class Informa-tion
Dataset III	1st	95	15.7 ± 4.2	55,871	5	Anatomy
Dataset IV	2nd	135	35.6 ± 8.4	15,384	14	Standard plane
Dataset V	2nd	212	37.6 ± 12.4	10,280	13	Standard plane

Table 3.3: Summary of the labelled US video dataset

In Dataset III, 95 first-trimester scans were used to create an annotated dataset on first-trimester anatomy, which was done by Robail, a former post-doc of our lab. According to the sonographers' workflow, they scan to locate the anatomy of interest, freeze the frame when a satisfactory view is obtained, and then perform measurements on the US video buffered in the machine. In the first trimester, the following actions are typically performed on the buffered video [260]:

- **Diagnostic inspection:** such as head, heart, abdomen
- **Biometric measurement:** such as NT, CRL, head circumference
- **Doppler or pulse-Doppler based measurement:** such as maternal uterine artery
- **3D-mode surface rendering of anatomy**

Hence, the dataset consists of five key anatomical categories, including crown rump length (CRL), nuchal translucency (NT), biparietal diameter (BPD), 3D-mode (3D) and other (BK). Although 3D mode is not an anatomical structure, it is included as an important aspect of the overall scanning process, allowing for the tracking and evaluation of the time spent scanning in 3D mode. The "other" class includes various minority classes, such as the placenta. In total, 55,871 labelled frames make up this first trimester anatomy dataset, as summarised in Table

Table 3.4: Standard plane classes with description

Class	Description
four views of heart	
3VT	Three-vessel and trachea view
4CH	Four-chamber
RVOT	Right ventricular outflow tract
LVOT	Left ventricular outflow tract
two views of brain	
BrainTv.	Brain view at the posterior horn of the ventricle
BrainTc.	Brain view at the level of the cerebellum
two views of spine	
SpineCor.	Coronal spine view
SpineSag.	Sagittal spine view
Abdominal	Standard abdominal view at stomach level
Femur	Stardard femur view
Kidneys	Axial kidneys view
Lips	Coronal view of the lips and nose
Profile	Medial facial profile
Background	Non-modelled standard views and background frames

3.3. This dataset is used to train and test the first-trimester anatomy recognition task described in Chapter 5, with 73 scans included in the training set, 16 in the validation set, and 6 in the test set.

3.4.2 Standard Plane Dataset (Dataset IV)

In general, obstetric US screening involves manual scanning, standard plane detection, biometric measurement, and diagnosis [3]. An US standard plane is defined as a specific 2D imaging plane that clearly depicts key anatomically meaningful structures or organs, used for subsequent biometric measurement or clinical condition diagnosis [239, 261].

During the routine scanning process, the sonographers scan to identify the anatomy. They freeze the frame when they approach a standard view and perform

subsequent biometric measurements. If the frozen frame does not provide a satisfactory standard plane, this search-freeze-measure procedure is repeated. This process always occurs in real time, with sonographers often revisiting the same anatomical area multiple times to acquire the best possible view—referred to as the standard plane—for improved anatomical assessment and measurement.

A large fraction of the frozen views saved by sonographers corresponded to standard planes and have been manually annotated during the scan allowing us to infer the correct ground-truth label. Based on these labels, 14 classes have been identified and split, with an overview provided in Table 3.4. It includes 13 anatomy classes, along with a a background class, which encompasses both non-standard frames and non-anatomy frames.

According to Table 3.3, annotations from 135 ultrasound scans were included to form Dataset IV, comprising a total of 15,384 labelled frames across defined 14 classes. In the downstream task of standard plane detection, which is used in both Chapter 4 and 5, the 135 scans are used with three-fold cross-validation by equally-divided subsets, resulting 90 training scans and 45 testing scans.

3.4.3 ICD Dataset (Dataset V)

In chapter 6, the definition of "Out-of-Clinical Distribution" is proposed based on the presence of a clinically interesting region within each medical image. According to this definition, the 13 standard plane views mentioned in 3.4.2 are considered ICD data, as they provide an optimal view of the anatomy. Therefore, Dataset V, comprising 212 scans and 10,280 labeled frames across 13 anatomical classes, is used as the ICD dataset. An additional OCD dataset was manually created, which will be discussed in detail in Chapter 6.

3.5 Evaluation Metrics

To quantitatively evaluate the model performance on above-mentioned downstream tasks, including standard plane detection task and anatomy recognition task, precision, recall, and F1 scores were reported. Precision measures how many

positive predictions are correct, recall measures how many positive cases have been correctly predicted and F1 score measures the harmonic mean of the precision and recall. The evaluation metrics are calculated as follows:

$$Precision = \frac{True\ positive}{True\ positive + False\ Positive}, \tag{3.1}$$

$$Recall = \frac{True\ positive}{True\ positive + False\ negative}, \tag{3.2}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{3.3}$$

Gazing at the banks of the River Qi, where the green bamboo flourishes. A noble man stands, like jade being cut and polished, like stone being carved and ground.

— *The Book of Songs, The Odes of Wei, The Banks of Qi.*

4

Fetal Ultrasound Video Representation Learning using Contrastive Rubik's Cube Recovery

Contents

4.1 Introduction

The notable success of deep learning across various domains has greatly relied on the availability of extensive annotated datasets. However, challenges arise when obtaining manual human annotations becomes non-trivial, as seen in domains like medical imaging. To address this, self-supervised representation learning is emerging as a promising alternative, demonstrating potential in both image and video domains [42, 262]. It releases the burden of heavy labour manual annotation and has gained significant attention across various medical imaging modalities to effectively obtain meaningful representations with large-scale unlabelled medical dataset and transfer the learnt representation to achieve promising results in clinical downstream tasks with small amount of data [44].

Self-supervised representation learning methods can be categorised into several classes based on their learning objectives, for instance, predictive pretext task (e.g. temporal sequence sorting [258], rotation prediction [263]) and contrastive learning (e.g. instance discrimination tasks [58]). Most prior works on SSL applied to US image analysis have focused on pretext tasks involving spatial transformations. As US scanning usually include a video recording of the US scan, some recent works explore SSL for the entire video instead of video frames, to learn both spatial and temporal representations. Jiao et al. propose a joint reasoning approach to learn representations from both order correction and geometric transformation [57]. As contrastive learning has become one of the leading paradigms of SSL and demonstrates stronger generalisation to downstream tasks [58], Chen et al. propose the US semi-supervised contrastive learning (USCL) method [264] and Zhang et al. design the hierarchical contrastive (HiCo) learning method [265] for US video, which provides state-of-the-art performance.

Most existing US video pre-training methods generate contrastive pairs using video-level data augmentations [124][264][265]. Two main types of augmentations are spatio-temporal transformations (e.g. cropping, shuffling) and colour transformations (e.g. solarisation) [266]. Normally, augmented views from the same video are referred to as positive pairs, and samples from different videos are referred to

as negative pairs. CL learns global representation by relying on the representation invariant of positive pairs [58]. However, given the fact that fetal US videos often share a global spatial pattern with significant local variations, I are motivated to explore the use of local contrastive pairs generated from sub-cubes of US videos for local representation learning.

For video self-supervised representation learning, [267] has demonstrated that directly applying temporal augmentations (e.g., shuffle and reverse) at the input level instead of the feature level could hinder feature learning. Representations learned from different network levels exhibit varying degrees of generalization and abstraction. Low-level features focus on simple structural patterns, while high-level features capture semantic concepts. High-level features are more representative of instances or semantics with context-aware representations, providing a global and abstract understanding of objects and concepts. In contrast, low-level features are more transferable since simple patterns are typically not task-specific. However, they lack structural consistency across samples and are particularly sensitive to temporal variations, as they tend to capture fast-changing details and are more prone to noise [268]. In fetal US, low-level features capture edges (e.g., tissue boundaries) and noise (e.g., acoustic shadows), middle-level features learn local anatomical patterns and structures (e.g., ventricles and heart chambers), and high-level features capture global context and semantic information (e.g., identifying complete anatomical structures). As a result, middle-to-high-level features tend to be more robust to noise and better at filtering out irrelevant variations as the network focuses on meaningful semantic regions. Given that fetal US videos are often degraded by acoustic shadows and speckle noise, I are particularly interested in transferring the pretext task at the feature level rather than the pixel level to enable more robust representation learning.

Meanwhile, SSL has been significantly influenced by the use of CNNs due to their ability to learn highly complex representations [269], including in medical imaging [57]. The CNN algorithm utilises convolutional operators to learn generalised representations for medical imaging tasks, but its limited local receptive field hinders

its ability to capture long-range spatial relationships in images. This limitation creates a demand for an alternative architecture that can effectively capture long-range dependencies and learn stronger feature representations. Vision transformer [270], one of the latest advancements in deep learning, can serve as an alternative to traditional CNNs. Their larger effective receptive fields enable better capture and understanding of contextual information, which is particularly important in medical imaging analysis [269]. In AI-assisted diagnosis, it is important not only to focus on the area of clinical interest but also to account for surrounding tissues and organs [271]. This broader contextual awareness could lead to more comprehensive diagnoses. Therefore, I are interested in exploring whether transformer models can help extract stronger features and benefit self-supervised medical imaging analysis.

As a result, in this chapter, I aim to take a further look at self-supervised representation learning by addressing the following specific questions:

- ***Question 1****: Would performing pre-text tasks at the feature level, instead of the input raw data level, enhance feature learning?*
- ***Question 2****: How might utilising a combination of pre-text tasks enhance self-supervised representation learning? How can I design an objective function to effectively combine multiple pre-text tasks?*
- ***Question 3****: Would stronger feature extractors, for instance, transformer-based models, contribute to better feature representations learned by self-supervised representation learning?*

By answering the above questions, our main contributions are as follows:

1. I propose a SSL framework, named **feature-level Contrastive Rubik's Cube Recovery (CRCR)** for fetal ultrasound video analysis, which is customised to leverage the advantage of contrastive learning with Rubik's cube recovery and cube reconstruction pre-text tasks for local representation learning;

2. I introduce an approach to generate local contrastive pairs from sub-cubes of US features, which facilitate discriminative and consistent local representation learning;

3. I empirically compare the effect of using feature-level pretext tasks and stronger feature extractor (e.g. 3D Swin Transformer [272]) to enhance feature learning;

4. The proposed method CRCR consistently outperforms existing SSL methods on both in-domain and cross-domain US clinical downstream tasks, showing its effectiveness and generalisability.

4.2 Originality and Individual Role

I preprocessed and selected raw fetal ultrasound video clips used in this chapter with a Python-based library developed by Richard Droste, a previous D.Phil student at the Institute of Biomedical Engineering, University of Oxford. I designed, trained, and tested the proposed self-supervised representation learning method using a Python-based open-source library PyTorch [273]. This work was published and presented in ASMUS 2024, MICCAI workshop [9]. The presentation was awarded the Best Presentation Runner-Up, and the paper received the Best Paper award.

4.3 Background

4.3.1 Contrastive Representation Learning

Contrastive representation learning is a self-supervised representation learning method that learns useful and transferable representations by teaching the machine to discriminate similar and dissimilar data pairs [58]. This method learns representation in a different way as compared to the more traditional generative or discriminative models and has become the leading paradigm.

A simple framework for contrastive learning (SimCLR [58]) is shown in Fig. 4.1. Given an augmentation family $\mathcal{T}$, two different data augmentations $T_1 \sim \mathcal{T}$ and

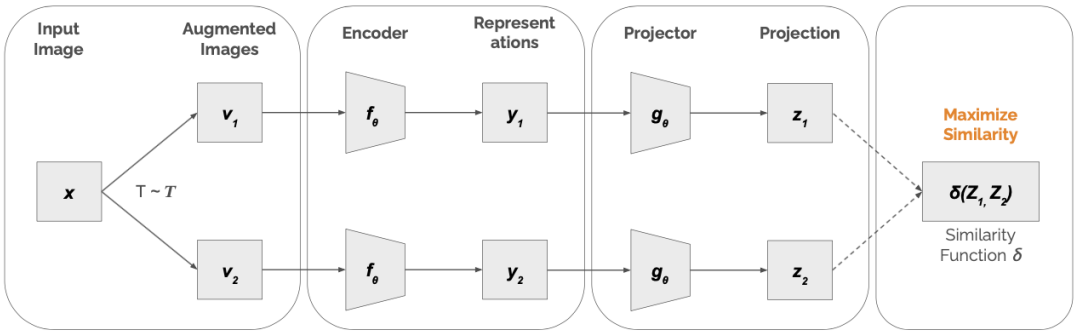


Figure 4.1: A simple framework of self-supervised contrastive representation learning. An image X is taken and random augmentations that are sampled from a family of augmentations T are applied to it to get a pair of two augmented images v_1 and v_2. Images from the pair are sent through the encoder f_θ to obtain representations and a following projector to obtain projections z_1 and z_2. The task is to maximise the similarity between these two representations with similarity function $\delta(.)$

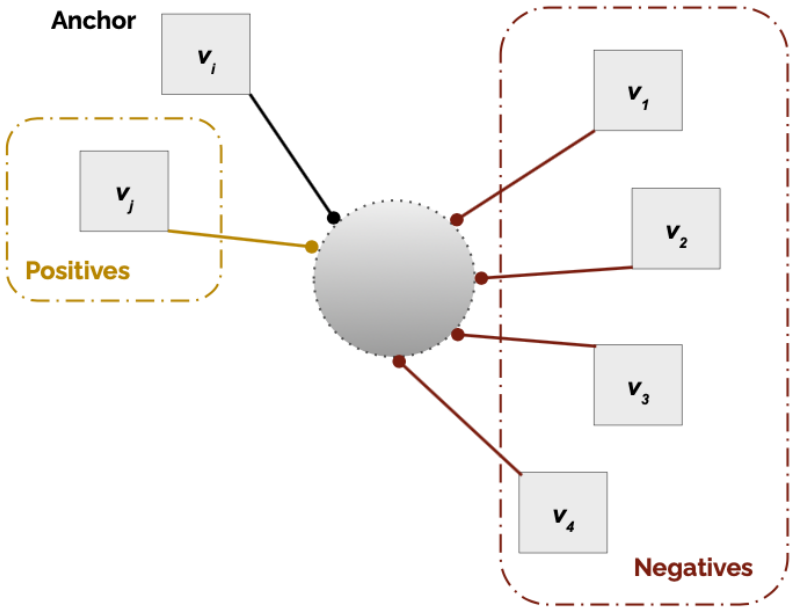


Figure 4.2: Visual demonstration of contrastive representation learning loss. For each anchor $X_i^{T_1}$, a single positive (e.g. the augmented pair of image X_i) is contrasted against a set of negatives, including the augmented pairs of the remaining images in the mini-batch $\{X_k\}_{k=1,k\neq i}^{n}$. The augmented images are projected onto the representation space for loss calculation, depicted as the grey sphere in the figure.

$T_2 \sim \mathcal{T}$ are sampled and performed for each input image x to obtain two augmented views of image, which are denoted as v_1 and v_2. Since the data augmentations are sampled from the same family of augmentation combinations, the augmented images are correlated views of the original image. The goal is to maximize the agreement between these pairs, which are referred to as positive pairs.

For a mini-batch of N samples that are randomly sampled from a pre-training dataset D, the framework yields $2N$ augmented images. A base encoder network f_θ is applied to each augmented image to extract a representation vector. While the augmented images v_i and v_j are treated as positive pairs, negative pairs are not sampled explicitly. Instead, the other $2(N-1)$ augmented images within the same mini-batch are regarded as negative pairs to x_i. A contrastive loss function is defined as

$$L_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{k \neq i} \exp(\text{sim}(z_i, z_k)/\tau)}. \tag{4.1}$$

where $1_{k \neq i}$ is an indicator function, τ is the temperature parameter, sim(.) is the pairwise cosine similarity function and z is the representation vector with $z_i = g_\theta(f_\theta(v_i))$. A detailed demonstration of the loss generation process is provided in Algorithm 1. A visual demonstration of the contrastive representation learning loss is illustrated in Fig. 4.2. For a given anchor image, $L_{i,j}$ tries to maximise its similarity with its augmented view while minimising its similarities with other input images in the mini-batch in the representation space.

4.3.2 Rubik's Cube Recovery

The Rubik's Cube Recovery method [50] is a patch-based method implemented in 3D. Inspired by the jigsaw puzzle, the authors [47] proposed a pretext task to deeply extract useful representations from the 3D medical data. The pipeline is illustrated in Fig. 4.3. By denoting $f_i \in \chi$ as the i-th frame of the video dataset χ, a video clip of length k, $x = \{f_1, f_2, ..., f_k\}$ forms a 3D medical volume (2D + temporal dimension). The 3D volume is partitioned into a grid (e.g. $2 \times 2 \times 2$) of cubes and the sub-cubes are then randomly shuffled and rotated. The task is to recover the

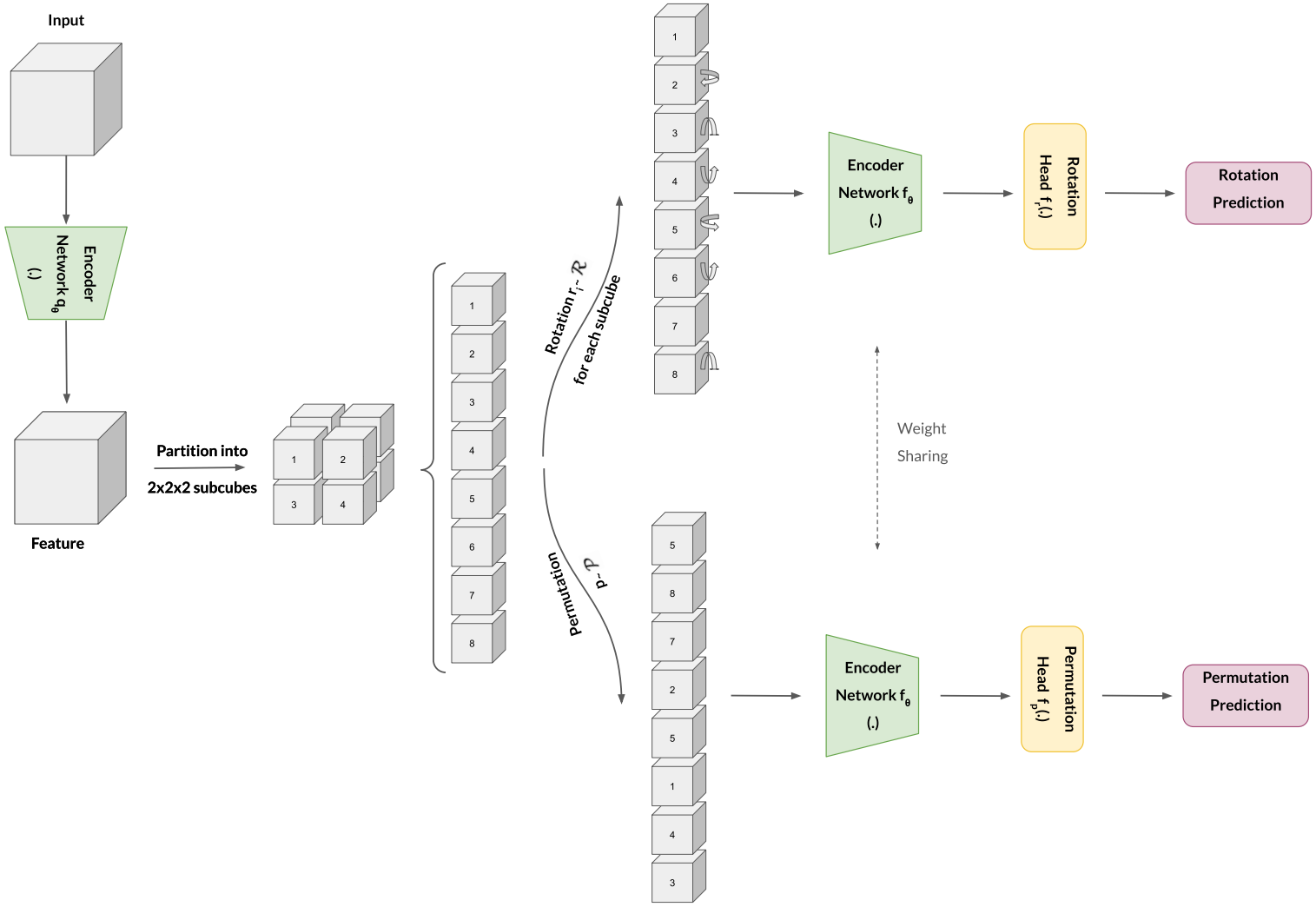


Figure 4.3: Pipeline of self-supervised representation learning via recovering a Rubik's cube puzzle. Two sub-tasks are involved, e.g. cube ordering and cube orientation. For cube ordering, the eight sub-cubes are rearranged according to a permutation selected from all possible options. In contrast, for cube orientation, rotations that are randomly drawn from all possible rotation options are applied to each sub-cube individually, as shown by the grey arrows in the figure.

Algorithm 1 Loss generation algorithm for self-supervised contrastive learning based on instance discrimination tasks

Input: Sampled mini-batch $\{x_i\}_{i=1}^{N}$, family of augmentation $\mathcal{T}$,encode network f_θ

for i from 1 to N **do**
  Sample $v_{2i-1} \sim \mathcal{T}(x_i)$
  Sample $v_{2i} \sim \mathcal{T}(x_i)$
  $z_{2i-1} = g_\theta(f_\theta(v_{2i-1}))$
  $z_{2i} = g_\theta(f_\theta(v_{2i-}))$
end for
for i from 1 to N **do**
  $L_{2i,2i-1} = -\log \frac{\exp(\text{sim}(z_{2k}, z_{2i-1}/\tau)}{\sum_{k=1}^{2N} 1_{k \neq 2i} \exp(\text{sim}(z_{2i}, z_k)/\tau)}$.
end for

Output: $L = \frac{1}{2N} \sum_{k=1}^{2N} [L_{2i,2i-1}, L_{2i-1,2i}]$

Algorithm 2 Loss generation algorithm for self-supervised learning via recovering a Rubik's cube puzzle

Input: Sampled mini-batch $\{x_i\}_{i=1}^{N}$, set of permutations $\mathcal{P} = \{p_1, p_2, ..., p_K\}$, set of rotations $\mathcal{R} = \{r_1, r_2, ..., r_M\}$, encoder network f_θ, rotation prediction head f_r, prediction prediction head f_p, loss weights α, β

for all i from $\in \{1, ..., N\}$ **do**
  sample permutation $p_j \sim \mathcal{P}$
  $x_{i,1}, ..., x_{i,8} = \text{partition}(x_i)$ ▷ Partition into 8 sub-cubes
  $\hat{x_{i,1}}, ..., \hat{x_{i,8}} = p_j(x_{i,1}, ..., x_{i,8})$ ▷ Permutation on sub-cubes
  $l_j = f_p(f_\theta(\hat{x_{i,1}}, ..., \hat{x_{i,8}}))$ ▷ Permutation prediction
  $L_{P,i} = -\sum_{j=1}^{K} p_j \text{log} l_j$ ▷ Permutation loss
  for all $k \in \{1, ..., 8\}$ **do**
    sample rotation $r_j \sim \mathcal{R}$
    $\tilde{x_{i,k}} = r_j(\hat{x_{i,k}})$
    $g_j = f_r(f_\theta(\tilde{x_{i,k}}))$ ▷ Rotation prediction
    $L_{R,i,k} = -\sum_{j=1}^{M} r_j \text{log} g_j$ ▷ Rotation loss
  end for
end for

Output: $L = \frac{1}{2N} \sum_{i=1}^{N} (\alpha L_{P,i} + \beta \sum_{k=1}^{8} L_{R,i,k})$

original configuration, which is the cube's ordering and orientation. Hence, the task is able to learn both the transnational and rotational invariant features. For the cube ordering task, the objective is to predict the permutation applied to shuffle the eight sub-cubes. With a total of K possible permutations, this task can be framed as a K-class classification problem. In the cube orientation task, the goal is to determine the specific rotation that has been applied to each sub-cube. With a total of M possible rotations, this task can be framed as eight M-class classification problems.

Given the set of permutations $\mathcal{P} = \{p_1, p_2, ..., p_K\}$ and rotations $\mathcal{R} = \{r_1, r_2, ..., r_M\}$ that may be applied to the 3D volume x, the loss functions could be summarised as

$$L = \alpha L_P + \beta L_R \tag{4.2}$$

where α and β are loss weights that control the influence of two tasks. L_P and L_R are the permutation loss and rotation loss. Given an input 3D medical volume x_i, let p_j and l_j denote its ground truth and predicted permutation classes, respectively, and r_j and g_j denote its ground truth and predicted rotation classes for the $k-th$ sub-cube. The permutation loss and rotation loss are defined as follows:

$$L_{P,i} = -\sum_{j=1}^{K} p_j \log l_j \tag{4.3}$$

$$L_{R,i,k} = -\sum_{j=1}^{M} r_j \log g_j. \tag{4.4}$$

Here, cross-entropy loss is used instead of MSE loss, as the rotation is constrained to horizontal or vertical flips to align with the nature of Rubik's cube transformations. Therefore, multi-class classification is used instead of regression, where the degree of rotation is unconstrained. Algorithm 2 provides a detailed description of the process for generating loss for self-supervised learning via recovering a Rubik's cube puzzle.

4.3.3 Transformers in Medical Imaging Analysis

Transformers, as an alternative network architecture to CNNs, is firstly introduced by [274] in natural language processing tasks. Attention-based blocks are used as layers of neural network in transformer. In NLP, these attention blocks aggregate information and correlations from the entire sequence and improved the performance significantly. The architecture is then extended to computer vision, where vision transformers (ViTs) [270] has been developed. For ViTs [270], images are split into patches, and the sequence of linear embeddings of these patches is provided as input. As a result, the global context of an input image is learned and captured within the transformer layers. ViTs have achieved state-of-the-art performance

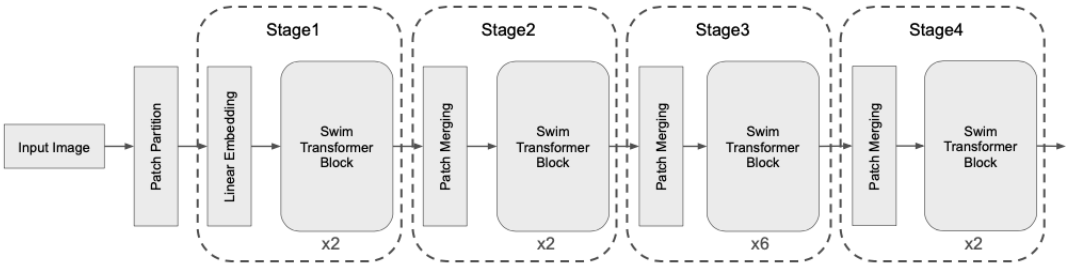


Figure 4.4: Pipeline of Swin transformer [272]

in various computer vision tasks [270] and therefore have gained a lot of interest in the field of medical imaging analysis.

A number of recent studies have focused on developing transformer-based models for medical imaging applications, which can be roughly classified into several categories: medical image segmentation (e.g., brain tumour [275], skin lesion [276]), medical image classification (e.g., COVID-19 [277], retinal disease [278]), medical object detection (e.g., MRI [279], spine CT [280]), and medical image reconstruction (e.g., MRI [281], CT [282]). Meanwhile, transformer-based models have also demonstrated their advantages in large-scale self-supervised pre-training on medical imaging datasets [283].

4.3.4 Swin Transformer

The Swin Transformer is an innovation in the field of vision transformer [272]. Most prior transformer-based architectures rely on fixed-scale patches as tokens, which are not ideal for general vision applications. Additionally, these architectures requires quadratic complexity when producing representations of a single token, as they compute the relationships between every pair of tokens. This poses a challenge for medical images, which are typically high-resolution and may be intractable for such transformer models. To address the challenge, the Swin transformer is proposed to construct hierarchical representations with vary-scaled tokens and has linear computational complexity to image size. Swin Transformer constructs a hierarchical representation by starting from treating small-sized patches as tokens and gradually merging neighbouring patches in deeper Transformer layers. The linear

computational complexity is achieved by computing self-attention locally within non-overlapping windows that partition an image. The number of patches in each window is fixed, and thus the complexity becomes linear to image size. Hence, the Swin Transformer is suitable as a general-purpose backbone for various vision tasks.

The general pipeline of the Swin transformer is illustrated in Fig. 4.4. An input image is first processed by the patch partition module, where it is split into non-overlapping patches. Each patch is treated as a token, and its feature is represented as a concatenation of the RGB values of raw pixels for RGB images, or simply the raw pixel values for greyscale images. A common choice of patch size is 4×4, resulting in a feature dimension of 48 for RGB images. Each patch is then passed to "Stage 1," which consists of a linear embedding layer and a Swin transformer block. The linear embedding layer projects the patch features into an arbitrary dimension C. The first Swin transformer block maintains the resolution at $\frac{H}{4} \times \frac{W}{4}$. "Stage 2" consists of a patch merging layer and a Swin transformer block. The patch-merging layer reduces the number of tokens by concatenating the features of each 2×2 neighboring patches into a $4C$-dimensional feature, followed by a linear layer that projects these features to a $2C$-dimensional space. The Swin transformer blocks are then applied at a resolution of $\frac{H}{8} \times \frac{W}{8}$. "Stage 3" and "Stage 4" repeat the same process, further reducing the resolution while increasing feature dimensionality. To produce a hierarchical representation, a shifting window partitioning strategy is employed. In the patch-merging layer, window-based partitioning and standard partitioning are used alternately. The window-based partition computes relationships within local windows, while the standard partition computes relationships between every pair of tokens. This window-based partitioning reduces computational complexity while maintaining cross-window connections by shifting the window partition at each layer. Such hierarchical representations could be beneficial for medical imaging by allowing multi-scale representations, anatomies of varying scales could be learnt for accurate analysis.

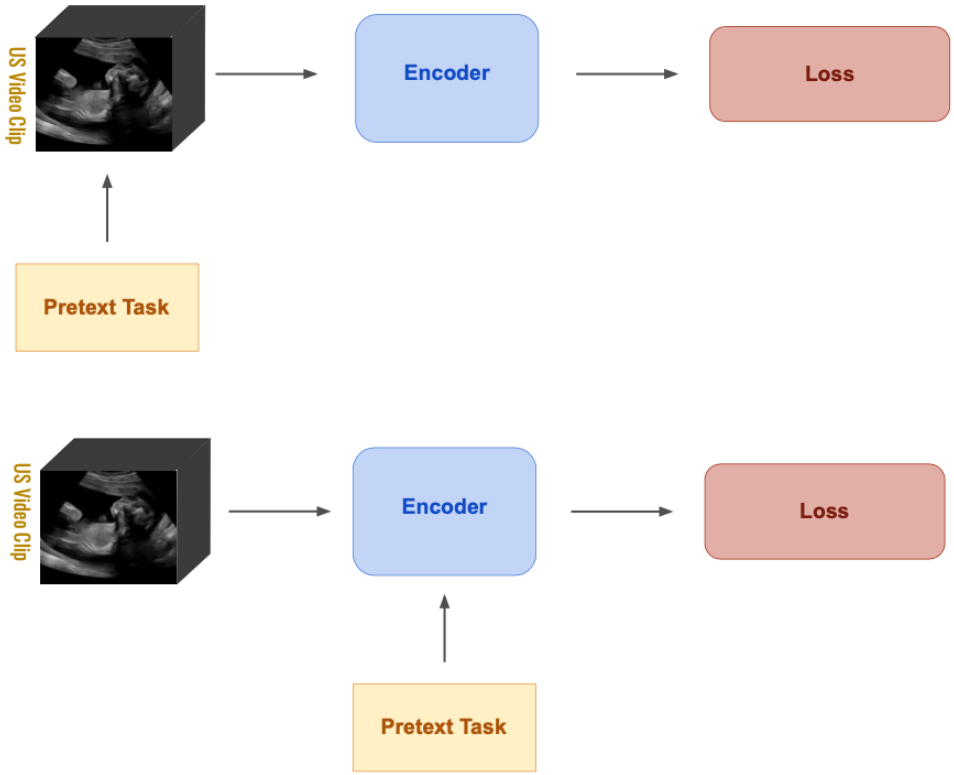


Figure 4.5: Pipeline of conventional self-supervised representation learning method on instance level and our proposed method on feature level.

4.4 Methods

4.4.1 Feature-level Pre-text Task

Contrastive representation learning based on instance discrimination has become the leading paradigm of self-supervised pre-training tasks. It is trained to pull together each instance and its augmented views, while push away from all other instances, in the embedding space. Since the quality of US videos is always affected by the extensive presence of speckle noises and acoustic shaded [8], whereas low-level information (e.g. boundaries) and local noise could be discarded at the middle to high-level feature level as discussed in the introduction. Therefore, it is worth considering operating a pre-text tasks at the feature level instead of the instance level.

Fig. 4.5 illustrates the general idea of our proposed self-supervised representation learning approach, which operates augmentations or distortions at the feature level instead of the original instance level. Here, I are particularly interested in feature-

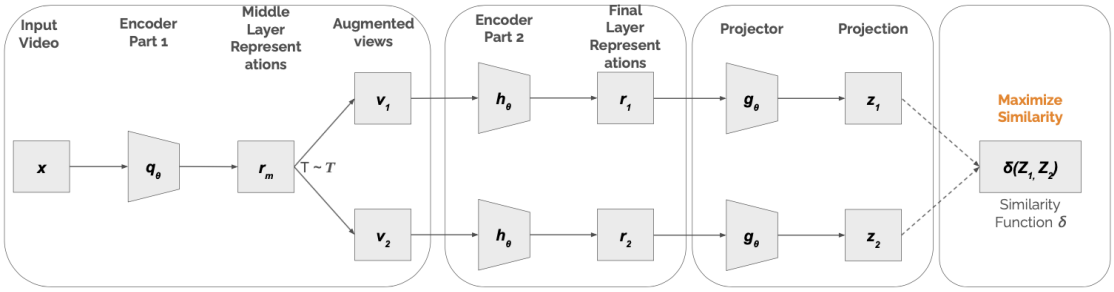


Figure 4.6: Pipeline of feature-level self-supervised contrastive representation learning.

level self-supervised contrastive representation learning. Fig. 4.6 illustrates the algorithm of the proposed method. Here, the encoder network is split into two smaller networks $q_\theta(.)$ and $h_\theta(.)$, where $f_\theta = h_\theta(q_\theta)$. For each input video x, instead of applying augmentations on x itself, here I send the x into $q_\theta(.)$ to get middle layer feature representations $r_m = q_\theta(x)$ and apply sampled augmentations on obtained feature representations to get augmented feature views v_1 and v_2. Similar to instance-level self-supervised contrastive representation learning, here I contrast the augmented feature pairs with the other features in the mini-batches and optimize the contrastive loss. For each input video x_i, the contrastive loss is calculated as follows:

$$L_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{k \neq i} \exp(\text{sim}(z_i, z_k)/\tau)}. \tag{4.5}$$

where $1_{k \neq i}$ is an indicator function, τ is the temperature parameter, sim(.) is the pairwise cosine similarity function and z is the projection vector with $z_i = g_\theta(h_\theta(v_i))$. A detailed demonstration of the loss generation process is shown in Algorithm 3.

4.4.2 Stronger Feature Extractor

Given that our study focuses on self-supervised video representation learning through feature-level optimization, it is worth considering whether using stronger feature extractors is beneficial. Vision transformers have shown strong performance in medical image analysis [269]. The Swin Transformer, a key innovation in vision transformers, is considered a highly effective feature extractor for images [272]. As introduced in Section 4.3.4, a Swin transformer can construct multi-scale

Algorithm 3 Loss generation algorithm for self-supervised contrastive learning based on feature discrimination tasks

Input: Sampled mini-batch $\{x_i\}_{i=1}^{N}$, family of augmentation $\mathcal{T}$,encode network q_θ and q_θ, projector g_θ

1: **for** i from 1 to N **do**
2: $r_{m,i} = q_\theta(x_i)$
3: Sample $v_{2i-1} \sim \mathcal{T}(r_{m,i})$
4: Sample $v_{2i} \sim \mathcal{T}(r_{m,i})$
5: $z_{2i-1} = g_\theta(h_\theta(v_{2i-1}))$
6: $z_{2i} = g_\theta(h_\theta(v_{2i-}))$
7: **end for**
8: **for** i from 1 to N **do**
9: $L_{2i,2i-1} = -\log \frac{\exp(\text{sim}(z_{2k}, z_{2i-1}/\tau)}{\sum_{k=1}^{2N} 1_{k \neq 2i} \exp(\text{sim}(z_{2i}, z_k)/\tau)}$.
10: **end for**

Output: $L = \frac{1}{2N} \sum_{k=1}^{2N} [L_{2i,2i-1}, L_{2i-1,2i}]$

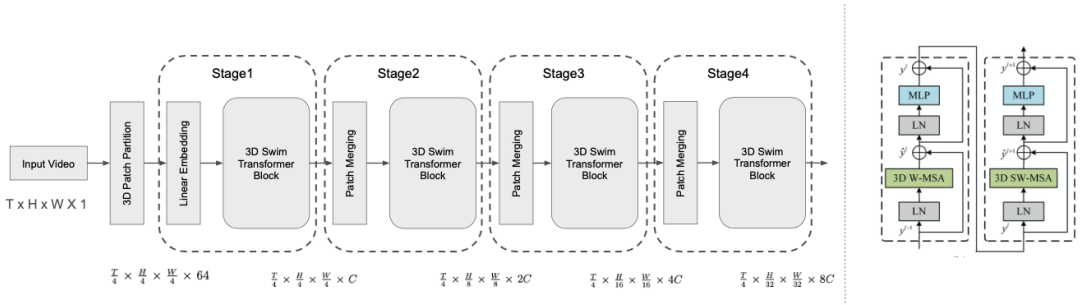


Figure 4.7: Pipeline of 3D Swin transformer

features by continuously fusing neighbouring patches and a window partition algorithm. I chose the Swin Transformer over other vision transformers for two main reasons: (1) Traditional vision transformers require quadratic complexity, whereas the Swin Transformer, through window-based self-attention, reduces this to linear complexity—critical for medical imaging, which typically involves high-resolution data. (2) Its hierarchical feature extraction structure enables multi-scale representation learning, improving the model's ability to capture complex anatomical patterns.

In this section, I consider the 3D Swin Transformer, based on the 2D Swin Transformer, as a stronger feature extractor to replace 3D CNNs. Here, the 2D Swin transformer is extended to a 3D structures (2D spatio-temporal structures) to capture rich spatial and temporal information from our 3D input, which is an

ultrasound video clip. This idea is inspired by early works on 3D Swin transformers, including [272] and [284]. Fig. 4.7 shows the architecture of the 3D transformer. I define each ultrasound video clip input of dimension $T \times H \times W \times 1$, where T is the number of frames in the clip, H and W denote the height and width of the frame, respectively, 1 represents the grey-scale property of ultrasound video. To adapt the 3D Swin Transformer for the Rubik's Cube recovery task, instead of directly extracting patch embeddings, I first partition the 3D cube into non-overlapping $2 \times 2 \times 2$ groups. Each group is treated as a macro-block before patch embedding, which aligns with the natural partitioning of a Rubik's Cube. A transformer encoder is applied at the macro-block level to learn the relationships between the four macro-blocks, capturing their relative positions through self-attention. This step ensures spatial consistency for Rubik's Cube recovery while reducing computational cost. After learning the macro-block relationships, I apply $2\times2\times2$ patch embedding within each macro-block to extract fine-grained local features. This results in $\frac{T}{2} \times \frac{H}{2} \times \frac{W}{2}$ patches, each with a feature dimension of 8. Then the feature is projected into a dimension of $C = 48$ and sent into the first 3D Swin transformer block. Compared to the traditional 2D Swin transformer, a temporal domain is added to the 3D Swin transformer, resulting a 3D window-based multihead self-attention (W-MSA) and 3D shifted window-based multihead self-attention (SW-MSA). The patching merging module works similarly to 2D but only downsamples the spatial dimension without temporal dimension to concatenate the neighboring 2×2 patches into a large patch. Therefore, for each stage, the size of the feature map halves in height and width, while is unchanged in the temporal dimension. For US video clips, if the duration is too long, the sonographer may perform multiple tasks within the clip, potentially leading to misleading representations during the learning process.

In general, the learning process for the 3D Swin transformer is represented by the following equation:

$$\hat{y}^l = \text{3D W} - \text{MSA}(LN(y^{y-1})) + y^{y-1} \tag{4.6}$$

$$y^l = \text{MLP}(LN(\hat{y}^l)) + \hat{y}^l \tag{4.7}$$

$$\hat{y}^{l+1} = \text{3D SW} - \text{MSA}(LN(y^y)) + y^y \tag{4.8}$$

$$y^{l+1} = \text{MLP}(LN(\hat{y}^{l+1})) + \hat{y}^{l+1} \tag{4.9}$$

where $\hat{y}^l$ and y^l represents the outputs of 3D W-MSA and MLP in block l, respectively.

4.4.3 Multi Pre-text Tasks

In the previous section, I proposed to move from video-level CL to feature-level CL to mitigate the influence of irrelevant low-level information and noise, and I introduced the use of the Swin transformer as a more powerful feature extractor. In this section, I explore how to further enhance the proposed method by considering the special properties that owned by US videos - fetal US often exhibit global patterns with significant local variations. Therefore, I are motivated to utilise multiple pre-text tasks to learn local features, in contrast to traditional CL methods, which primarily capture global frame representations.

In this section, I introduce an approach that integrates other pre-text tasks, such as Rubik's cube recovery and cube reconstruction, with feature-level contrastive learning. Rubik's cube recovery is chosen for its ability to partition data into sets of local features, with each feature subjected to both translational and rotational distortions, allowing for further local contrast to be drawn. Cube reconstruction is chosen because a distorted cube loses all spatio-temporal information, making predictions unreliable, while predictions from the original cube retain valuable information. This enhances the contrast between distorted and original data. This combined method is referred to as *Contrastive Rubik's Cube Recovery (CRCR).*

CRCR Framework

The overall framework of our method is illustrated in Fig. 4.8, which consists of two encoders $q_\theta(.)$ and $h_\theta(.)$, one projection head $g_\theta(.)$ for contrastive learning, two MLP head $f_r(.)$ and $f_p(.)$ for Rubik's cube recovery task and one decoder $d(.)$ for the cube reconstruction task.

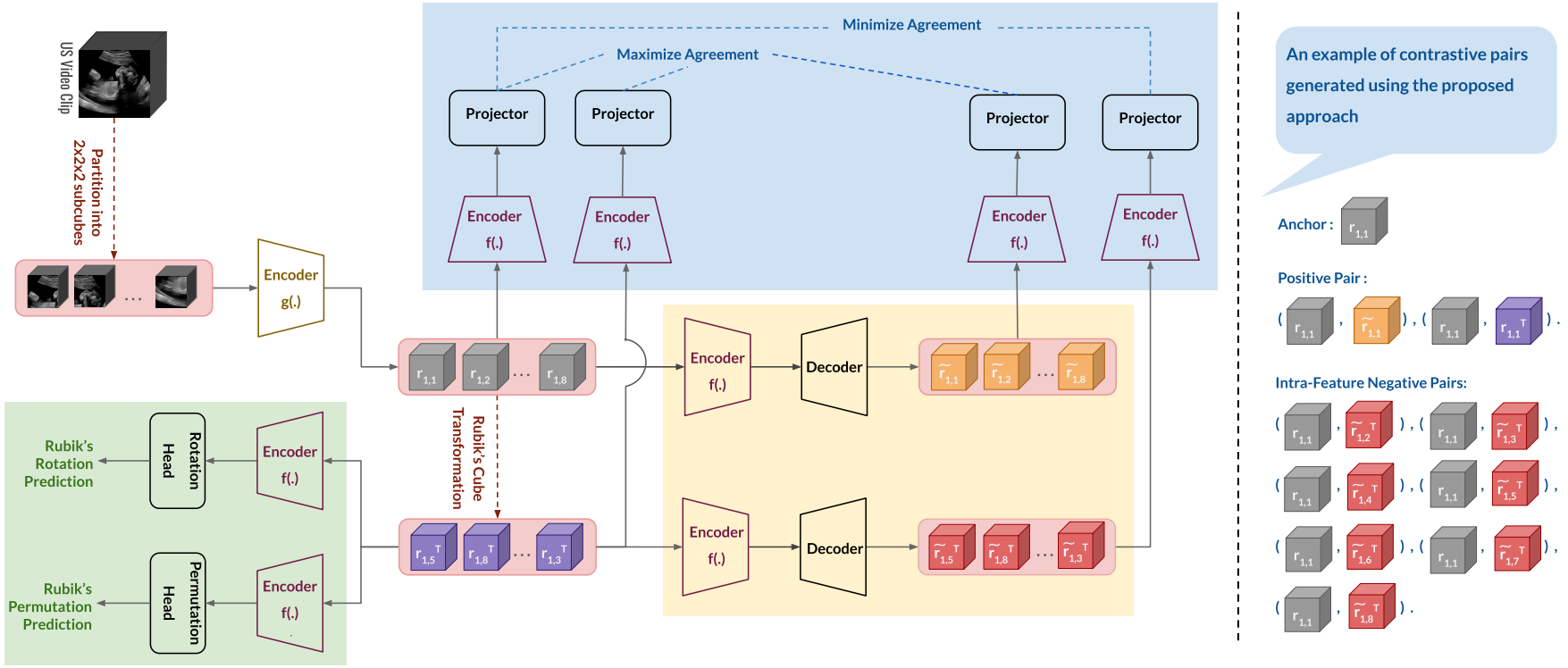


Figure 4.8: The pipeline of the proposed Contrastive Rubik's Cube Recovery (CRCR) framework, consisting of three pretext tasks: contrastive learning (the blue block), cube reconstruction (the yellow block), and Rubik's Cube recovery (the green block). An example of our proposed local contrastive pair generation approach is demonstrated with a given anchor.

Different from the normal CL paradigm [58], CRCR introduces two novel designs for effective local representation learning. Firstly, I perform pretext tasks on the feature level instead of the video level, and secondly, I proposed an approach to generate local contrastive pairs from sub-cubes of US video. As shown in Fig. 4.8, an input US video clip x_i is firstly partitioned into $2 \times 2 \times 2$ sub-cubes $\{x_{i,j}\}_{j=1}^{8}$, following the partition of Rubik's cube. The sub-cubes of US video along with its position embedding are then sent into the encoder $q_\theta(.)$ to get sub-cubes of US features $\{r_{i,j}\}_{j=1}^{8}$. Distortions (Rubik's cube transformation and reconstruction) are operated on the obtained features. Local contrastive pairs are generated from sub-cube of US features with the designed distortions, as in Sec. 4.4.4.

4.4.4 Training Objectives

Rubik's cube recovery.

RCR, which includes both cube rotation and rearrangement, are designed to recognize the correct orientation after rotation and the correct order after permutation of the sub-cubes. Here, I select the set of permutations $\mathcal{P} = \{p_1, p_2, ..., p_K\}$ to have the largest Hamming distance for sub-cubes shuffling. Hamming distance is a metric for comparing two equal-length strings, it is the number of positions at which the corresponding symbols are different [285]. Therefore, largest Hamming distance ensures that every single sub-cube has been moved. The set of rotations $\mathcal{R} = \{r_1, r_2, ..., r_M\}$ is defined to ensure either a horizontal or vertical flip for each sub-cube. Hence, the distortions are maximised, leading to the greatest loss of spatio-temporal information.

Assume that RCR is operated on sub-cubes $\{r_{i,j}\}_{j=1}^{8}$. Let P_k denotes the permutation among the 8 sub-cubes and $\{R_{m,j}\}_{j=1}^{8}$ denote the relative rotation of each sub-cube. I sample $P_k \sim \mathcal{P} = \{p_1, p_2, ..., p_K\}$ and $R_{m,j} \sim \mathcal{R} = \{r_1, r_2, ..., r_M\}$. The transformed sub-cubes are denoted as $\{r^T_{i,(P_k)_j}\}_{j=1}^{8}$, with position embedding updated to aligned with P_k. Among these, $\{r^T_{i,j}\}_{j=1}^{8}$ represents the sub-cubes after rotation only and $\{r_{i,(P_k)_j}\}_{j=1}^{8}$ represents the sub-cubes after permutation only.

We assume that by both translational and rotational representations could be learnt by the task. The predicted results l_j and $\{g_{m,j}\}_{j=1}^{8}$ are obtained from the rotation and permutation heads, respectively. RCR loss is calculated as the sum of rotation and permutation loss as follows:

$$L_{RCR} = -\{\sum_{k=1}^{K} P_k \log l_k + \sum_{j=1}^{8}\sum_{m=1}^{M} R_{m,j} \log g_{m,j}\}. \tag{4.10}$$

Cube reconstruction.

Cube reconstruction is operated on both the original feature sub-cubes $\{r_{i,j}\}_{j=1}^{8}$ and the sub-cube obtained after Rubik's sub-cube $\{r_{i,j}^{T}\}_{j=(P_k)_1^8}$, with the position embedding j updated with the selected permutation. The obtained reconstructions are denoted as $\{\tilde{r_{i,j}}\}_{j=1}^{8}$ and $\{\tilde{r_{i,j}}^{T}\}_{j=1}^{8}$ respectively, with the latter's position embedding updated.

By completing the task, two sets of predicted sub-cubes are obtained: the predicted original sub-cubes and the predicted distorted sub-cubes. The reconstruction loss is calculated as follows:

$$L_{Recon.} = \alpha_1 \times \sum_{j=1}^{8} MSE(\tilde{r_{i,j}}, r_{i,j}) + \alpha_2 \times \sum_{j=1}^{8} MSE(\tilde{r_{i,j}}^{T}, r_{i,j}^{T}). \tag{4.11}$$

A grid-search hyperparameter optimisation was performed, where $\alpha_1 = 1, \alpha_2 = 0.4$. As in the grid search, if too much weight is placed on predicting distorted sub-cubes, it could negatively impact the overall performance.

Contrastive learning.

With the assumption that US videos share a global spatial pattern with local divergence, I propose generating local contrastive pairs using sub-cubes of US videos, instead of global pairs using the entire video.

Normally, CL considers different US videos as negative pairs, which might result in the high similarity between negative pairs (e.g. videos from the same scan or performing the same measurement task) and potentially mislead representation learning [264]. The proposed approach is designed to generate two sets of

strongly discriminative negative samples $I^- = \{I^-_{intra}, I^-_{inter}\}$ by introducing the aforementioned pretext tasks as effective distortion tools. I assume that rotating and shuffling the cube would cause severe spatial and temporal distortion, resulting in the loss of both spatial and temporal information and leading to poor-quality reconstruction. A Local positive samples set, I^+, is generated with appropriate similarities. For a given anchor sub-cube $r_{i,j}$, the local contrastive pairs generated from the proposed approach consist of:

- **Positive sample** $I^+ = \{r^T_{i,j}, \tilde{r_{i,j}}\}$: rotated and reconstructed views of anchor
- **Intra-feature negative samples** $I^-_{intra} = \{\tilde{r_{i,m}}^T\}_{m \neq j}$: distorted sub-cubes from the remaining sub-cubes within the same video
- **Inter-feature negative samples** $I^-_{inter} = \{\tilde{r_{k,l}}^T\}_{k \neq i, l=1}^{8}$: distorted sub-cubes from different videos

The above two sets of negative samples enable the model to learn representations from different perspectives. Inter-feature negative pairs enhance instance discrimination, while intra-feature negative pairs provide local contextual information.

An example of contrastive pairs generated using the proposed strategy is demonstrated in Fig. 4.8 and 4.9. Here, sub-cube $r_{1,1}$ is chosen as an anchor, generating 2 positive pairs, 7 intra-feature negative pairs, and $8 \times (N_b - 1)$ inter-feature negative pairs, where N_b denotes the batch size.

- First positive sample is the rotated view of the anchor $r^T_{1,1}$, which is the green sub-cube on the top-right. I select it as a positive pair because rotation is commonly used as a data augmentation technique in contrastive learning, and I assume that US videos remain semantically similar after rotation. This is due to the constant movement of the probe during an ultrasound scan, making it likely to capture different views from various angles.

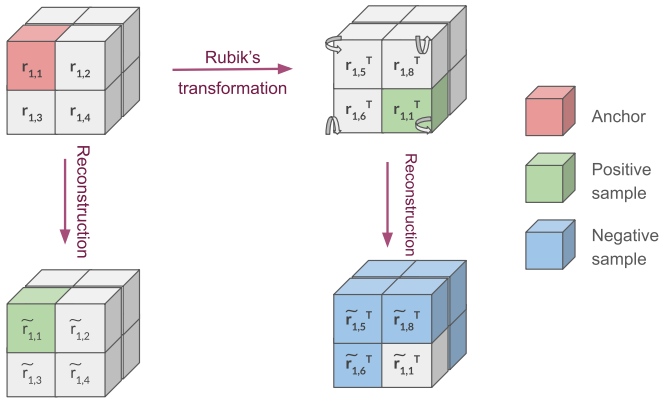


Figure 4.9: An example of local contrastive pairs generated using the proposed strategy, with a given anchor.

- Second positive sample is the reconstructed view of the anchor $\tilde{r_{1,1}}$, which is the green sub-cube on the bottom-left. It is the reconstruction based on the anchor $r_{1,1}$, with the objective to minimise the distance between the reconstruction and the original input. Hence, it shares the same objective as CL - to maximise the similarity between them.

- Seven intra-feature negative samples are $\{\tilde{r_{1,j}}^T\}_{j=2}^8$, which are the blue sub-cubes on the bottom-right. They are the reconstructions of the transformed sub-cubes based on the remaining sub-cubes within the same video clip, which are $\{r_{1,j}^T\}_{i=2}^8$. The reconstructed sub-cubes are expected to be lack of geometric and semantic information due to enforced spatio-temporal transformation. Therefore, the prediction would be unreliable and unrelated to the original input. I intentional does not include $\{\tilde{r_{1,1}}^T\}$ to maximise the divergence from the original input.

- $8 \times (N_b - 1)$ inter-feature negative pairs are $\{\tilde{r_{i,j}}^T\}_{i \neq 1 l=1}^8$, which are the reconstructions of the transformed sub-cubes based on the sub-cubes from different

video clips. The negative pairs are used to provide instance discrimination, which used the idea adapted from SimCLR.

Referred to [58][286], I adapt NT-Xent for our contrastive loss function, which is calculated as follows:

$$\begin{aligned} L_{CLR} &= -\sum_{j=1}^{8} \sum_{i^+ \in I^+} \log \frac{\exp(\varphi(r_{i,j}, i^+)/\tau)}{\sum_{i^- \in I^-} \exp(\varphi(r_{i,j}, i^-)/\tau)} \\ &= -\sum_{j=1}^{8} \sum_{i^+ \in I^+} \{\varphi(r_{i,j}, i^+)/\tau - \log \sum_{i^- \in I^-} \exp(\varphi(r_{i,j}, i^-)/\tau)\} \end{aligned} \tag{4.12}$$

where τ and $\varphi(.)$ denote the temperature parameter and the pairwise cosine similarity function, respectively.

Overall loss function.

The overall learning objective of CRCR is a weighted combination of Rubik's cube recovery loss, reconstruction loss, and contrastive loss, which is calculated as follows:

$$L_{CRCR} = \alpha \times L_{Rubik} + \beta \times L_{Recon.} + \eta \times L_{CLR}. \tag{4.13}$$

A grid-search hyperparameter optimization was performed which estimated the optimal values of $\alpha = \eta = 1, \beta = 0.5$.

A detailed description of loss generation process of the proposed CRCR approach is shown in Algorithm 4.

4.5 Experiments

4.5.1 Data

Our experiments are based on the large-scale fetal US video dataset, named PULSE, as described in Chapter 3. Dataset I, discussed in 3.3, consists of a large number of unlabelled US video clips and is used as the pre-training dataset. Dataset III, discussed in 3.4.1, consists of labelled first-trimester scan frames and is used for three-fold cross-validation on the cross-domain downstream task—anatomy

Algorithm 4 Loss generation algorithm of self-supervised Rubik's contrastive predicting

Input: Sampled mini-batch $\{x_i\}_{i=1}^{N}$, encode network $f_\theta = q_\theta(h_\theta)$, set of permutations $\mathcal{P} = \{p_1, p_2, ..., p_K\}$, set of rotations $\mathcal{R} = \{r_1, r_2, ..., r_M\}$, rotation prediction head f_r, permutation prediction head f_p, cube reconstruction network d, weight parameter $\alpha_1, \alpha_2, \alpha, \beta, \eta$

for i from 1 to N **do**
  $x_{i,1}, ..., x_{i,8} = \text{partition}(x_i)$ ▷ Partition into 8 sub-cubes
  $r_{i,1}, ..., r_{i,8} = q_\theta(x_{i,1}, ..., x_{i,8})$ ▷ Feature of each partition
  sample permutation $P_k \sim \mathcal{P}$
  for all $j \in \{1, ..., 8\}$ **do**
    sample rotation $R_{m,j} \sim \mathcal{R}$
    $r_{i,j}^{T} = R_{m,j}(r_{i,j})$ ▷ Rotated sub-cube
    $r_{i,(P_k)_j} = P_k(r_{i,j})$ ▷ Permuted sub-cube
    $r_{i,(P_k)_j}^{T} = \textsc{Rubik Transform}(r_{i,j})$ ▷ Rubik transformed sub-cube
    $g_{m,j} = f_r(h_\theta(r_{i,j}^{T}))$ ▷ Rotation prediction
    $(l_k)_j = f_p(h_\theta(r_{i,(P_k)_j}))$ ▷ Permutation prediction (j-th element)
    $\tilde{r_{i,j}} = d(h_\theta(r_{i,j}))$ ▷ Original sub-cube reconstruction
    $\tilde{r_{i,(P_k)_j}}^{T} = d(h_\theta(r_{i,(P_k)_j}^{T}))$ ▷ Transformed sub-cube recon.
    $I^+ = \{r_{i,j}^{T}, \theta(r_{i,j})\}$ ▷ Positive samples
    $I- = \{\tilde{r_{i,m}}^{T}\}_{m \neq j} \cup \{\tilde{r_{k,l}}^{T}\}_{k=1 \neq i, l=1}^{N \quad 8}$ ▷ Negative samples
    $L_{rot,i} = -\sum_{j=1}^{8} \sum_{m=1}^{M} R_{m,j} \log g_{m,j}$ ▷ Rotation recognition loss
  end for
  $L_{perm,i} = -\sum_{k=1}^{K} P_k \log l_k$ ▷ Permutation recognition loss
  $L_{Rubik,i} = L_{rot,i} + L_{perm,i}$ ▷ RCR Loss
  $L_{Recon.i} = \alpha_1 \sum_{j=1}^{8} MSE(\tilde{r_{i,j}}, r_{i,j}) + \alpha_2 \sum_{j=1}^{8} MSE(\tilde{r_{i,j}}^{T}, r_{i,j}^{T})$. ▷ Recon. Loss
  $L_{clr,i} = -\sum_{i^+ \in I^+} \{\varphi(r_{i,j}, i^+)/\tau - \log \sum_{i^- \in I^-} \exp(\varphi(r_{i,j}, i^-)/\tau)\}$ ▷ CLR Loss
end for

Output: $L = \frac{1}{N} \sum_{k=1}^{N} \{\alpha \times L_{Rubik,i} + \beta \times L_{Recon.,i} + \eta \times L_{CLR,i}\}$

recognition task. Dataset IV, discussed in 3.4.2, consists of labelled second-trimester scan frames and is used for three-fold cross-validation on the in-domain downstream task—standard plane detection. All frames were centrally cropped to remove the fan shape and resized to 224 × 224 pixels.

4.5.2 Model Implementation and Training

Both 3D ResNet-18 [287] and 3D Swin Transformers [284] are utilised as the backbone networks for the proposed self-supervised learning method. ResNet-18 is chosen based on prior work [264][51], while Swin Transformer is chosen for its

strong feature extraction capability. I return each sub-cube to its embedded position before passing it to 3D ResNet encoders to include positional information.

For SSL pretraining, I employ an AdamW optimiser [288] to maintain consistency with [283]. The initial default parameters are adopted from their work and further refined in our setting through grid search. Both networks are trained with a momentum of 0.9, a warm-up cosine scheduler of 500 iterations and a mini-batch of 32 for 40 epochs. After parameter tuning, an initial learning rate of 1×10^{-3} is used with decay of 0.1 for every 25K iterations for 3D ResNet-18 and an initial learning rate of 1×10^{-3} is used with decay of 0.01 for every 45K iteration for 3D Swin Transformer. All models are implemented with PyTorch [273], with our methods taking around 180 hours to run on a single NVIDIA Titan V GPU.

For transfer learning, I fine-tune the combined encoder $q_\theta(h_\theta(.))$ along with an attached classifier head. The following parameters have been fine-tuned by previous lab members for the corresponding tasks. For the standard plane detection task, networks are trained with SGD optimizer with momentum of 0.9 and mini-batch of 16 for 70 epochs. An initial learning rate of 0.01 is used with 0.1 decay at epochs 30 and 55. For first-trimester anatomies recognition, networks are trained with SGD optimizer with momentum of 0.9 and mini-batch of 16 for 200 epochs. An initial learning rate of 0.1 is used with a decay of 0.1 at epochs 150.

4.5.3 Comparison Methods

In this section, the following five methods have been selected for comparison, including those that use 3D ResNet18 and 3D Swin Transformer as backbones.

- **RCR [50]:** Rubik's cube recovery, which consists of two tasks, cube rearrangement that focuses on identifying the correct order and cube rotation that focuses on recognising the orientation.
- **SimCLR [58]:** A simple framework for contrastive learning of visual representations, which include one task of maximising agreement between differently augmented views of the same image.

- **DiRA [289]:** The first SSL framework that integrate discriminative, restorative, and adversarial learning in a unified manner.

- **USCL [264]:** An US CL method that avoid similarity conflict of CL by using only negative samples negative pairs coming from different videos.

- **HiCo [265]:** A hierarchical contrastive learning method with the objective to minimise the peer-level semantic alignment loss and cross-level semantic alignment loss.

- **SwinUNETR [283]:** A dubbed Swin UNEt TRansformers with tailored proxy tasks, including contrastive learning, masked volume inpainting and 3D rotation prediction.

These five methods have been chosen for comparison because they represent different perspectives:

1. **Baseline Comparison:** Our method builds on SimCLR and RCR; therefore, I include them as baseline methods for direct comparison.

2. **Multi-Pretext Task Integration:** Our method aims to improve performance by incorporating multiple pretext tasks. DiRA and Swin UNETR are two state-of-the-art methods that leverage other pretext tasks within CL frameworks. Additionally, they employ ResNet and Swin Transformer as backbones, respectively, making them suitable for comparison due to their similar design principles.

3. **Ultrasound-Specific Design:** Our method is specifically designed to address the unique properties of US data. USCL and HiCo are two state-of-the-art SSL methods developed for US video analysis, making them valuable for comparison as they are designed for the same medical imaging modality.

4.6 Results

4.6.1 Transfer Learning on Standard Plane Detection

Task description.

We first evaluate the pre-trained representation by transferring it to the in-domain second-trimester standard plane detection task. Similar to [57][290], 14 classes are considered, which includes four cardiac views, three-vessel and trachea (3VT), four-chamber (4CH), right ventricular outflow tract (RVOT), and left ventricular outflow tract (LVOT), two views of brain, transventricular (BrainTv.) and transcerebellum (BrainTc.), two views of spine, coronal (SpineCor.) and sagittal (SpineSag.), abdomen, femur, kidneys, lips, profile and background class. Precision, recall and F1-scores are used as the evaluation metrics.

Results.

Table 4.1 shows a quantitative comparison of fine-tuning performance on the standard plane detection task. The table indicates that CRCR generally outperforms four state-of-the-art CL-based methods, e.g. SimCLR [58], DiRA [289], USCL [264], and HiCo [265], by 3.8%, 1.9%, 2.0% and 1.1%, respectively, in F-1 score with the 3D ResNet backbone. Additionally, CRCR improves the performance of SimCLR and Swin UNETR by 3.7% and 1.9%, when using 3D Swin Transformer as the backbone. These improvements are statistically significant using the Wilcoxon signed-rank test [291], validating the effectiveness of our proposed method. The Wilcoxon signed-rank test is selected here as a non-parametric, robust statistical analysis test. In contrast, methods like the paired t-test require the assumption that the underlying distribution follows a t-distribution to ensure reliable results, which may not always hold in real-world data.

In particular, CRCR performs better with 3D Swin Transformer as the backbone network, demonstrating the benefit of utilising a stronger feature extractor to enhance feature learning. Table 4.2 presents the F1-score for each class, in which CRCR owns the best performance in the majority of classes. The improvement is particularly notable for *cardiac* views, which share a global heart perspective but

each has a local focus on specific structures, making them difficult to distinguish even for experts. This finding is in line with our assumption that by contrasting local contrastive pairs, the proposed method can learn semantically meaningful information from these local areas.

4.6.2 Transfer Learning on First-Trimester Anatomy Recognition

Task description.

We explore the generalisability of the pre-trained representation to a cross-domain first-trimester anatomy recognition task. Similar to [260], five key anatomy categories are considered, which include crown rump length (CRL), nuchal translucency (NT), biparietal diameter (BPD), 3D-mode (3D) and other (BK) for first-trimester fetal biometry measurements.

Results.

Table 4.1 demonstrates quantitative results of fine-tuning performance on the first-trimester anatomy recognition task. From the results, I observe that our proposed method achieves the best performance among all the compared methods with both 3D ResNet and 3D Swin Transformer as backbone networks. Most improvements are statistically significant with paired t-tests. This task of anatomical recognition could be challenging due to the small fetal size in the first trimester. While the compared CL-based methods focus on global representation learning, CRCR leans local patterns through local contrastive pairs, which could be valuable when the clinical region-of-interest is small. Among the compared methods, USCL and HiCo, which are designed specifically for US videos, perform better than those designed for general computer vision. Table 4.3 shows the F1-score for each anatomy category, in which CRCR achieves the highest F1 scores in all categories. This finding demonstrates the effectiveness and generalisability of our proposed method CRCR on first-trimester US video.

Table 4.1: Quantitative comparison of downstream task performance (mean±std.[%]) on second-trimester standard plane detection task and first-trimester anatomy recognition task. *Rand. Init.* indicates the 3D ResNet18 trained from scratch, and ‡ denotes the use of 3D Swin Transformer as backbone network. The best results for each backbone network are marked in **bold**. P-values are calculated between our CRCR results and the previous top-1 result for each backbone network. Any $p < 0.05$ represents a statistically significant improvement and is highlighted in blue.

	Second-trimester Standard Plane Detection			First-trimester Anatomy Recognition		
Pretrain Method	Precision	Recall	F1	Precision	Recall	F1
Rand. Init.	69.3±1.6	58.8±3.1	59.1±3.1	80.8±3.2	78.4±0.7	81.1±0.5
RCR [50]	70.0±0.8	66.3±3.5	66.4±2.3	89.2±1.4	88.6±1.6	89.5±1.7
SimCLR [58]	70.9±0.5	68.9±1.7	68.7±1.2	95.5±0.5	94.7±0.3	94.8±0.8
DiRA [289]	72.0±2.5	70.4±2.3	70.6±2.3	95.7±1.4	95.3±1.3	95.1±1.2
USCL [264]	71.6±1.1	70.2±1.5	70.5±0.9	95.9±0.7	95.2±0.8	95.3±1.5
HiCo [265]	72.3±1.7	71.7±1.9	71.4±2.0	96.3±0.3	95.7±0.7	95.8±0.9
CRCR (ours)	**73.1 ±2.3**	**72.6±2.7**	**72.5±2.2**	**96.8±1.2**	**96.1±1.4**	**96.2±1.8**
P-value	0.048	0.021	0.010	0.018	0.027	0.035
SimCLR[58]‡	72.8±0.8	70.7±1.2	70.7±1.6	95.9±1.3	95.3±0.8	95.1±1.5
SwinUNETR[283]‡	73.6±1.2	73.2±3.5	72.5±2.1	96.7±2.7	96.9±2.4	96.5±2.3
CRCR (ours)‡	**74.9±2.5**	**74.8±3.0**	**74.4±2.6**	**97.1±2.3**	**97.1±1.6**	**97.2±1.3**
P-value	0.003	0.001	<0.001	0.042	0.062	0.009

Table 4.2: Quantitative performance results on second-trimester standard planes detection using 3D ResNet as the backbone. F1-score per class (%)is used as the evaluation metric. *Rand. Init.* indicates training from scratch. The best results for each class are marked in **bold**.

Class	Rand. Init.	SimCLR	DiRA	USCL	HiCo	CRCR(ours)
3VT	30.2	39.2	41.2	41.7	43.0	**49.2**
4CH	25.1	33.6	35.1	35.6	37.1	**41.7**
LVOT	30.7	38.5	39.7	40.1	40.8	**42.5**
RVOT	28.8	35.3	36.8	36.9	37.6	**40.1**
BrainTv.	72.8	81.4	83.3	83.0	84.3	**85.3**
BrainTc.	69.9	81.2	83.1	82.8	83.1	**83.6**
SpineCor.	72.9	82.5	85.2	85.3	86.2	**86.4**
SpineSag.	77.3	84.3	86.5	85.4	86.9	**87.2**
Abdominal	65.0	76.5	78.2	78.5	79.4	**79.6**
Femur	80.9	88.7	89.8	89.4	**89.8**	89.6
Kidneys	62.4	74.6	77.4	77.1	78.3	**81.2**
Lips	76.7	84.5	87.3	86.9	86.7	**90.8**
Profile	70.5	79.9	80.9	81.3	**81.9**	81.5
Background	79.3	88.6	90.3	89.9	90.5	**90.6**

Table 4.3: Quantitative performance results on first-trimester anatomy recognition using 3D ResNet as the backbone. F1-score per class (%)is used as the evaluation metric. *Rand. Init.* indicates training from scratch. The best results for each class are marked in **bold**.

Class	Rand. Init.	SimCLR	DiRA	USCL	HiCo	CRCR(ours)
CRL	85.9	99.1	99.4	99.6	**99.8**	**99.8**
NT	78.2	92.6	92.9	93.2	93.6	**93.9**
BPD	79.6	92.5	92.8	93.3	93.8	**94.1**
3D	79.2	95.7	96.5	96.9	97.4	**98.7**
BK	86.7	99.4	99.5	99.5	99.7	**99.9**

4.6.3 Ablation Study

Two ablation studies discussed in this subsection are conducted using 3D ResNet18 as the backbone network.

Table 4.4: Ablation studies on the effectiveness of each self-supervised objective and the effectiveness of performing pretext tasks at the feature level. Experiments are fine-tuned using 3D ResNet for the standard plane detection task.

			Feature-level			Video-level		
L_{CLR}	L_{Rubik}	L_{Rec}	Precision	Recall	F1	Precision	Recall	F1
✓			71.2	69.4	69.5	70.9	68.9	68.7
	✓		70.4	66.7	66.4	70.0	66.3	66.4
		✓	70.3	66.1	65.8	69.8	64.2	64.5
	✓	✓	70.9	69.0	68.5	70.6	68.5	68.1
✓	✓		72.3	71.1	70.9	71.4	69.7	69.8
✓		✓	71.8	70.5	70.3	71.0	69.5	69.1
✓	✓	✓	**73.1**	**72.6**	**72.0**	**72.4**	**71.1**	**71.3**

Efficacy of self-supervised objectives.

We perform an empirical study on pre-training with different combinations of self-supervised objectives used in CRCR loss. Results are shown in Table 4.4. An improvement is observed with the addition of pretext tasks. By incorporating Rubik's recovery and cube reconstruction into contrastive learning, performance increases by 1.1% and 0.6%, respectively, with the combination of contrastive learning and Rubik's cube recovery emerges as the best-performing pair. Furthermore, the triple objective—combining contrastive learning, Rubik's recovery, and cube reconstruction—surpasses the best pair by 1.8% in precision. These results indicate that three selected tasks harmonise with each other and the proposed collaborative learning methods enhance representation learning and downstream performance.

Efficacy of feature-level pretext task.

We investigate how performing pretext tasks at the feature-level impacts representation learning compared to the video level. As illustrated in Table 4.4, performing pretext tasks at video level degrades the performance of the standard plane detection task to feature level under different combinations of SSL objectives. The largest drop occurs with the triple objective—combining contrastive learning,

which is our designed method—resulting in a decrease of approximately 1.5 % in recall. This aligns with our hypothesis that performing pretext tasks at the feature level enables the model to be less sensitive to superficial changes and focus more on informative regions, as local noise and irrelevant features could be discarded through encoder optimisation.

4.7 Conclusion

In this chapter, I have presented a novel self-supervised learning method named feature-level Contrastive Rubik's Cube Recovery (CRCR) for local representation learning of fetal US video. The proposed method leverages the advantages of contrastive learning with Rubik's cube recovery task and cube reconstruction task. A local contrastive pair generation approach is introduced to contrast sub-cube pairs from US video. This feature-level approach is motivated by the observation that mid-level features can effectively filter out noise in US videos, which otherwise significantly degrades video quality and impairs feature learning. The design of local representations is based on the inherent property of US videos, which consistently exhibit a global pattern with significant local variations.

Through extensive experiments, it can be concluded that operating pretext tasks at the feature level enables better representation learning than at the instance level for noisy medical imaging, in our case, fetal US video. Employing a more powerful feature extractor, such as the Swin Transformer, can further enhance feature learning. Additionally, integrating multiple pretext tasks can be beneficial, provided that the learning target is well-designed for the research question.

Furthermore, the results demonstrate that the proposed CRCR achieves state-of-the-art performance on both second-trimester standard plane detection and first-trimester anatomy recognition tasks, indicating that the quality of the learned representations has been improved during pretraining for both in-domain and cross-domain downstream tasks. This underscores both the generalisability and transferability of the learned representations.

Behold the vast Milky Way, shining and turning in the heavens. The king said, 'Oh! What crime have the people committed to deserve this?'

— The Book of Songs, Minor Odes of the Kingdom, The Milky Way.

5

Explainability of Self-Supervised Representation Learning for Fetal Ultrasound Video

Contents

5.1 Introduction

In the previous chapter, I investigated the performance of various self-supervised learning methods on fetal US video. Despite some of the methods demonstrating promising performance in different tasks, one of the limitations of end-to-end learning is that the process is not transparent and not well understood. Clinicians express concern over the opaque nature and desire for explanations to better integrate it into the clinical decision process. Therefore, in this chapter, my aim is to explore the question: *how can I reach a deeper understanding of self-supervised representation learning in the application of ultrasound video analysis by incorporating gaze-informed, clinically driven insights.*

As discussed in Chapter 2, SSL is a machine learning approach that generates an unlabelled data representation via pretext task training and transfers the learnt representation to make predictions on downstream tasks. At the time I started this work, SSL methods had started to be shown to be promising in the field of medical imaging analysis [92] to release the burden of manual data annotation, with the first implementation in fetal US video analysis commenced in 2020 by Jianbo Jiao [57], my co-supervisor. These studies suggest that SSL methods could learn useful representations from large unlabelled healthcare datasets and then utilise these learnt representations to make more complex clinical predictions, achieving performance levels comparable to, or even surpassing, those of clinicians in certain healthcare tasks [107]. Fig. 5.1 illustrates how SSL can be translated into clinical practice to support and assist the clinical decision-making process. These methods have the potential to reduce diagnostic and therapeutic errors that are often inherent in human practice [292], while also addressing critical challenges in the healthcare system—such as workforce shortages [293] and inequities in access to care [294].

As the potential of SSL models in clinical practice continues to grow, a significant gap remains in our ability to understand, interpret, and trust these models at

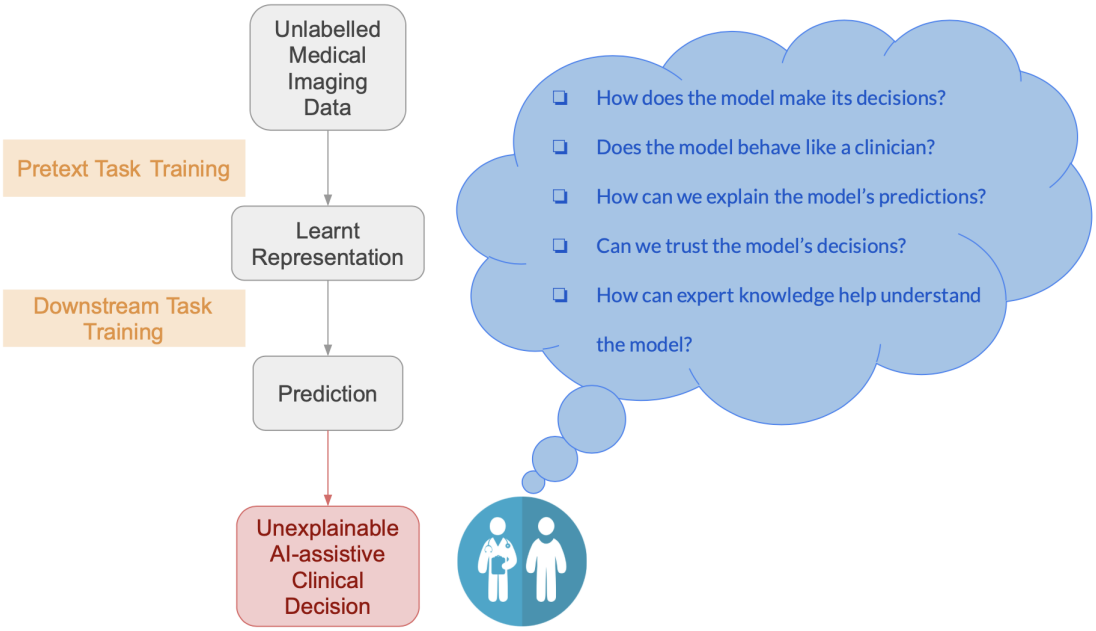


Figure 5.1: Illustration of SSL applied in clinical practice, showing interactions with patients and clinicians. Yellow arrows indicate the training process, with outputs used to assist clinical decisions without explanation. Common questions from patients and clinicians regarding unexplained decisions are listed on the right.

the same pace. A deep learning method is often regarded as a "black box", meaning the algorithm's internal mechanisms and decision-making processes remain fundamentally opaque to human understanding [295]. For the self-supervised representation learning model, the processes by which the model extracts features from raw data, transfers learned representations to different tasks, and utilises these transferred features for prediction remain largely unknown and unexplainable. Fig. 5.1 presents several common questions raised by patients and clinicians in response to unexplained model decisions, including:

- *How does the model make its decisions?* Specifically, what information is the model using to arrive at its predictions?

- *Does the model behave like a clinician?* In particular, is it focusing on the same image regions and patterns that experts rely on during real-time interpretation?

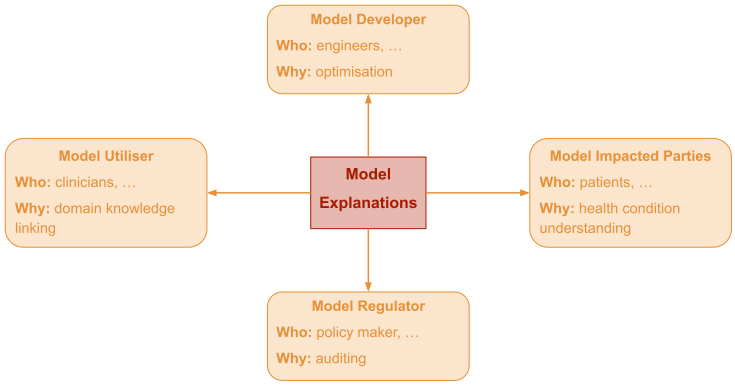


Figure 5.2: Target audiences for model explanations and their specific needs. This overview is partially inspired by [296].

- *How can we explain the model's predictions?* Can we make its reasoning understandable and clinically meaningful?
- *Can we trust the model?* Are its decisions based on clinically relevant information, and are the explanations reliable?
- *How can expert knowledge help understanding the model?* Can tools such as gaze tracking show whether the model aligns with expert visual attention?

In this chapter, I aim to address these key questions to better understand the model's behaviour and decision-making process, with the goal of building trust in the model's clinical applicability.

In safety-critical fields such as healthcare, decisions made with the assistance of black-box models can have significant consequences, potentially directly affecting human lives. Therefore, model explanation plays a crucial role in the advancement of deep learning in clinical settings, offering a rationale that helps target audiences understand the model's decisions and build trust in automated, learning-based systems. As shown in Fig. 5.2, in clinical practice, the target audience for model explanations typically includes four key groups: model users (clinicians), impacted parties (patients), model regulators (policy-makers), and model developers

(engineers) [296]. Each group has specific needs and expectations regarding the nature and purpose of model explanations. As the end-user of the model, clinicians need to establish trust in the model's use. Model explanations help bridge specific predictions with clinical domain knowledge, therefore validating the accuracy and relevance of those predictions. Since the model's output can directly influence patient health, providing clear and meaningful explanations is also essential for patients, enabling them to understand their condition and build confidence in the predicted outcomes. This "right to explanation" for decisions made by automated systems has been legally mandated under the European General Data Protection Regulation (GDPR) since 2018 [297]. Additionally, model developers can use explanations to better understand and optimise model behaviour, ultimately improving model performance and reliability.

In this study, the target audience for model explanations includes all key end-users involved in the use of deep learning-assisted fetal US scans: sonographers, expectant mothers, regulators, and engineers. My focus on explainability is driven by the need to provide each of these groups with meaningful, clinically grounded insights. Specifically, I generate explanations that assess how well the model's attention aligns with the gaze patterns of experienced sonographers. For sonographers—the primary users—trust in the model depends on its ability to mirror their clinical reasoning. I address this by evaluating the correspondence between the model's focus and expert gaze during scanning. Expectant mothers, whose care is directly influenced by these models, benefit from meaningful explanations that support their understanding and confidence in the diagnostic process. Regulators require evidence that AI tools are reliable, and clinically valid, which our anatomy-aware, gaze-informed metrics help to demonstrate. Finally, engineers benefit from explainable feedback to better understand and refine model behaviour. By grounding model explainability in clinician gaze data and anatomical relevance, my work aims to meet the diverse expectations of all these end-users, supporting more trustworthy and clinically integrated DL systems.

At the time of this work, most research in model explanations concerned understanding general deep learning models with a focus on model visualisation [162–164, 298, 299]. The visual-based explanations are generally coarse, especially when the information is diffuse. Quantitative metrics for explainability have been proposed [162, 299], but are still limited. The explainability of self-supervised representation learning models, on the other hand, was under-explored.

This chapter explores how self-supervised representation learning in fetal US video analysis can be better understood by incorporating clinically relevant information, specifically sonographers' gaze tracking data. I define explainability as the model's ability to capture anatomy-aware knowledge and propose quantitative metrics based on visually salient landmarks—regions that clinicians naturally focus on during scanning. These metrics tell, how well the model's learned features align with sonographers' gaze and anatomical relevant information by measuring the clustering quality of fine-tuned CNN features at these landmarks. Validated on mid-pregnancy fetal US videos, my approach offers valuable insights into model behaviour and supports more trustworthy, clinically aligned explanations of self-supervised learning.

5.2 Originality and Individual Role

I preprocessed and selected raw fetal ultrasound video clips used in this chapter with a Python-based library developed by Richard Droste, a previous D.Phil student at the Institute of Biomedical Engineering, University of Oxford. I re-implemented the saliency prediction model that was developed by Richard Droste. I designed, trained, and tested the proposed quantitative metrics using a Python-based open-source library PyTorch [273]. This work was accepted as extended abstract in the Medical Imaging Meets Neurips (MED-NEURIPS) 2021, NEURIPS workshop.

5.3 Background

5.3.1 Self-supervised Representation Learning

The general pipeline of self-supervised learning, typically consists of two steps: (1) train a CNN model with a pre-defined task on a large set of unlabelled data, and (2) fine-tune the pre-trained network for the high-level downstream task with a small set of annotated data. The core to the approach is the principle of providing "labels for free" from unlabelled data, which is advantageous as it makes deep learning no longer annotation dependent. There has been growing interest in self-supervised representation learning applications in the field of medical image analysis due to the relatively limited availability of medical image annotations. Therefore, it is crucial to select appropriate approaches when adapting and applying current methods that were originally developed for natural images to medical image analysis. Otherwise, it may result in visual representations that lack consistency in appearance and semantics.

5.3.2 Model Explainability

Model explainability refers to methods that help humans understand how a deep learning (DL) model makes decisions, and is often discussed alongside interpretability. While the terms are sometimes used interchangeably [300, 301], recent studies distinguish them more clearly [158, 302]. As defined by [158], interpretability typically refers to the transparency of a model at the task definition level, whereas explainability involves post hoc methods that aim to clarify the behaviour of black-box models. Montavon et al. [302] describe interpretation as "the mapping of an abstract concept (e.g., a predicted class) into a domain that the human can make sense of," and explanation as "the collection of features of the interpretable domain that have contributed to a given decision."

In my work, I refer explainability in the clinical setting to the model's ability to capture anatomy-aware knowledge. To assess this, I use sonographers' gaze-tracking data as a reference for clinically meaningful regions. My proposed quantitative

metrics measure how well the model's learned representations align with these gaze-informed landmarks, revealing whether the model's attention corresponds to expert reasoning. This approach provides a form of post hoc explainability by helping to interpret the features the model relies on, making its decisions more understandable and trustworthy in clinical ultrasound analysis.

5.3.3 Saliency Prediction

According to cognitive studies, the human optic nerve receives a vast amount of data that exceeds its processing capacity and only pays attention to the necessary information [303]. Hence, when interpreting an image, the human eye tends to focus on semantically informative regions [304], leading to its close relationship with image analysis and understanding. The attention of the eye across the image could be quantified through the distribution of gaze points, which could be recorded using gaze-tracking techniques. The study presented in this chapter is part of the PULSE project. A key novelty of this project is the simultaneous recording of both full-length ultrasound videos and the sonographer's eye movements while performing routine obstetric scans on pregnant women [5].

A machine learning-based model that imitate human visual attention is referred to as *saliency prediction model*. It predicts the probability that the human eyes would look at each pixel of a given image [305], which has been referred to as a *visual saliency map*. Huang et al. firstly introduce CNN as an effective saliency prediction model [306]. Since then, most state-of-art saliency prediction models have been developed based on CNN.

Saliency prediction is useful in ultrasound image analysis. [248, 253] model sonographer visual attention to assist in ultrasound image analysis tasks. In my work, I utilise saliency prediction in a different perspective to aid in the explainability of self-supervised representation learning for ultrasound.

5.3.4 Landmarks

An *anatomical landmark*, or *landmark*, is defined as "a point of correspondence on each object that matches between and within populations" [307] and "a biologically-meaningful point in an organism" [308]. The extraction and localisation of landmarks are critical in numerous medical image analysis tasks, aiding in visual navigation [309] and providing evidence for anomaly diagnosis [310]. Therefore, anatomical landmarks detection has become a fundamental task. Traditionally, human experts manually identify these landmarks, which is often tedious and cumbersome [311, 312]. To address this, machine learning-based models have been developed for fast and accurate automatic landmark localisation, offering a valuable alternative to manual identification, particularly when precise localisation of multiple landmarks is required [313]. These methods could be either classification-based [314, 315], detecting the presence of landmarks in image voxels, or regression-based [316, 317], predicting heatmaps that indicate landmark locations.

5.3.5 Clustering

Clustering is an unsupervised machine learning technique that assigns a set of data points into self-similar groups, which are called *clusters*, so that the data points in the same cluster are more similar to each other than those in other clusters [318]. Traditionally, clustering is performed on observed data points using various methods. These include hierarchical clustering (i.e. BIRCH [319]), centriod-based clustering (i.e. K-means [320]), density based clustering (GMM [321]) and density-based clustering (i.e. DBSCAN [322]). He et al. suggests that a non-linear transformation could simplify the data structure, making it more linear separable in the transformed feature space [323]. Clustering performed in this transformed feature space is referred to as *feature clustering*. A CNN model, which operates as a non-linear model, is designed to learn meaningful feature representations, ensuring that meaningful information in the data is preserved in the feature space. Consequently, it simplifies the associated data structure and preserves the inherent data groupings. Clustering

also plays a key role in XAI, by offering insights of relationships between data and simplify complex data to uncover patterns for explainability.

5.4 Methods

An overview of the proposed explainability method for self-supervised learning framework is presented in Fig. 5.3. As shown in the figure, our method consists of six steps:

1. Pre-training a CNN model $f(\theta)$ that transferable to downstream tasks with a selected self-supervised pretext task on unlabelled ultrasound video clips;

2. Supervised transfer learning with end-to-end fine-tuning on a selected clinical downstream tasks with a small amount of labelled ultrasound video frames;

3. Saliency map prediction on selected ultrasound video frames using a Saliency-VAM model [248] with automatically recorded eye-gaze tracking data and extracting visual salient landmarks;

4. Extracting features that correspond to visually salient landmarks from the penultimate layer of the fine-tuned CNN model $g(\theta)$;

5. Clustering the extracted CNN feature vectors across the input ultrasound video frames, with each cluster represents a landmark;

6. Evaluating the quality of clustering as the proposed explainability metrics.

In this section, I describe our proposed explanability method in general term. Let $\chi \subset \mathbb{R}^{1\times H\times W}$ be the set of US video frames with height H, width W and monotone color, let $\mathcal{V} \subset \mathbb{R}^{1\times H\times W\times K}$ be the set of US video clips with number of frames K in each video clips $v = \{x_i | x_i \in \chi\}_{i=1}^{K}$ and let $\mathcal{G} = [0, W] \times [0, H]$ be the set of valid gaze points.

Dataset $V = \{v_i | v_i \in \mathcal{V}\}_{i=1}^{N_v}$ consists of N_v raw US video clips of length N_f is available for self-supervised pretext training. Dataset $X = \{x_i | x_i \in \chi\}_{i=1}^{N_d}$ consists of N_d labelled US video frames is available for supervised downstream training. Suppose each video frame x_i has a corresponding gaze point set $G_i = \{g_j | g_j \in \mathcal{G}\}_{j=1}^{N_g}$,

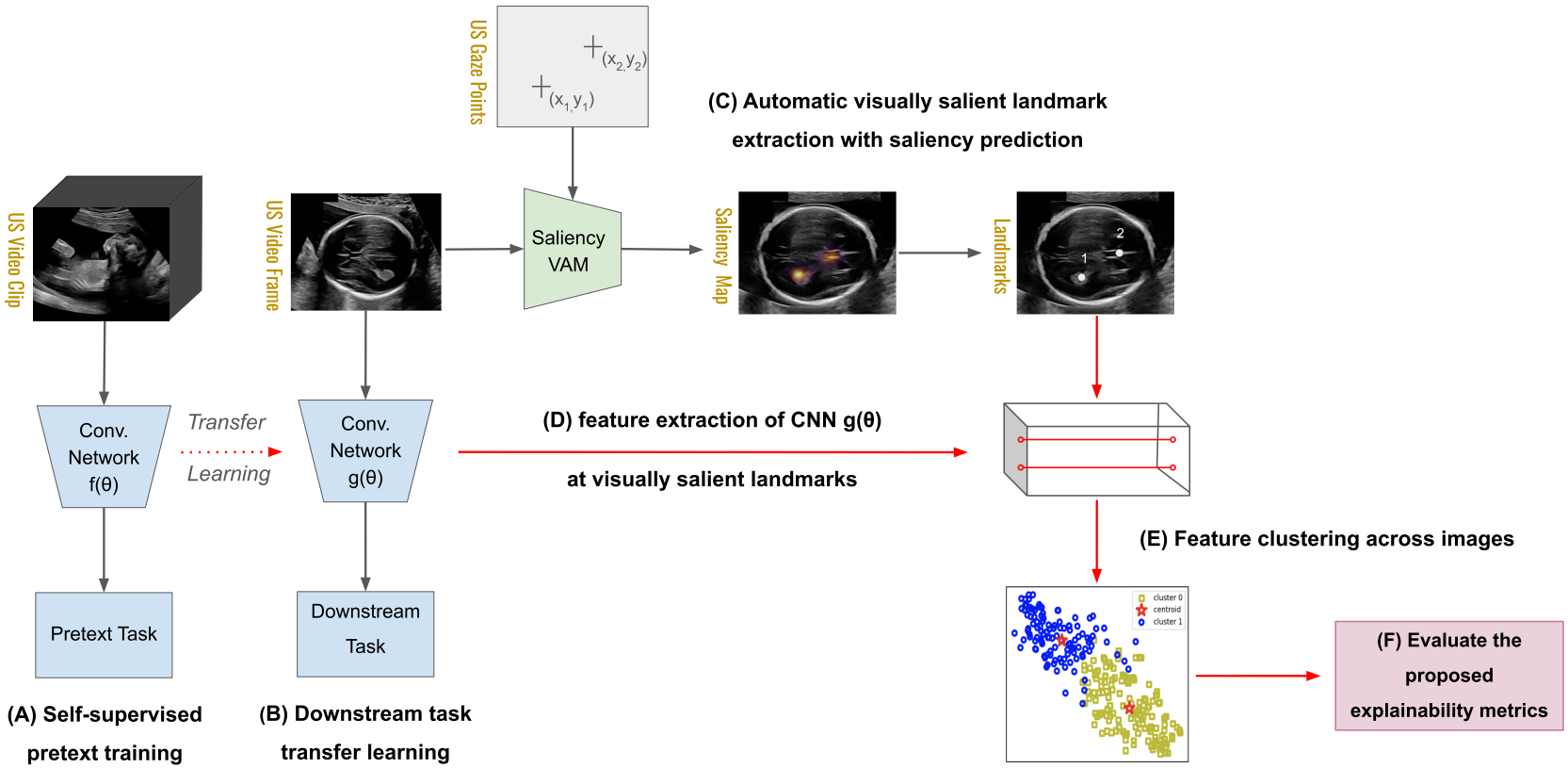


Figure 5.3: Proposed explainability method framework for self-supervised learning on medical ultrasound video. In step A, ultrasound video clips are used as input for self-supervised representation learning. Ultrasound video frames are used as input for a downstream task in step B and automatic visually salient landmarks extraction in step C together with ultrasound gaze points. In step D, features at these visually salient landmarks are extracted, followed by clustering across images in step E. Finally, step F measures the quality of feature clustering as for explainability metrics.

with $N_g \geq 1$. Dataset $D = \{(x_i, G_i)\}_{i=1}^{N_d}$ consists of N_d pairs of video frames and gaze point sets is available for further explanability analysis on downstream task.

5.4.1 Self-supervised Pretext Training

Firstly, I pre-train self-supervised pretext tasks on raw ultrasound video clips. The following three self-supervised pretext tasks are selected with the reason for selection with the specific US properties.

- **Temporal sequence sorting [258]**: US is a real-time operator dependent imaging modalities, where the temporal coherence is crucial as the probe continuously capture frames information in a sequential manner. Leveraging this temporal consistency allows the model to better understand frame order and motion patterns, further encouraging the learning of spatio-temporal representations, which are important for recognising anatomical structures in dynamic scans.

- **Rotation prediction [263]**: Unlike the natural image, US image lack a canonical orientation, as the appearance of anatomical structures is directly influenced by the probe's rotation. Clinicians rely on rotation consistency to accurately interpret anatomy. Therefore, it is important for the model to learn rotation-aware features to enhance its spatial understanding and improve anatomical recognition.

- **Pace prediction [324]**: In fetal US scans, the frame acquisition rate is not uniform; it varies depending on probe speed, anatomical complexity, and user interaction. The proposed method encourages the model to understand temporal dynamics and motion variations, which are the key factors in real-time analysis. Therefore, enables more effective motion-aware feature learning.

These methods encompass both spatial (rotation) and temporal (sequence and pace) transformations that have proved to be effective in other medical image modalities. I assume that if a model can successfully complete one of the selected

pretext tasks, it should be capable of capturing underlying anatomical-aware information from a specific perspective. Hence, it is valuable to explain and compare these methods as they offer distinct insights.

Furthermore, some of the other pretext tasks mentioned in the literature review are not applicable here due to certain unique characteristics of fetal ultrasound. For instance, colourisation-based methods are not suitable because of the monochrome nature of ultrasound images, and tracking-based methods face challenges due to difficulties in visual tracking of ultrasound scans under B mode (e.g., complex and variable natural of US image). At the time this research was conducted, SimCLR [58] had only recently been published and its strong performance in medical imaging analysis had not yet been widely demonstrated. As a result, contrastive learning methods were not included in this study.

Here, I apply three selected methods to a set of raw fetal ultrasound video clips V with clip length $N_f = 6$. Fig. 5.3 briefly illustrates the algorithms of three methods, which are more explicitly discussed next.

Temporal Sequence Sorting [258]

The raw ultrasound video clip undergoes pre-processing through random shuffling to form an input for the 3D CNN model $f(\theta)$. Subsequently, the model is trained to arrange the shuffled sequence in ascending order. Since video clips of length 6 are utilized, there are a total of 720 ways to permute the clip. To maximize the deviations between permutations and control the task's difficulty and efficiency, only permutations exhibiting the largest Hamming distance with the original order are considered during pre-processing. [325] defines the Hamming distance between two strings or vectors of equal length as the count of positions where the corresponding symbols differ. When considering permutation of a set of length 6, the maximum Hamming distance is 5, attained by the following permutations: $(2,1,4,3,6,5), (2,3,5,6,1,4), (2,4,6,5,3,1), (3,4,1,6,2,5), (3,5,2,1,6,4), (3,6,4,5,1,2)$, $(4,5,6,2,1,3), (4,6,5,3,2,1), (5,1,2,6,4,3), (5,6,1,2,3,4), (6,5,1,3,4,2)$. Therefore, the temporal sequence sorting task has been formulated as an 11-class classification problem.

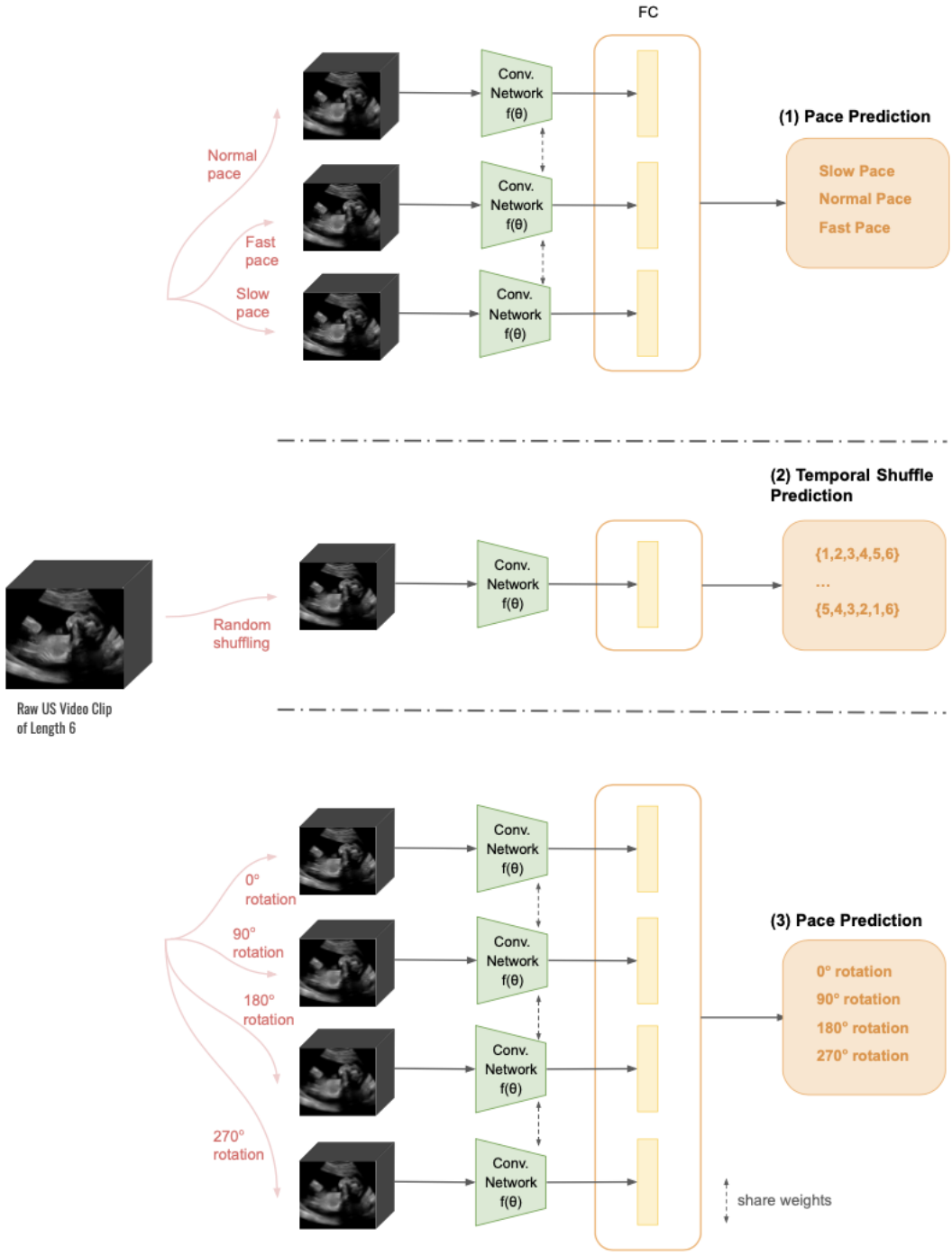


Figure 5.4: Illustration of three selected self-supervised video representation learning methods on ultrasound scan video clip.

Rotation Prediction [263]

A set of rotations $\{R^1, R^2, .., R^M\}$ are applied to a video clip and the resulting transformed clips are utilized as inputs for 3D CNN model $f(\theta)$. The task is formulated to identify the specific types of rotations that have been applied to each of the input clips, which is a M class classification. Intuitively, if a model could accomplish this task, it should be aware of both semantic representations and also geometric transformations of the video clip. Here, I consider four degrees of in plane rotation $\{0^o, 90^o, 180^o, 270^o\}$ to apply to ultrasound video clips.

Pace Prediction [324]

3D CNN model $f(\theta)$ takes inputs of video clips sampled at different paces and trains to identify the corresponding pace category. Three pace candidates {slow, normal, fast} are considered with the corresponding down-sample rates of {2, 4, 8}. The assumption is that if a model can learn the task, it should be able to understand the underlying video content and learn the spatio-temporal features. Each input clip has six frames and the starting frame is randomly sampled with the prerequisite that the remaining frames are longer than the desired training clips.

5.4.2 Downstream Task Fine-tuning

Secondly, I fine-tune the pre-trained network $f(\theta)$ for the clinical downstream task with a small set of annotated ultrasound video frames. In this work, I focus on a clinical classification task. During fine-tuning, the weights of the convolutional layer are retained, and a classifier head with randomly initialized weights is attached for classification training. The resulting 3D CNN model, denoted as $g(\theta)$, is employed to carry out the downstream task and assess the meaningfulness and transferability of the pre-trained representations

5.4.3 Visually Salient Landmark Extraction

Thirdly, I predict the saliency map and extract the visually salient landmarks.

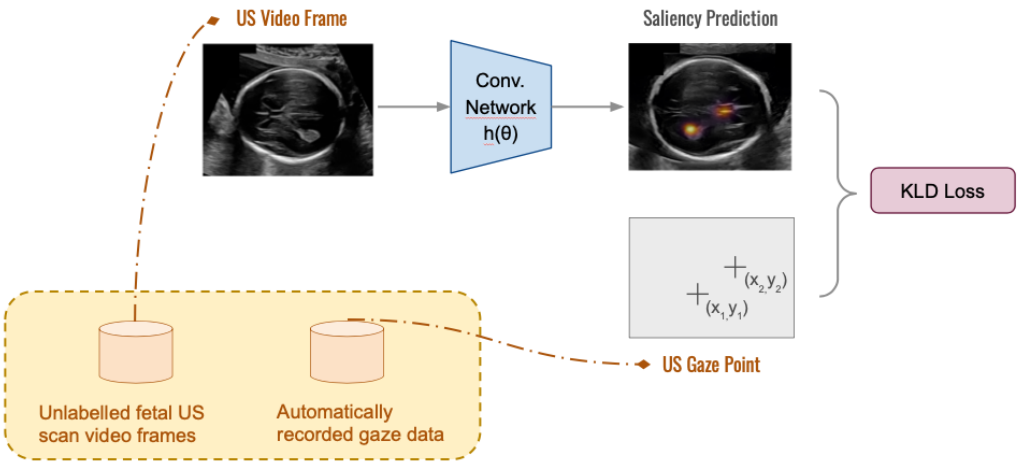


Figure 5.5: Pipeline of saliency map prediction using the Saliency-VAM model.

Saliency Map Prediction

As mentioned in Sec. 5.3.3, human tend to look at semantically informative regions to understand an image. In the context of ultrasound scanning, sonographers tend to focus on anatomically informative regions to understand the image.

The Saliency-VAM model [248] is an original framework proposed by Richard Droste, a previous member of the lab, which models sonographer visual attention on static fetal ultrasound video frames through visual saliency prediction. Fig. 5.5 illustrates the pipeline of the Saliency-VAM model, where a CNN is trained to predict gaze point distribution (saliency map) of sonographers observing the image from the domain of interest. This model is based on unique datasets from the PULSE project, where gaze-tracking data is automatically recorded during sonographers' routine clinical fetal anomaly screenings. The model takes the unlabelled fetal US scan video frames as input and predicts visual saliency to learn the attention patterns corresponding to each frame. It is trained using a KLD loss, computed between the predicted saliency maps and gaze points automatically recorded during ultrasound scanning.

Given an ultrasound video frame and its corresponding gaze point set $\{(x_i, G_i)\} \in D$, a visual saliency map $S_i \in [0,1]^{H \times W}$ is generated, where $S_{i,j}$ denotes the probability that pixel $x_{i,j}$ is fixed upon. $\S_i$ is created as a sum of Gaussian around

the gaze points in G_i, normalized so that $\sum_{i,j} = 1$. By down-scaling visual saliency map S_i to the size of $\hat{S}_i \in [0,1]^{H_d \times W_d}$, it is used as a target to predict the saliency map $\hat{S}_i$. The training loss is computed using the Killback-Leibler divergence (KLD) between the predicted saliency map and the down-scaled probability map.

In my work, I choose to use the Saliency-VAM model for saliency map prediction for two primary reasons. Firstly, it operates in a self-supervised manner, relying solely on unlabelled fetal US video frames and automatically recorded gaze data, without the need for manual annotations. This aligns with my objective of explaining self-supervised learning models while avoiding the introduction of additional annotation procedures during the explanation stage. Secondly, the model has already demonstrated strong performance in visual saliency prediction on the same fetal US dataset used in my study, making it a ready-to-use solution that requires no further adaptation.

Visually Salient Landmark Extraction

Here I use the learnt visual saliency to discover landmarks that attracts human visual attention, which are termed as *visually salient landmarks*. As the defined landmarks are detected by predicting the sonographer's gaze directly from saliency measures, they are practically significant and salient to operators.

The saliency-VAM model takes ultrasound video frame x_i of dimension 288×244 as input and outputs saliency maps $\hat{S}_i$ of dimensions $W_d \times H_d = 36 \times 28$ with three two-fold down-sampling operations. The highest visual attentions are located on the local maxima $max_{i,j}$ of the generated saliency map $\hat{S}_i$ and extracted by *peak_local_max* algorithm. Here, j represents the number of local maxima. Each video frame x_i has a corresponding set of visual salient landmarks $L_i = \{l_{i,j}\}_{j=1}^{N_l}$, with $N_l \geq 1$. Visual salient landmarks $l_{ij} \in [1, W_d] \times [1, H_d]$ are obtained by fitting a $2D$ Gaussian peak to a 3×3 neighbour around the local maxima $max_{i,j}$, and are proved to correspond to key anatomical structures in US scans [326]. Fig. 5.6 shows examples of predicted saliency maps and the discovered visually salient

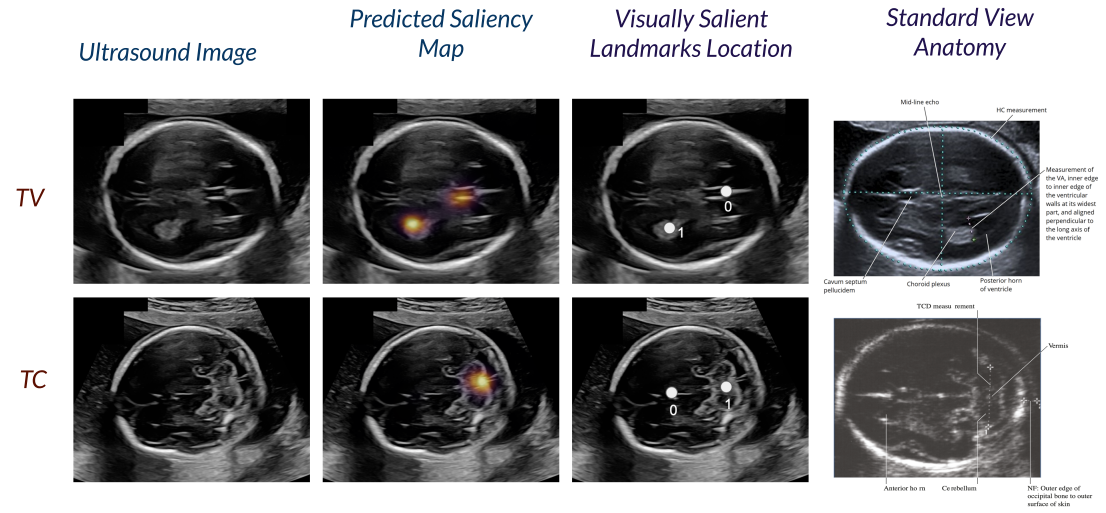


Figure 5.6: Examples of predicted saliency map (second column) and visually salient landmarks (third column) on ultrasound standard plane TV and TC. The predicted saliency map of each example has two peaks and two generated visually salient landmarks, which are labelled as 0 and 1. The first column shows a selected image of each standard plane and the last column illustrates the anatomy of each standard view.

landmarks from two selected ultrasound standard planes, the transventricular (TV) and the transcerebellar (TC) plane.

5.4.4 Self-supervised Feature Extraction

Fourthly, I extract CNN features from the fine-tuned model $g(\theta)$ at the location of visually salient landmarks, which process has been demonstrated in Fig. 5.7.

With the obtained set of visual salient landmarks L_i for the US frame x_i, I extract the corresponding feature set F_i from the penultimate layer of the CNN model $g(\theta)$. This is done using the function $f : [1, W_d] \times [1, H_d] \rightarrow \mathcal{R}^{N_f}$, which maps location of each visual salient landmark on the saliency map to the corresponding feature activation of the penultimate CNN layer, where N_f is the number of channels.

Given that the location of visually salient landmarks correspond to the areas of maximum visual attention of sonographers, it is reasonable to hypothesize that if the model $g(\theta)$ successfully initiates sonographers' performance on clinical

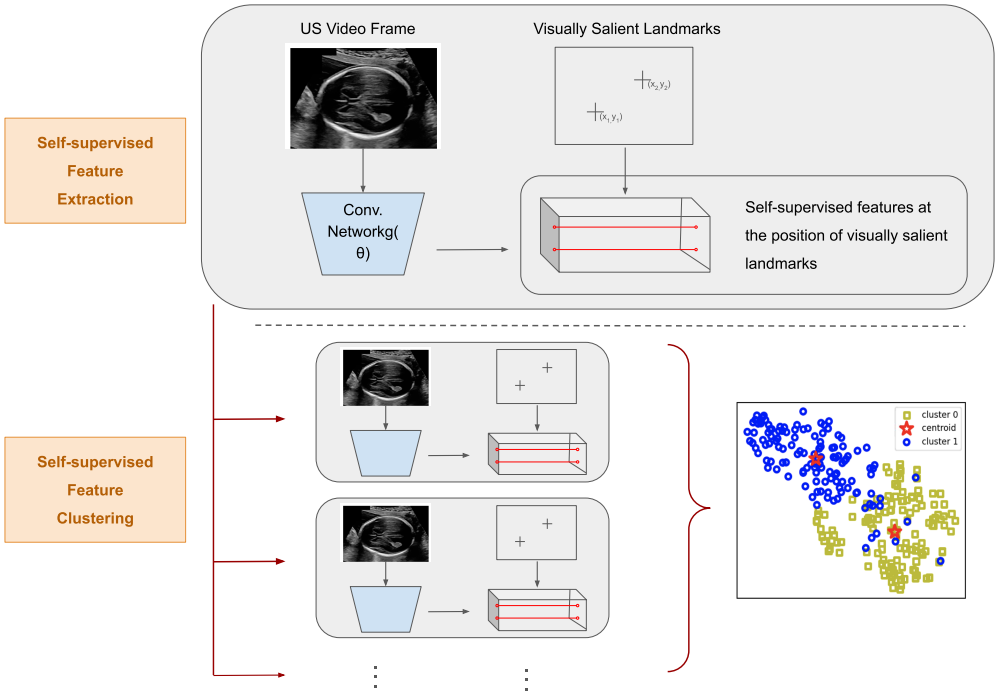


Figure 5.7: Pipeline of self-supervised CNN feature extraction (after fine-tune) at visual saliency landmarks and self-supervised CNN clustering across images.

downstream tasks, the CNN features at these salient landmarks should contain meaningful anatomical information.

5.4.5 Self-supervised Feature Clustering

Fifthly, I cluster the set of extracted feature vectors at the locations of visual salient landmarks across US video frames as depicted in in Fig. 5.7.

We consider a set of US video frames $X_c = \{x_i | x_i \in \chi\}_{i=1}^{N_c} \subset X$, which is a subset of labelled frames available for supervised downstream training. The set of all landmarks features vectors across these N_c number of US video frames is obtained as $\mathcal{F} = \bigcup_{i=1}^{N_i} F_i$. I use K-means clustering [320] to classify $\mathcal{F}$ into K number of clusters $\{C_1, C_2, ..., C_K\}$, where the number of cluster K is defined as the number of landmark labels contained in the dataset X_c. K-means, as a centroid-based clustering algorithm, partitions the dataset based on the closeness of data points to the centroid of each cluster. I choose it due to its efficiency, effectiveness, and simplicity. Here, the Euclidean norm is used as a distance metric.

5.4.6 Explainability Metric

Finally, I quantitatively measure explainability using a proposed set of metrics based on the clustering quality of self-supervised CNN features at visually salient landmarks across different US video frames.

As discussed in earlier sections, visually salient landmarks refer to locations that attract sonographers' visual attention during real-time US scanning and interpretation. In clinical practice, sonographers tend to focus on anatomically informative regions to interpret the image. Based on this, I hypothesise that these gaze locations correspond to regions containing key anatomical information.

Since self-supervised models are designed to closely mimic how a human makes decisions from US images, I hypothesise that an effective model should align with the reasoning of sonographers. Therefore, the focus of the learned representations should correspond closely with gaze distribution. The fine-tuned CNN model, denoted as (g_θ), should therefore be capable of capturing meaningful anatomical features at visually salient landmarks, indicating that the model's attention aligns with the clinician's gaze focus. Moreover, the features extracted from different landmarks should corresponding to their underlying anatomical relevant information.

Accordingly, I define explainability as the model's ability to capture anatomy-aware knowledge. I assess this through the quality of clustering of the learned features at visually salient landmarks. The proposed metrics offer a quantitative evaluation of explainability by measuring how well the model's focus aligns with sonographers' gaze in terms of capturing anatomically meaningful information.

Here, I use three quantitative metrics to measure the clustering quality: the Silhouette Coefficient [327], together with two additional new measures - cluster Compactness and cluster Uniqueness.

Silhouette coefficient (S) aims to explain the consistency of clustering and quantifies how similar a data point is to its cluster compared with the other

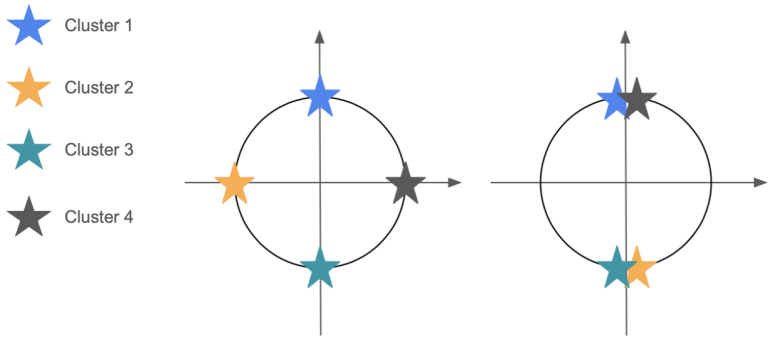


Figure 5.8: Example when two clusters share the same distance deviation, but different intra-class deviation. I assume that each star represents a cluster, with one data point per cluster located at the intersection of the circle with a radius of 1 and the axis.

clusters [327]. S is defined as

$$S = \frac{1}{n}\sum_{i=1}^{n}\frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}}, \tag{5.1}$$

where $a(x_i)$ represents the average distance between x_i and $x_{j,j\neq i}$ that are in the same cluster, $b(x_i) = \min_{k\neq l}\frac{1}{|C_k|}\sum_{x_o\in C_k}||x_i, x_o||_2^2$, represents the minimum average distance from each data point to the data points in another cluster, and n is the number of feature vectors in the set F. S is always between -1 and 1. $S = 1$ indicates a well-separated clustering.

Compactness (D_C) quantifies the inter-class deviation, which measures how similar a data is to its own cluster. It is defined as the weighted average of the square distance to the centroid. Here, τ is a scaling factor, $\hat{\mu_k}$ is the centroid of k-th cluster C_k. D_C is defined as

$$D_C = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k}\frac{||x_i - \hat{\mu_k}||_2^2}{\tau} \tag{5.2}$$

where τ denotes the temperature parameter.

Uniqueness (D_U) quantifies the intra-class deviation, which measures how separable the clusters are. Here, I define it to include both distance deviation and angular deviation. A classical measure of distance deviation is the weighted

sum of squared distance between global average of the dataset and each centroid $\frac{1}{n}\sum_{k=1}^{K}|C_k| \times ||\hat{\boldsymbol{\mu}}_{\mathbf{k}} - \bar{\mathbf{x}}||_2^2$. Fig. 5.8 illustrates an example where two clusters own the same distance deviation. I assume that each star represents a cluster, with one data point per cluster located at the intersection of the circle with a radius of 1 and the axis. Both datasets have a data mean of (0,0), thus both clusters have the same distance separation of 2. However, observe that the clustering on the left appears to be better separated than the one on the right, which indicates different intra-class deviations. Hence, distance deviation alone cannot fully represent the intra-class deviation; angular deviation also plays a crucial role. Here, I use the cosine similarity between centroid and data mean as a measure of deviation measure. Since cosine similarity measures the closeness of two clusters in distance space and distance measure tells their separation in angular space, I introduce an additional exponential term and assign opposite signs to each term to leverage their opposite trends. Here, $|C_k|$ denotes the cardinality of C_k. D_U is defined as

$$D_U = \frac{1}{n}\sum_{k=1}^{K}|C_k| \times \frac{\exp\{||\hat{\boldsymbol{\mu}}_{\mathbf{k}} - \bar{\mathbf{x}}||_2^2 - \frac{(\hat{\boldsymbol{\mu}}_{\mathbf{k}}-\bar{\mathbf{x}})\cdot(\hat{\boldsymbol{\mu}}_{\mathbf{1}}-\bar{\mathbf{x}})}{||\hat{\boldsymbol{\mu}}_{\mathbf{k}}-\bar{\mathbf{x}}||||\hat{\boldsymbol{\mu}}_{\mathbf{1}}-\bar{\mathbf{x}}||}\}}{\tau} \tag{5.3}$$

where τ represents the temperature parameter, which is the same as in D_c.

5.5 Experiments

5.5.1 Data

Our experiments are based on the large-scale fetal US video dataset, named PULSE, as described in Chapter 3. Dataset II, discussed in 3.3, consists of a large number of unlabelled US video clips and is used as the pre-training dataset. Dataset IV, discussed in 3.4.2, consists of labelled second-trimester scan frames and is used for three-fold cross-validation on the downstream task—standard plane detection. All frames were centrally cropped to remove the fan shape and resized to 288×224 pixels with further data augmentation including horizontal flipping, brightness variation.

5.5.2 Model Implementation and Training

Self-supervised Pre-training Stage

We use a 3D ResNet-18 [287] as the backbone network for self-supervised representation learning. The models were trained with an SGD optimizer with momentum of 0.9 and weight decay 1×10^{-4}, which is aligns with [57]. The learning rate was set to be 0.001 initially by grid search and is decayed by 0.1 after every 24,000 iterations. The models were trained for 50 epochs, with batch size of 32. The models were implemented with the PyTorch framework. Each training run took 13-18 hours on a single NVIDIA GTX 1080 Ti.

Supervised Fine-tuning Stage

In terms of the clinical downstream task, during the fine-tuning stage, the weights of the convolutional blocks are preserved from the self-supervised representation learning networks, while the weights of the fully-connected layers are initialized randomly. Subsequently, the network is trained in a supervised manner using cross-entropy loss. The data pre-processing and training strategy remains consistent with the self-supervised pre-training stage, with the exception of setting the initial learning rate to 0.01. The parameters have been fine-tuned by previous lab members for this specific tasks.

5.5.3 Evaluation Metrics

- **Model Explainability** Here, both quantitative and qualitative methods are employed to provide explanations for the SSL model in fetal US analysis, offering a more comprehensive understanding from multiple perspectives. Qualitative explanations, such as highlighted regions, are intuitive and visually interpretable, while quantitative explanations offer objective, comparable metrics that facilitate straightforward comparison between different models.
  - **Quantitive Explainbility**: The proposed quantitative metrics are used to measure model explainability, incorporating clinical domain knowledge into the explanation process. Specifically, the quality of feature clustering

is used to assess the model's ability to capture anatomy-aware knowledge, providing a quantitative measure of explainability.

- **Qualitative Explainbility**: Visualising CNN representations is a direct way to explore hidden patterns inside ML algorithms. Here, two types of visulisation methods are selected to provide qualitative explanations for fetal US analysis. Firstly, I visualise the CNN decisions using guided backpropagation [163]. Secondly, I plot the corresponding activation-based spatial attention maps, which are considered as the sum of absolute values of each feature map here [328].

- **Model Performance** To evaluate the model performance on downstream tasks, precision, recall, and F1 scores are reported. Precision measures how many positive predictions are correct, recall measures how many positive cases have been correctly predicted and the F1 score measures the harmonic mean of the precision and recall. Detailed formulas and discussion could be found in Chapter 3.

5.6 Results

5.6.1 Quantitative Evaluation of Model Explainability

Quantitative evaluation of the proposed explainability metrics has been conducted across selected self-supervised pretext tasks, as presented in Table 5.1. Both lower and upper baselines are included, represented by random initialisation and a supervised model, respectively. As shown in the table, the upper and lower baselines show the best and worst clustering quality, as expected. In the case of random initialisation, the model has not been exposed to any anatomical information during training, and thus performs poorly in capturing anatomy-aware knowledge. In contrast, the supervised model is trained on a large number of labelled ultrasound frames and is therefore expected to demonstrate stronger clinical reasoning and better feature clustering. Among self-supervised learning methods, the sequence sorting method has the best clustering quality in terms

Table 5.1: Quantitative evaluation of the proposed explainability metrics for selected self-supervised pretext tasks, including lower and upper baselines.

	Rand.Init.	Pace Pred.	Rot.Pred.	Seq.Sort	Supervised
Uniqueness ↑	1.132	2.677	2.587	**3.512**	4.032
Compactness ↓	1.805	1.596	1.378	**1.375**	1.358
Silhouette ↑	0.164	0.237	**0.244**	0.243	0.278

of uniqueness and compactness, while the rotation prediction method has the best with Silhouette coefficient.

For the three selected pretext tasks, the models are trained to learn representations from different perspectives. The results indicate that, among the tested self-supervised methods, rotation prediction and sequence sorting achieved higher clustering performance, suggesting superior explainability—specifically, a stronger ability to capture anatomy-aware knowledge and better alignment between the model's reasoning and the gaze focus of sonographers. This implies that the attention pattern of these models during prediction more closely resemble the visual focus of clinicians during scanning and interpretation. The stronger clustering performance of rotation prediction and sequence sorting tasks can be explained by how well these tasks match the nature of US imaging. Sequence sorting benefits from both spatial and temporal information, which is especially useful in US videos where frames are captured in a smooth and continuous order as the probe moves. This natural flow helps the model learn meaningful anatomical changes over time, similar to how sonographers interpret scans. Rotation prediction is also effective because US images do not have a fixed orientation. Rotating an image can significantly change how anatomy appears, so training the model to predict rotation encourages it to focus on important visual features. On the other hand, pace prediction is less helpful. The scanning speed in US is often irregular and varies depending on the operator. Sudden stops or changes in speed may not reflect any anatomical structure, making this task harder to learn from. Overall, rotation prediction and

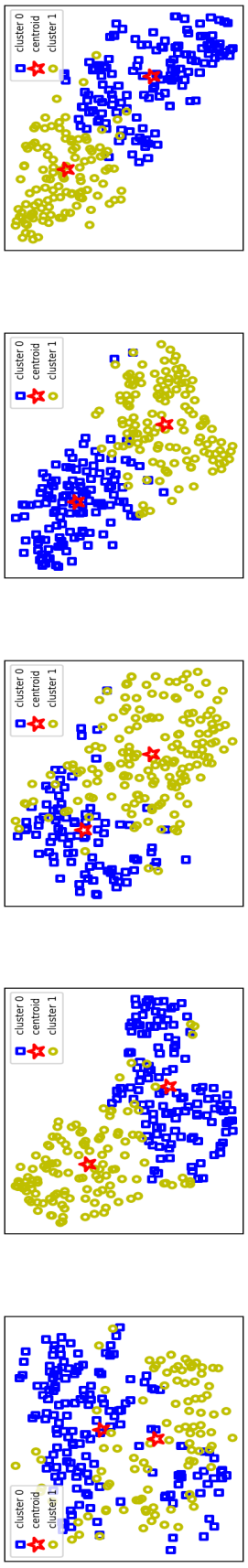
sequence sorting help the model focus on anatomy in a way that better matches how clinicians view and interpret US scans.

Fig. 5.9 presents the t-SNE plots of the clustering results. A clear separation between clusters emerges progressively from (a) to (e), with the cluster centroids becoming further apart. This provides a visualisation that aligns well with the quantitative results discussed above. These findings suggest that, based on our proposed explainability metrics, either of the top-performing methods would be preferable for the selected ultrasound analysis task to achieve improved explainability.

5.6.2 Qualitative Evaluation of Model Explainability

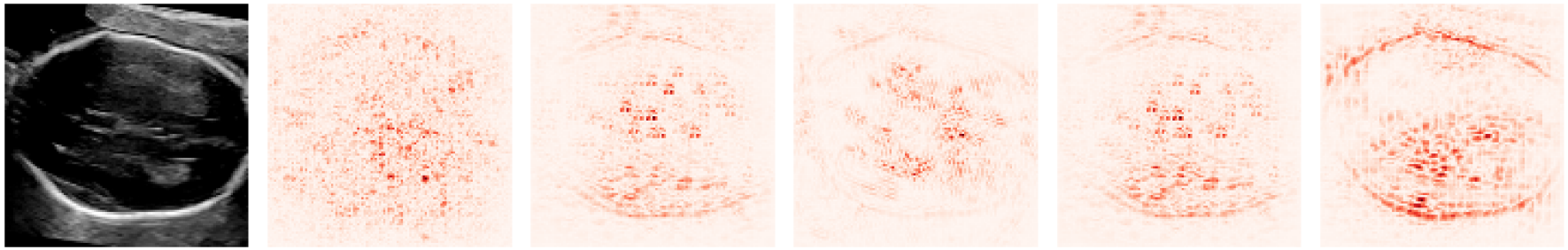
Fig. 5.10 shows heatmaps generated using guided backpropagation, which indicates which part of the image is used to make a decision. Taking the fetal head as an example, the visualisations demonstrate that both supervised and self-supervised methods are able to capture key anatomical structures, such as the outer edge of the skull. In contrast, the model with random initialisation exhibits a uniformly low response across the image, indicating a lack of focus on anatomically relevant regions. Fig. 5.11 presents the mid-level activation-based spatial attention maps for the spine, illustrating the regions where the network primarily focuses when performing the task. The spatial attention responses show a progressively stronger correlation with anatomical structures, transitioning from random initialisation to self-supervised learning, and further to supervised learning. These visual-based explanations are generally consistent with the results of our proposed explainability metric.

However, as discussed earlier, visual-based explanations are often coarse and limited in their ability to capture subtle differences. Similar visual outputs can be produced across different models, making comparisons difficult, as minor variations may be easily overlooked by the human eye. This limitation could be found in the two examples presented above, where it is challenging to distinguish between methods based solely on visual inspection. This highlights a major drawback of qualitative explainability approaches and emphasise the motivation behind our proposed quantitative evaluation framework. Quantitative metrics provide precise,



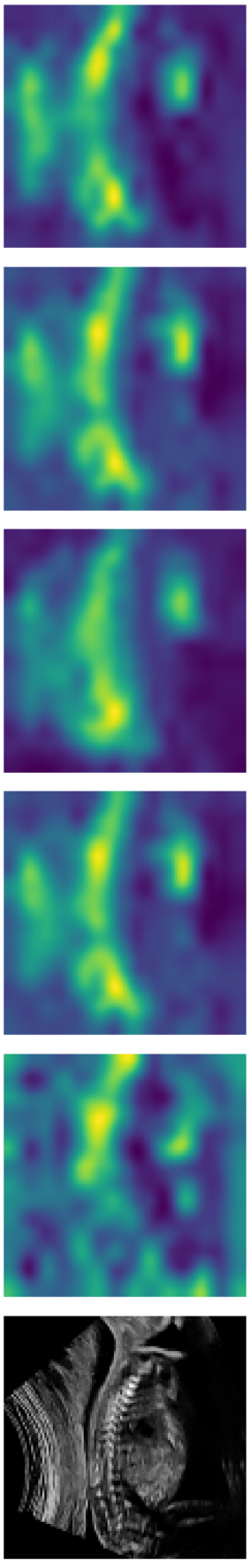
(a) Rand. Init. (b) Pace Pred. (c) Rot.Pred. (d) Seq.Sort (e) Supervised

Figure 5.9: t-SNE plots of feature clustering based on the proposed explainability metrics. Cluster centroids are indicated by red stars.



(a) Image **(b)** Rand. Init. **(c)** Pace Pred. **(d)** Rot.Pred. **(e)** Seq.Sort **(f)** Supervised

Figure 5.10: Visualization of downstream tasks using guided-backpropagation. Dark colors mean high-response regions.



(a) Image (b) Rand. Init. (c) Pace Pred. (d) Rot.Pred. (e) Seq.Sort (f) Supervised

Figure 5.11: Attention visualization on the last convolutional layer.

Table 5.2: Evaluation results on standard plane detection task.

	Rand.Init.	Pace Pred.	Rot.Pred.	Seq.Sort	Supervised
Precision	0.625	0.673	0.667	**0.679**	0.681
Recall	0.580	0.584	**0.637**	0.622	0.627
F1-score	0.596	0.614	**0.644**	**0.644**	0.646

objective measurements that can capture subtle variations in model behaviour and allow for straightforward comparisons across different methods.

5.6.3 Evaluation on Downstream Task Performance

A second-trimester standard plane detection task is used to evaluate downstream performance and further validate the effectiveness of our proposed explainability metrics. Evaluation results are shown in Table 5.2. The results demonstrate that, the self-supervised learning models outperform random initialisation, while the supervised model performs the best. The sequence sorting and rotation prediction methods have better results than the pace prediction method. The pace prediction method is designed to learn representations through understanding video content. I suspect the ultrasound video, describes the motion of the probe might not be appropriate for the design (i.e. pause during biometric measurement). Better performance on the downstream task suggests the model learns better anatomical meaningful representations compared to the others, which corresponds to the high clustering quality and better anatomic-aware knowledge capture in Sec. 5.6.1. Hence, the results demonstrate the effectiveness of our proposed metrics, and the recommended method choice based on the metrics is valid.

5.7 Conclusion

In this chapter, I addressed the challenge of explainability in self-supervised learning for fetal US analysis. Explainability is defined within the clinical context as the ability to capture anatomy-aware knowledge, with clinical domain expertise incorporated directly into the explanation process. Rather than treating explainability

as generic feature visualisation, I adopt a clinician-driven approach by integrating visually salient landmarks—regions that naturally attract sonographers' attention during ultrasound interpretation. I proposed a novel set of quantitative metrics to assess model explainability, evaluating how well the learned representations align with gaze-tracking data and how effectively the features at these landmarks are clustered and correspond to anatomically relevant information. This ensures that the model's reasoning aligns with sonographers' visual attention during real-time US interpretation, offering a clinically meaningful explanation framework. The proposed methods are evaluated on fetal US scan undertaken within mid-pregnancy. Experimental results demonstrate that the proposed quantitative metrics effectively measure explainability across selected methods with different pretext tasks. Compared to conventional visual explanations, these metrics provide a more straightforward, precise, and objective assessment, while also aligning well with downstream task performance.

The valley wind blows gently, bringing clouds and rain. Let us strive together with united hearts, there should be no anger between us.

— *The Book of Songs, The Odes of Bei, The Valley Wind.*

6

Out-of-Clinical-Distribution Detection with a Softmax-Conditioned Variational Autoencoder Boosted Model

Contents

6.1 Introduction

Deep learning models, including self-supervised learning models, have demonstrated effectiveness with state-of-art performance [329, 330]. In Chapter 4, I demonstrate its effectiveness in medical imaging analysis, specifically in US video analysis. However, such models are always "black boxes" with obscure decision-making algorithms and minimal human intervention. This can be problematic in safety-critical fields like healthcare. For this reason, it is essential to make machine learning trustworthy and responsible for users [331]. In Chapter 5, I explore this question within the subfield of explainable DL. I investigate the explainability of the model to advance understanding of model decisions and build human trust. As argued in [332], one of the other key aspects of building trustworthy deep learning models is ensuring their reliability. [332]. An important topic is improving the reliability of prediction confidence [333], which remains a significant challenge. DL models are trained in a closed-world setting, with the assumption that the testing data belongs to the distributions that have been encountered during the training process [334]. However, in open-world applications, DL models might be exposed to unseen data from different data distributions. A reliable DL model should not only maintain good performance with in-distribution (ID) data, but also be detect and abstain from making prediction based on out-of-distribution (OOD) data [335]. OOD detection, which aim to classify whether an input belongs to the training distribution is an active area of research [193] in various safety-critical applications, e.g. cybersecurity [336], autonomous driving [337] and medical image analysis [338]. OOD performance is therefore an important measure for the safe deployment of DL models, ensuring they can reliably handle unexpected or unfamiliar data. This is particularly important for medical image analysis, as a safe DL system should withhold any decisions related to assistive diagnosis for cases that fall outside their validated expertise, ensuring patient safety and clinical accuracy [338].

Majority of OOD detection methods in medical image analysis are adapted from those used in natural image analysis. However, this adaptation is not straightforward. Medical images posse unique characteristics across different modalities. For example, cell images are invariant to orientation changes, X-ray images exhibit high variance in feature scale, and CT scans have location-specific features [339]. Additionally, defining the concept of "out-of-distribution" data is more challenging in the medical domain. The definition must be clinically meaningful, considering factors such as data acquisition biases, which can complicate the process of identifying what constitutes out-of-distribution data. Cao et al. [338] identify several common choices for OOD data in medical image analysis, which include: 1) inputs that are unrelated to the evaluation, such as data from a different imaging modality; 2) inputs that are incorrectly prepared, such as blurry images, incorrect anatomical views, or different data acquisition protocol; and 3) inputs that are unseen, such as cases involving different or unknown diseases.

In this chapter, I introduce a new choice of OOD definition in the clinical setting. Different from the traditional OOD detection method, that relying solely on statistical differences, our definition is based on the presence of a clinically relevant region. Here, the concept of a "*clinically interesting region (CIR)*" is introduced to refers to an area within a medical image that holds particular clinical significance. These regions are assumed to contain critical, high-quality anatomical information used for clinical decision-making, for instance, abnormalities or important anatomical views. Samples contain a CIR are referred to as *clinical interesting samples*. In real-world scenarios, such samples demand extra attention from clinicians, as they are more likely to be linked to diseases or require further treatment. For instance, in a CT scan, abnormal organs might signal the presence of cancer, necessitating extra scrutiny. Similarly, in ultrasound scans, standard planes with key anatomical views require special attention for biometric measurements to assess fetal health. Samples contain the "clinically interesting region" are defined as *in clinical-distribution* (ICD) data. *Out-of-clinical distribution* (OCD) data refer to sample that lack of "clinically interesting region" - lacking key anatomical information. Examples of

OCD data include images from healthy samples or background images that do not provide clinically relevant features. The goal of OCD detection is to distinguish ICD samples from OCD data.

Here, I focus on OCD detection in fetal US application. During scanning, the sonographer adjusts the probe based on their expertise and standard plane guidelines to capture the optimal view of the fetal anatomy [340]. Once the desired standard plane is achieved, a freeze-frame is saved and annotated for measurements and the medical report. Sonographers may zoom in, zoom out, and fine-tune the probe to achieve the optimal view, often holding the probe near standard planes for extended periods. Therefore, frames captured before and after the freeze-frame may contain relevant anatomical views, even if they don't fully meet standard criteria. These frames are referred to as "Approximate standard plane" because they contain valuable, though not perfectly aligned, anatomical information. Fetal US examinations typically involve 13 anatomical views. These views include both standard planes and "approximate standard planes," which are considered ICD data as they contain relevant anatomical information. Background frames, on the other hand, contain no anatomical information and exhibit a distribution distinct from clinically relevant frames are considered OCD data. A reliable deep learning model that mimics the sonographer's scanning and detection process should not only accurately detect and classify anatomical views but also remain vigilant to background frames.

A reliable deep learning system designed to mimic the sonographer's scanning and detection process should confidently detect and classify standard anatomical views while remaining vigilant to background frames, which includes noisy and non-anatomical frames. Hence, I define frames contains meaningful anatomical information as "clinically interesting" ICD data and frames with no anatomical information as OCD data. By performing OCD, a decision boundary is learned for clinically interesting regions, and "approximate standard planes" could be selected from the background, which provides additional information for standard

planes and can be used as a guide in US scan with signal given when potential "clinically interesting region" might occur.

Similar to OOD detection [335], an OCD detector defines a scoring function that maps input features to an uncertainty score, enabling the differentiation of inputs based on this score. Numerous studies have sought to improve the performance of OOD detection score functions through techniques such as stronger data augmentation and the addition of noise at various levels [217, 335]. Early approaches primarily focused on classification-based models [149, 194, 221], whereas more recent work has explored generative models, which have shown promising results despite their increased complexity [341, 342]. However, in the context of high-resolution medical imaging, this added complexity imposes significant computational demands, raising practical concerns. Mangaokar et al. [343] argued that in medical imaging, disease characteristics are often subtle—such as minor changes in heart shape on a chest X-ray—and that generative models may oversimplify the task by producing images that appear similar but fail to capture these critical diagnostic variations. As a result, I adopt a simple classification-based backbone for the design of our OCD detection model, enabling the joint optimisation of both ICD classification and OCD detection objectives. However, prior work has shown that a trade-off often exists between classification accuracy and OOD detection performance [190].

To address this, Liu et al. [344] proposed integrating image classification with open-set recognition to allow knowledge transfer between ID and OOD classes while preserving class discrimination. Building on this idea, I aim to improve OCD detection by enhancing the cluster compactness of ICD features, based on the assumption that enhancing the cluster compactness of the ICD classes benefits both classification and distance-based OCD detection. This is achieved by incorporating a feature regulariser to encourage more discriminative representations.

Moreover, anatomical views in fetal ultrasound are highly variable due to differences in probe positioning and fetal movement. These variations often result in ICD frames that are contextually similar to adjacent background (near-OCD)

frames captured during continuous scanning. This high intra-class heterogeneity and proximity to near-OCD frames significantly increase the challenge of detection [345].

To address these challenges, I propose a softmax-conditioned variational autoencoder (VAE) boosted model, in which a classification-based model is augmented with a softmax-conditioned VAE acting as a feature regulariser. By incorporating softmax scores into the latent space, the VAE constructs a mixture of learnable class-conditioned Gaussian priors. I hypothesise that this approach enhances the constraint on ICD features and improves the accuracy of both classification and OCD detection.

In summary, this chapter makes the following contributions:

1. I define a new concept of "out-of-clinical-distribution" detection, in which samples contain "clinical interesting region" is referred as in-clinical-distribution data. OCD detector helps to distinguish samples that contain clinically interesting information used for clinical decision-making, and could be used to assist the first-stage medical examination.

2. I introduce an out-of-clinical-distribution (OCD) detection framework that integrates a classification-based backbone with the proposed softmax-conditioned VAE in the feature space. The VAE incorporates classification prediction information in both inference and generative models and acts as a feature regulariser to enforce ICD clusters' compactness.

3. Experiments on fetal US video frames show the effectiveness of the proposed OCD detection method in learning more compact ICD features and improving OCD detection performance. "Approximate Standard Planes" could be selected via OCD scores from the "background" class to enrich the dataset for anatomy learning.

6.2 Originality and Individual Role

I manual selected the OCD dataset defined and used in this chapter. I preprocessed the fetal ultrasound frames used in this chapter with a Python-based library developed by Richard Drose, a previous D.Phil student at the Institute of Biomedical Engineering, University of Oxford. I defined the concept of OCD detection. I designed, trained and tested a OCD detection method using a Python-based open-source library Pytorch [273]. This work was published and presented in MICAD 2024 [10] and the paper was awarded **the Best Paper award.**

6.3 Background

6.3.1 OOD Detection in Medical Image Analysis

Most of the recent researches about OOD detection in medical image analysis are focus on image classification. [346] adapted Gram matrix-based method [347] to skin disease classification, where OOD samples are defined as corrupted skin image, natural image and healthy skin image. The Mahalanobis-based method is deployed in [348] for unseen anomalies detection and in [349] for semantic shift detection with chest X-ray image. [338] define 3 types of OOD examples and use a feature-based binary classifier, AE and VAE for OOD detection. [342] propose to train Denoising Diffusion Probabilistic Models (DDPM) exclusively on in-distribution data (Chest X-ray) to detect OOD samples and consider all other datasets of the same dimension as OOD datasets.

6.3.2 OOD Detection Using Union of One-Dimensional Subspaces (U1D)

Classfication-based models are one of the important classes in OOD detection. In general, OOD detection techniques try to either use the class membership probabilities as a measure of uncertainty or define a measure of similarity between the input samples and the training dataset in a feature space. [221] proposed a modification based on the traditional classification-based methods to project

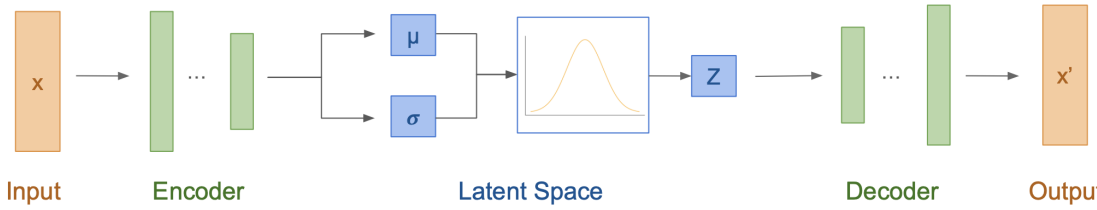


Figure 6.1: The general pipeline of VAE.

the training dataset onto a low-dimensional embedding. In particular, a union of 1-dimensional embedding is chosen to constrain the representation of ID data in the feature space. The authors hypothesise that by enforcing each know class onto a 1-dimensional subspace, the OOD data are less likely to be at the same region of these know classes, therefore the detection could be more robust and accurate. A first singular vector is selected as a robust representative of each 1-dimensional subspace and cosine similarity is used as OOD detection score.

6.3.3 Variational Autoencoder

Autoencoder is a type of neural network that designed to efficiently learn feature representation from unlabelled data [350]. An AE consists of two parts: the encoder and decoder, where the encoder compresses the input data into a latent space and the decoder reconstructs the input based on the compressed latent feature. While the standard AE maps an input into fixed latent representations - it outputs an encoding vector of size n, VAE extends the capability of AE by incorporating probability manner for describing dataset the latent space, as illustrated in Fig. 6.1. Its encoder maps an input to a distribution over the latent space by outputting two vectors of size n - the vectors of mean and standard deviation. At the decoder process, VAE generates a new sample from the distribution based on the mean and standard deviation, instead of using a single fixed point. The mean vector determines where the distribution should centred and the standard deviations determine how much from the mean it could vary. As the decoder is now exposed to a degree of local variation by sampling, a continuous and smooth latent space representation is enforced. Similar reconstructions are expected to correspond to the

input for any sampling drawn from the same latent distributions. Therefore, VAE is currently widely used for anomaly detection [351, 352], latent space manipulation [353], and data imputation and denoising [354, 355].

6.4 OCD Detection

6.4.1 Definitions

Clinically Interested Region: In the field of medical imaging analysis, I term a *clinically interesting region (CIR)* as an area within a medical image that is of particular interested to clinicians due to its relevance to patients' diagnosis or treatment, and need to be selected for closer examination or analysis. This term could be used in medical imaging analysis to put focus on areas with particular key anatomical structures, which might be clinically significant. Examples of CIR applications include:

- **Disease Diagnosis - Abnormal Anatomical Structures:** The regions where abnormality is likely located, which might indicate a disease and requires attention for further investigation, such as tumor or infections in a CT scan)

- **Quantitative Measurement - Key Anatomical Views:** The targeted region for characteristics measurement (i.e. size, shape, intensity), such as standard planes in ultrasound scans for biometric measurements

- **Treatment Planning - Post-operative Regions:** Targeted areas for treatment (i.e. surgery, radiation, therapies), such as a specific tumor for radiation.

In summary, CIRs include samples that are of particular clinical interest and hold the most relevance to the clinical decision-making process.

Figure 6.2: Examples of OCD frames in fetal US scan

Out-of-clinical Distribution: We define in-clinical distribution (ICD) and out-of-clinical distribution (OCD) in the following way based on the presence of a CIR:

- **ICD Data:** An image contains a CIR, where an abnormality or clinically significant finding is present and requires further examination, diagnosis, or treatment. Examples include standard planes during US scanning.

- **OCD Data:** An image without a CIR, where no abnormality or clinically significant finding could be detected. It is a normal image which does not contain significant medical concerns or require further clinical investigation. An example is a normal chest X-ray scan.

Here, anatomical views showing identifiable fetal structures are considered ICD data, while background frames lacking clear anatomical details are considered OCD data. Fig. 6.2 shows three examples of OCD frame within a fetal US scan. The goal of OCD detection is to distinguish anatomical views from background frames in real-time ultrasound videos, which is crucial for two reasons. Firstly, only standard planes are typically annotated and saved during scanning, yet anatomically informative frames appear throughout the scanning. Identifying these additional views offers deeper insights into fetal health and development. Secondly, detecting these frames provides valuable visual guidance for recognising critical anatomical information in real-time, potentially reducing the need for extensive training and experience for sonographers.

6.4.2 Rationale

The rationale for proposing OCD detection is as follows:

- **Dataset Bias:** In clinical practice, abnormal data, which requires clinical attention, typically represents a minority class, and the overwhelming number of normal samples can lead to class imbalance, which may adversely affect model performance. In the context of fetal US scans, a similar imbalance exists due to the predominance of background frames—such as those showing only amniotic fluid, acoustic shadows, or frames captured during probe repositioning. These non-informative frames vastly outnumber diagnostically relevant, anatomically informative ones in routine scans. If left unfiltered, this imbalance can bias the model towards learning superficial patterns from background frames, rather than focusing on features that are critical for clinical interpretation.

- **"Background Noise":** The presence of *background noise*—non-diagnostic frames during training and inference—can degrade the model's ability to learn consistent, anatomy-aware representations, ultimately reducing both performance and reliability. These background frames often contain noise or clinically irrelevant information, and including a large number of such frames can hinder the model's focus on meaningful content, limiting its capacity to extract anatomically informative features. OCD detection addresses this challenge by identifying and excluding non-diagnostic frames, allowing the model to concentrate on high-quality, semantically rich data. This improves model robustness and efficiency, while also supporting safer and more trustworthy deployment in clinical settings.

6.5 Methodology for OCD Detection

Let us denote D_{train} and D_{test} as dataset used for classification training and OCD detection test, respectively. As shown in Fig. 6.3, D_{train} contains labelled in-clinical distribution data only, where $D_{train} = \{(x_n, y_n)\}_{n=1}^{N}$ and $y_n \in \{1, 2, ..., L\}$.

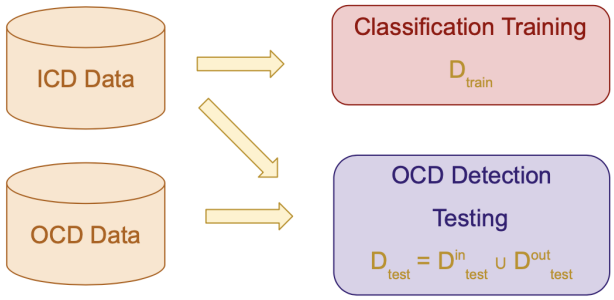


Figure 6.3: Demonstration of the composition of training and testing dataset

D_{test} consists of both ICD and OCD data, and the task is to distinguish whether a testing input contains the "clinically interesting region" and consists of three steps as depicted in Fig. 6.4.

6.5.1 Training Stage

In the first step, a classification model is trained simultaneously with the proposed softmax-conditioned VAE on ICD training data D_{train}.

Classification Model

We adapted the approach from [221] to project ICD features obtained from the classification model onto a union of 1-dimensional subspaces, with each class occupying a distinct, mutually orthogonal subspace. Such projection provides a stronger constraint on the feature space compared to traditional approach. To achieve this feature embedding, two key adjustments were made. Firstly, the weights of the last fully connected layer are initialised with orthogonal vectors and kept frozen during training; secondly, cosine similarity is utilised in the calculation of the softmax function.

For a given input x_i, the predicted probability of belonging to class l is firstly calculated as $p_{i,l} = \frac{c_{i,l}}{\sum_j c_{i,j}}$, where $c_{i,l} = \frac{w_l^T f_i}{||f_i||}$ denotes the cosine similarity between the weight vector w_l of the final fully connected layer (corresponding to class l) and the feature vector f_i of input x_i. This followed by a sharpening operation defined as $P_{i,l} = \frac{p_{i,l}}{s_{i,l}}$, where the sharpening factor is given by $s_{i,l} = \sigma(BN((w_s)_l^T f_i))$.

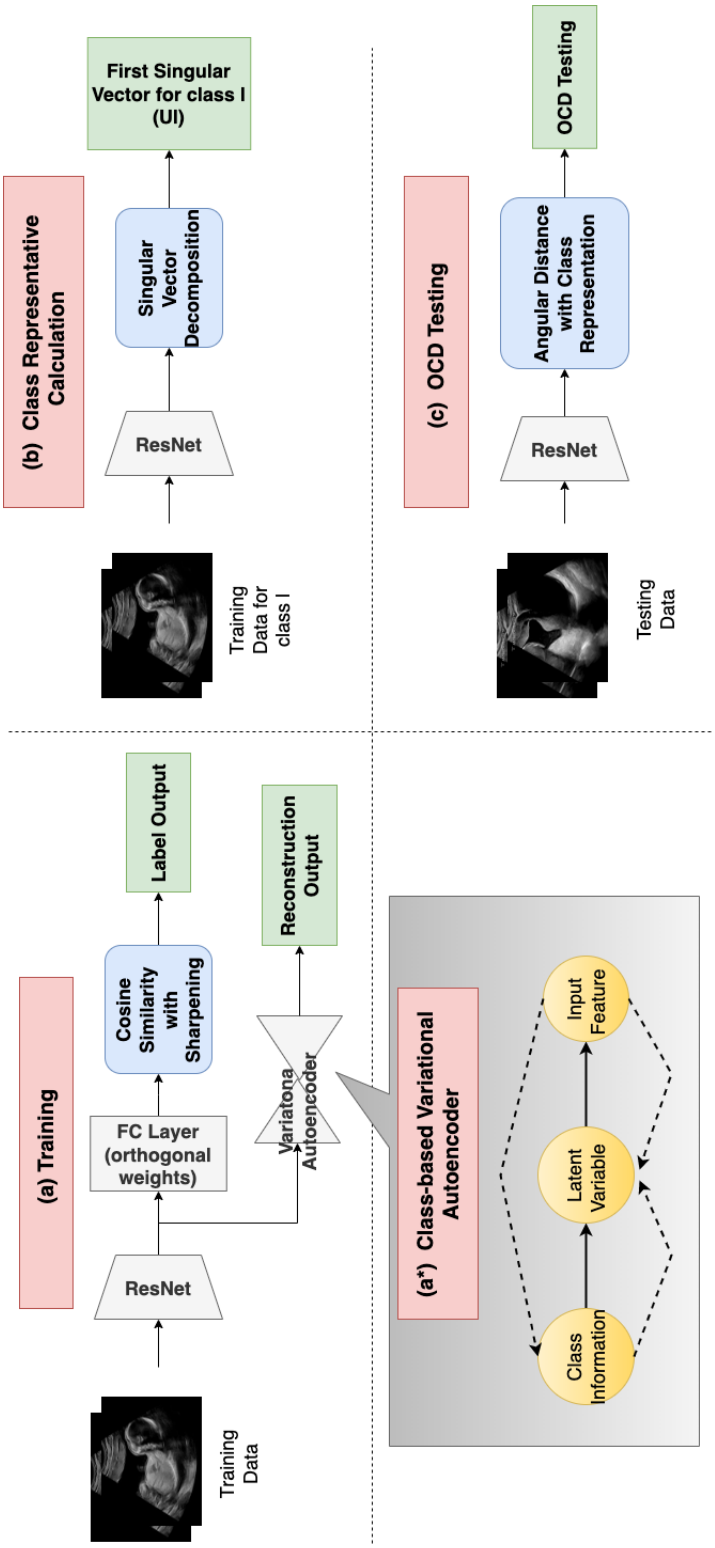


Figure 6.4: OCD detection framework. (**a**): A classification model and variational autuencoder are trained simultaneously using ID training data. (**b**): First singular vector is calculated as class representative. (**c**): OCD detection based on angular distance between testing features and class representatives. (**a***): graphical model of softmax-conditioned VAE (→: generative model; - - →: inference model).

Here, $(w_s)_l$ represents the weight vector of the sharpening layer corresponding for class l, $f_i \in \mathbb{R}^d$ is the learned feature of input, σ denotes the sigmoid function and BN refers to batch normalisation.

Cosine similarity is employed to encourage alignment between feature vectors and their corresponding class weight directions in the final layer, promoting class-specific feature representation and effectively forming a union of one-dimensional subspaces. An orthonormal weight matrix is used to maximise inter-class separation, thereby enhancing feature discrimination and reducing class overlap in the embedding space. The sharpening layer is introduced to improve the model's ability to extract fine-grained features and enhance robustness. The use of a union of one-dimensional feature subspaces aims to increase the compactness of ICD samples while minimising the error rate in OCD detection. Therefore, the input features of each group are imposed to lie on a subspace in the direction of w_l $(l = 1, 2, ..., L)$, which are 1-dimensional and mutually orthogonal. The classification loss is calculated based on modified output logit $P_{l,n}$.

Softmax-Conditioned Variational Autoencoder (SC-VAE)

In a standard VAE, an isotropic Gaussian is typically used as the prior over the latent space. While this simplifies the modelling process, it limits the model's ability to capture complex, class-specific representations [356]. In multi-class settings, a shared prior forces all latent features into a single overlapping distribution, which can blur important class-related characteristics and lead to the loss of class discriminative information during encoding and reconstruction. To address this limitation, I propose a softmax-conditioned VAE, which explicitly incorporates soft probabilistic class information into the latent space. By leveraging softmax outputs from a pre-trained classifier, the SC-VAE encourages the latent space to segregate into distinct, class-specific regions, therefore preserving key semantic features and improving the intra-class compactness.

In the SC-VAE, the prior is modelled as a mixture of class-conditioned Gaussian distributions, where each class is associated with its own learnable mean and

covariance. Rather than assuming a single prior for all inputs, the model assumes that each class corresponds to a distinct latent distribution. The softmax scores from the classifier serve as weights over these class-specific Gaussians: for a given input, the probability of sampling from the Gaussian of class l is given by the model's softmax probability for that class. Consequently, each input is associated with a weighted mixture of class-conditioned priors, encouraging more structured and semantically meaningful latent representations for each class. I hypothesise that this design also acts as a feature regulariser, guiding the model to learn compact, well-separated clusters for each ICD class.

Importantly, the SC-VAE uses the full softmax probability distribution as the conditioning signal, rather than relying on a single hard class label. This choice enables the model to account for prediction uncertainty, which is particularly relevant in clinical applications such as fetal US, where certain standard planes—such as cardiac views—are challenging to distinguish, even for experienced sonographers. In these cases, the model may assign high probabilities to multiple classes instead of one clear winner. Conditioning on only the most likely class could lead to biased or misaligned latent representations. By incorporating the entire softmax vector, the model captures the range of plausible class memberships, therefore generating a more robust encoding that reflects the model's confidence. This soft class-conditioning is especially beneficial in borderline cases, where OCD detection tends to fail if one relies only on confident predictions.

Another key advantage of the SC-VAE framework is its capacity to mitigate posterior collapse—a common issue in VAEs, where the posterior distribution becomes too similar to the prior, leading the latent variable to stop encoding meaningful information. This typically occurs when the prior is overly simplistic or the decoder is excessively powerful. In contrast to a fixed isotropic Gaussian, the prior in SC-VAE is learnable and class-specific, making it more informative and better aligned with the underlying data distribution. During training, the class-conditioned prior can adapt to the structure of the encoder's posterior, preventing

trivial convergence and enabling the model to learn richer and more informative latent representations.

We further hypothesise that OCD detection is improved by enforcing greater intra-class compactness within the latent space, leading to better separation between ICD and OCD samples. Consider two multivariate Gaussian distributions representing the latent features of OCD and ICD samples of class i: $\mathcal{N}(\mu_O, \Sigma_O)$ and $\mathcal{N}(\mu_i, \Sigma_i)$, respectively. Their separation could be measutred using the Bhattacharyya distance, which is calculated as the sum of Mahalanobis distance between means and compactness (covariance) of each distributions. Increasing the compactness of $\mathcal{N}(\mu_i, \Sigma_i)$ thus increase the Bhattacharyya distance, improving separability. The SC-VAE explicitly encourages this compactness through the KL divergence term during training, ensuring that each ICD class occupies a small, well-defined volume in latent space. This principle underpins the SC-VAE's role as a feature regulariser: by enforcing the compactness of each ICD class, the model enhances its ability to detect clinically irrelevant frames.

The overall SC-VAE architecture is illustrated in Fig. 6.4(a*), which shows how class information is incorporated into both the inference network (encoder) and the generative model (decoder). By using softmax outputs as soft class-conditioning for both stages, the model maintains flexibility and avoids the rigidity of hard class labels. This approach allows the model to emphasise the most likely class while still accounting for less proable predictions, resulting in more stable and informative latent features. Integrating class context into both the encoding and decoding processes ensures that the structure of the latent space and the quality of reconstruction are both grounded in clinically meaningful information.

In the generative model, the latent variable z_i is sampled from class-conditioned Gaussian distributions, with the probability of sampling from a specific class distribution determined by its softmax score. The probability density function (pdf) of z_i conditioned on class l is formalised as:

$$p_\theta(z_i|l) \sim N(\mu_\theta(l), diag(\sigma^2_\theta(l))), \tag{6.1}$$

where $\mu_\theta(\cdot)$ and $\sigma^2_\theta(\cdot)$ are generated by a class-conditioned encoder parameterised by $\theta(\cdot) : \mathbb{R}^L \rightarrow \mathbb{R}^e$. $p_\theta(z_i|l)$ is used as the Gaussian prior for class l.

In the inference model, the posterior of z_i is dual-conditioned on both the input feature and class, with its pdf given by:

$$q_\phi(z_i|f_i, l) \sim N(\mu_\phi(f_i, l), diag(\sigma^2_\phi(f_i, l))), \tag{6.2}$$

where $\mu_\phi(\cdot)$ and $\sigma^2_\phi(\cdot)$ are generated by a feature-class dual-conditioned encoder parametrised by $\phi(\cdot) : \mathbb{R}^{d+L} \rightarrow \mathbb{R}^e$, d denotes the dimension of latent space.

The designed SC-VAE is trained using evidence lower bound optimisation (ELBO) [357], which involves minimising two losses: the reconstruction loss between the input features and generated targets (measured by MSE) and the distribution discrepancy between the prior and posterior in the latent space (measured by KL divergence). Given the softmax score $p_{i,l}$, which indicates the likelihood of input x_i belonging to Gaussian distribution conditioned on class l in the latent space. The loss function of the designed SC-VAE is defined as:

$$\mathcal{L}_{vae} = \frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L} p_{i,l} \times \{MSE(f_i, \hat{f}_{i|l}) + KL[q_\phi(z_i|f_i, l)||p_\theta(z_i|l)]\}. \tag{6.3}$$

We jointly train the baseline classification model and softmax-conditioned VAE with loss:

$$\mathcal{L}_{train} = \alpha_{cls} \times \mathcal{L}_{cls} + \alpha_{vae} \times \mathcal{L}_{vae}, \tag{6.4}$$

where α_{cls} and α_{vae} are the weights for classification loss and VAE loss. KL distances are estimated using a single latent sample [356]. A detailed demonstration of the loss generation process is provided in Algorithm 5.

6.5.2 Class Representative Calculation

In statistics, the first singular vector has been used as a rubost estimator of mean and covariance [358]. According to [359], the first singular vector can also be treated as a class representative, as it preserves most of the information and is robust to noise and augmentation.

Algorithm 5 Loss generation algorithm for joint training of SC-VAE with classification model

Input: Dataset $X = \{x_i\}_{i=1}^{N}$, class labels $Y = \{y_i\}_{i=1}^{N}, y_i \in \{1, \ldots, L\}$, batch size B, number of training iterations T
Initialize: Feature Extractor $f_\psi(x)$, Classifier $h_\xi(f)$, Prior $p_\theta(z|y)$, Encoder $q_\phi(z|x, y)$, Decoder $g_\eta(x|z, y)$
for each training iteration $t = 1, \ldots, T$ **do**
    Sample mini-batch $\{x_1, x_2, ..., x_B\}$ from dataset X
    for each x in the mini-batch **do**
        Extract feature representation: $f = f_\psi(x)$
        Predict class probability: $\hat{y} = h_\xi(f)$
        for $y = 1$ to L **do**
            Compute mean and log-variance from encoder:

$$\mu_\phi, \log \Sigma_\phi = q_\phi(z|f, y)$$

            Sample $\epsilon \sim \mathcal{N}(0, I)$
            Compute latent variable:

$$z = \mu_\phi + \sigma_\phi \odot \epsilon, \quad \text{where} \quad \sigma_\phi = \sqrt{\text{diag}(\Sigma_\phi)}$$

            Compute mean and log-variance of learnable conditional prior:

$$\mu_\theta, \log \Sigma_\theta = p_\theta(z|y)$$

            Compute reconstructed feature: $\hat{f} = g_\eta(f|z, y)$
            Compute Reconstruction Loss: $\mathcal{L}_{\text{rec}} = \text{MSE}(f, \hat{f})$
            Compute KL Loss: $\mathcal{L}_{\text{KL}} = \text{KL}(q_\phi(z|f, y)||p_\theta(z|y))$
            Compute VAE Loss for given x, y: $\mathcal{L}_{x,y} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}}$
        end for
    end for
    Compute classification loss:

$$\mathcal{L}_{\text{cls}} = -\sum_{i=1}^{B}\sum_{y=1}^{L} y_i \log \hat{y}_i$$

    Compute VAE loss:

$$\mathcal{L}_{\text{vae}} = \sum_{x=1}^{B}\sum_{y=1}^{L} \hat{y}_i^{\{y\}} \times \mathcal{L}_{x,y}$$

    Compute total loss:

$$\mathcal{L} = \alpha_{cls} \times \mathcal{L}_{\text{cls}} + \alpha_{vae} \times \mathcal{L}_{\text{vae}}$$

    Compute gradients $\nabla_{\psi,\xi,\phi,\eta,\theta}\mathcal{L}$
    Update parameters $\psi, \xi, \phi, \eta, \theta$ using SGD optimiser
end for

Here, I use it as a representatives for each ICD class. For ICD class l, its corresponding first singular vector u_l is calculated using singular value decomposition (SVD). In this process, the matrix containing the aggregated input features of the training data from that class is decomposed, and the first singular vector is extracted to represent the class.

6.5.3 OCD Detection

As the features of each input lie on a 1-dimensional subspace and are regulated by the proposed VAE to ensure intraclass compactness, I assume they will occupy a tiny region with no volume in the space, concentrated around the first singular vector. Therefore, the angular distance between the first singular vector and the input sample can be used as the score for OCD detection.

At test time, I use the minimum angular distance δ_n between the extracted test feature and the first singular vector of each class as an uncertainty score

$$\delta_n = \min_l(arccos(\frac{f_t^T u_l}{||f_n^T||})) \tag{6.5}$$

Given testing dataset D_T and a chosen threshold δ_T, y_{pred}, which is the prediction of our proposed OCD detection method is defined as:

$$y_{pred} = \begin{cases} 0(\text{ICD}) & \text{if } \delta_n < \delta_T \\ 1(\text{ICD}) & \text{otherwise} \end{cases} \tag{6.6}$$

The probability of a test input x_n belongs to ICD could also be calculated using probability function $P(\delta_n < \delta_T | n \in D_T)$. By employing Monte Carlo sampling on δ_n, the probability function is estimated as $\frac{1}{M}\sum_{m=1}^{M} 1(\delta_n^m < \delta_T)$, where M is the number of Monte Carlo samples and δ_n^m is the m-th sample.

6.6 Experiments

6.6.1 Dataset

Our experiments are based on the large-scale fetal US video dataset, named PULSE, as described in Chapter 3. Dataset V, discussed in 3.4.3, consists of a large number of labelled anatomical views are used as the ICD dataset. I split the ICD dataset so

that 80% is used for the training stage and the remaining 20% is used as the ICD testing dataset. As discussed in Chapter 3, less than 20% of the frames are labelled either during scanning or through manual annotation by sonographers. A large number of background frames and non-standard anatomical views remain unlabelled and largely indistinguishable, resulting in the absence of a ready-to-use OCD dataset. To address this, I manually created an OCD dataset by selecting 1,085 background frames based on three specific criteria: frames containing only dark amniotic fluid without visible fetal anatomy, frames with ultrasound artifacts (such as acoustic shadows or speckle noise) that obscure anatomical details, or frames with empty regions when the probe was briefly moved away from the fetus. These criteria ensured that the OCD dataset contained no clear anatomical information. The testing dataset for OCD detection includes both ICD testing data and OCD data.

6.6.2 Implementation Details

We used ResNet-18 [287] as the backbone architecture for the classification model. The classification model described in Sec. 6.5.1 was jointly trained with the proposed SC-VAE. These models were implemented with the PyTorch framework and trained with an SGD optimiser for 200 epochs with a batch size of 32, a moment of 0.9, a weight decay of 1×10^{-4} and an initial learning rate 0.1 that decayed by 0.1 after every 60 epochs. The choice of initialisation parameters and optimiser followed the methodology proposed by [57]. Data augmentation was performed including horizontal flipping, brightness variation, resizing, and centre cropping.

6.6.3 Evaluation Metrics

During evaluation, OCD images are treated as positive examples, meanwhile, ID images from all classes are treated as negative for global metrics and ID images from each class are treated as negative for class-wise metrics. Four global and one class-wise metrics are used to evaluate OCD detection performance. These are summarised in Table 6.1.

Table 6.1: Global and Class-wise Evaluation Metrics

Metric	Description
Detection Error (%) ↓	Minimum misclassification rate over all possible thresholds.
AUROC ↑	Area under the FPR against TPR curve.
δ_T at 95% TPR ($\times 10^{-2}$) ↑	Angular distance based decision boundary at 95% TPR.
Global FPR at 95% TPR (%) ↓	Percentage of ID data that wrongly classified as OCD at 95% TPR.
Class-wise FPR at 95% TPR (%) ↓	Percentage of ID data in each class that wrongly classified as OCD at 95% TPR.

6.6.4 Comparison Methods

In this section, I compare seven selected methods, including both classification-based and generative-based OOD detection approaches. In our experiments, these methods have been adapted to the OCD setting, specifically tailored to detect OCD samples rather than general OOD data.

- **MSP [194]**: The maximum softmax probability is proposed as an OCD score for OCD detection.
- **ODI [195]**: Temperature scaling and small input perturbations are used to compute more reliable OCD scores.
- **Mahalanobis [212]**: Mahalanobis distance is used to calculate the OCD score by measuring how far a sample's feature representation deviates from the distribution of ICD features.
- **U1D [221]**: A union of one-dimensional projections is employed to project the learned features onto orthogonal 1D subspaces, enhancing the separability between ICD and OCD samples for more effective OCD detection.
- **AE: [360]**: An autoencoder is used to reconstruct the input, and the reconstruction error is calculated as the OCD score for detection.

- **MemAE [341]**: In a memory-augmented autoencoder, instead of directly using the encoder output for reconstruction, a query is generated from the encoded input to retrieve the most relevant items from an external memory.
- **DDPM OOD [342]**: A DDPM is used to reconstruct inputs corrupted with different levels of noise. The reconstruction errors across these noise levels are combined into a multi-dimensional score, which is used for OCD detection

The first four classification-based methods are selected for comparison because our proposed approach also falls within this category, and these represent state-of-the-art techniques in classification-based OOD/OCD detection. In addition, I include several generative-based methods, as they are emerging as strong alternatives in the field. Specifically, the AE serves as a classic baseline, while MemAE shares a similar idea to our design—enhancing reconstruction using the most relevant retrieved data, although it does not explicitly incorporate class information. DDPM is a recently proposed and highly promising method that achieves strong performance despite its high computational cost. Including DDPM in our comparison allows us to assess whether our method offers a favourable trade-off between performance and complexity.

6.7 Results

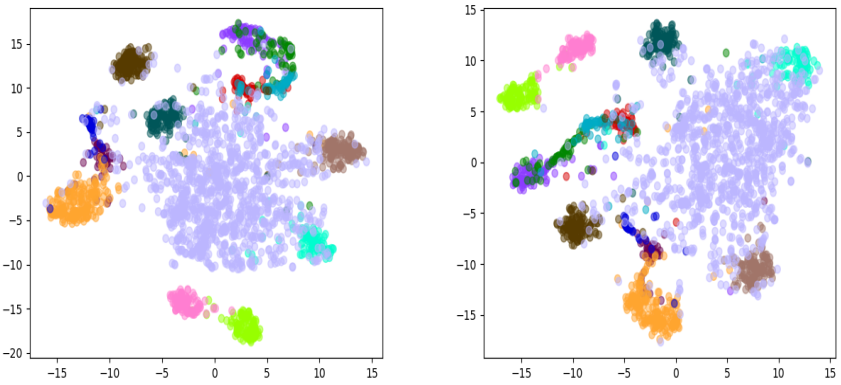
6.7.1 Quantitative Results

Table 6.2 shows a quantitative evaluation of OCD performance compared between different methods. Form the results, I observe that our proposed method achieves the best performance among all the compared classification-based methods, with an improvement of 4.46, 2.06, 8.35, 1.00 in detection error, respectively. The results indicate the effectiveness of our design in enhancing the separation between ICD and OCD data. The comparison between U1D and our method validates the effectiveness of our proposed VAE head in improving the compactness of ICD data, leading to enhanced OCD detection performance.

Method	**FPR at 95% TPR (%) ↓**	**δ_T at 95% TPR ($\times 10^{-2}$) ↑**	**Detection Error (%) ↓**	**AUC ↑**
Classification-based				
MSP [194]	25.8	38.92	11.32	94.32
ODIN [195]	18.64	43.58	8.92	96.13
Mahalanobis [212]	32.57	29.37	15.21	87.62
U1D [221]	12.7	50.25	7.86	97.19
Ours	9.17	57.45	6.86	97.53
Generative-based				
AE [360]	47.85	21.68	19.82	81.55
MemAE [341]	21.58	39.73	12.88	88.69
DDPM OOD [342]	12.06	54.18	7.41	95.51
Dual Heads				
DDPM + ours	9.12	57.92	6.69	97.64

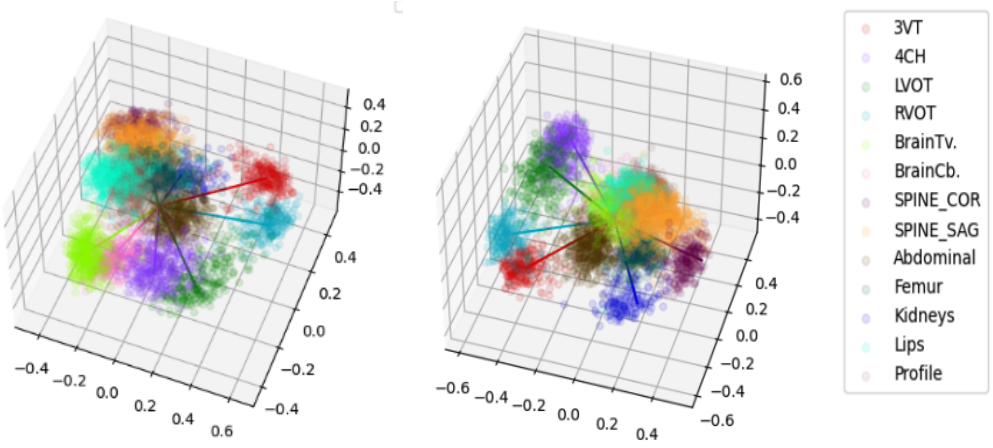
Table 6.2: Performance comparison between different methods

For generative-based methods, the AE offers the worst performance. As a simple reconstruction method, it may not be well-suited for recognising and reconstructing complex medical images or detecting outliers, especially since outliers can appear similar to ICD data. The improved performance of MemAE compared to AE highlights the importance of incorporating class information in the OCD detection process, as it enhances the model's ability to differentiate between ICD and OCD data. Notably, our method also outperformed method that using DDPM for OCD detection. In DL, there is often a trade-off between computational complexity and performance. Due to the computational complexity of DDPM, our method offers a simpler yet effective alternative that produces comparable or even better results. I also tested a model that simultaneously trained DDPM and our model as dual heads. There is an improvement in performance, but it is not statistically significant. However, the increase in computational cost is significant, suggesting that the added



(a) t-SNE (U1D) **(b)** t-SNE (ours)

Figure 6.5: PCA and t-SNE visualisations of baseline and SC-VAE boosted methods



(a) PCA (U1D) **(b)** PCA (ours)

Figure 6.6: PCA and t-SNE visualisations of baseline and SC-VAE boosted methods

complexity of the model may not be worth the trade-off. This indicates that more complex models do not always result in proportionally greater accuracy.

6.7.2 Qualitative Results

Here, a qualitative comparison is made between our model and U1D, as the same classification backbone and feature representatives are used, making the visualisation comparison more straightforward. A t-SNE visualisation and a 3D PCA plot of class representatives and input features are included are qualitative results. Fig. 6.5 shows the t-SNE visualisation, in which our proposed method shifts the OOD

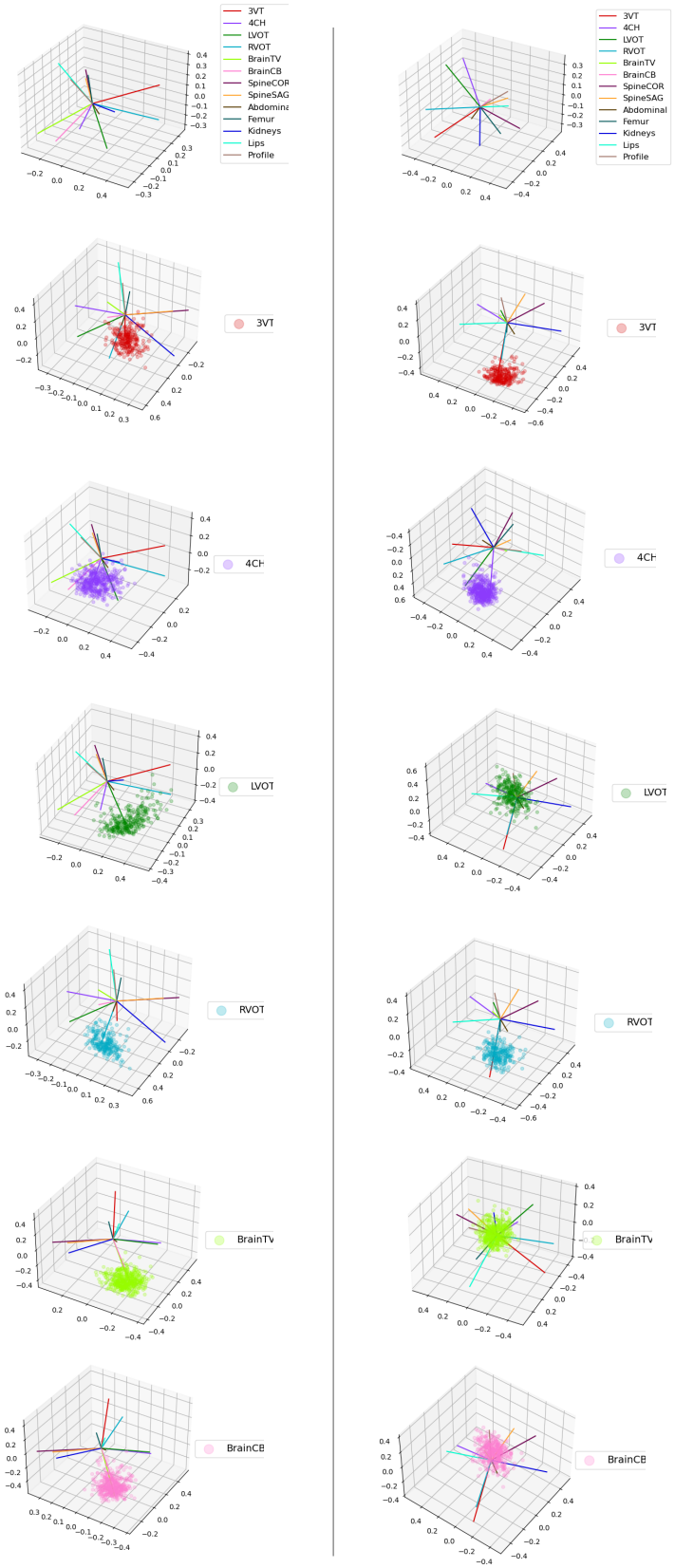


Figure 6.7: Class-wise PCA Visualisations of ICD testing data (Left: U1D; Right: ours).

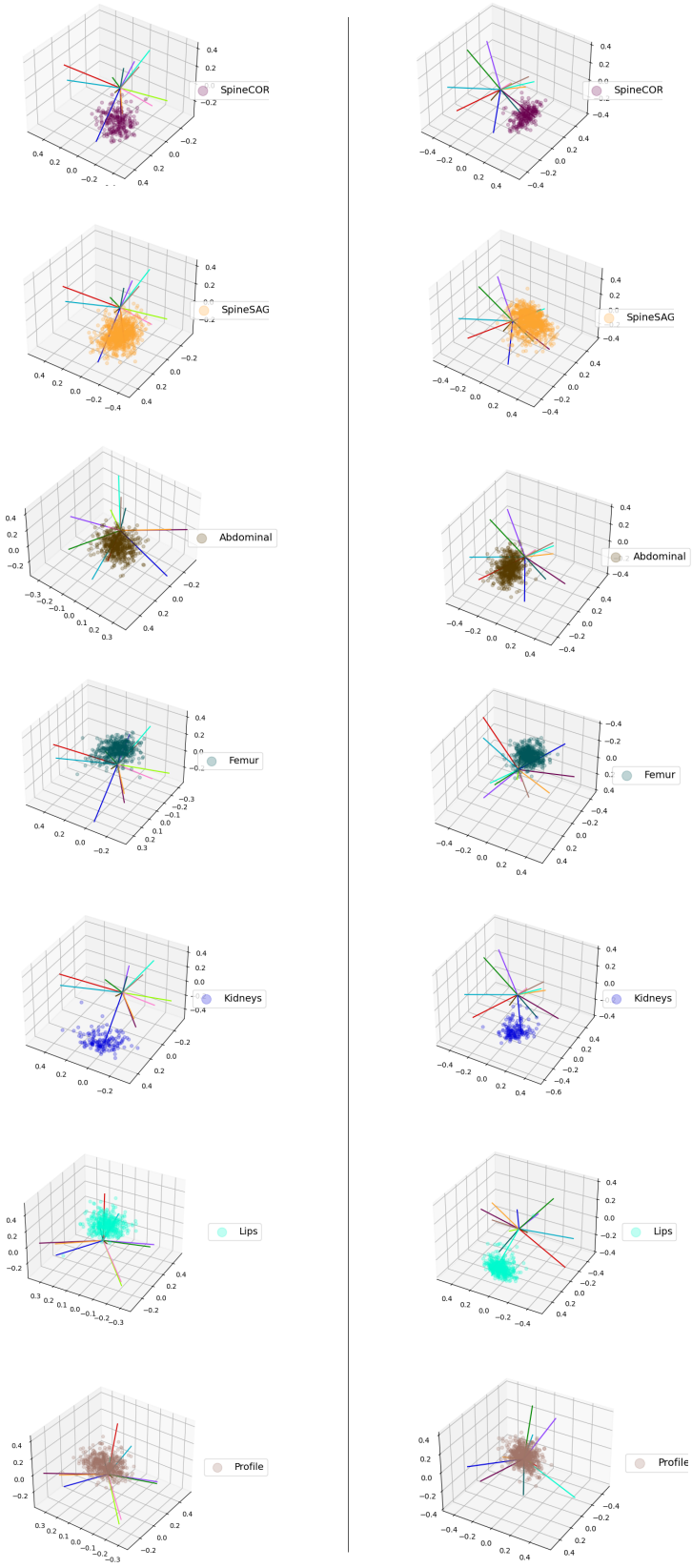


Figure 6.8: Class-wise PCA Visualisations of ICD testing data (continued).

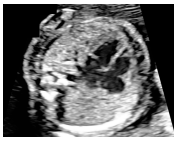
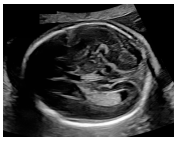
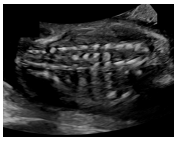
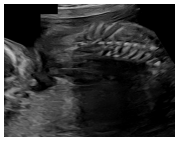
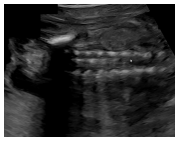
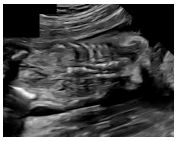
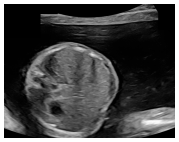
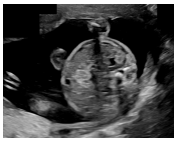
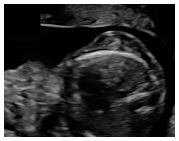
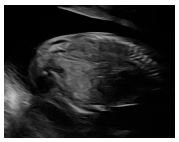
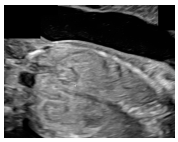
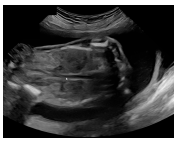
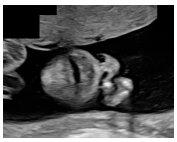
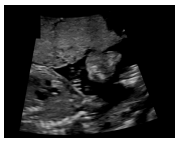
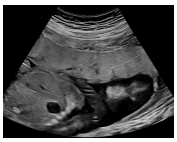
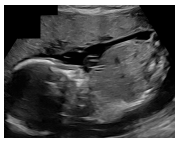
Standard Plane View	Annotated Training Data	"Approximate Standard Planes" (from Background)		
4CH				
BrainTv.				
SpineCor				
Abdominal				
Femur				
Kidney				
Lips				
Profile				

Table 6.3: Labelled ICD data and select "Approximation Standard Planes" for each standard plane view. **1st Column**: Standard plane (SP) views (8 out of 13 classes selected). **2nd Column**: Labelled ICD data, which is SP view. **3rd Column**: "Approximation Standard Planes", which are frames in the background (i.e. real-time ultrasound scan video) that are predicted as ICD, and further classified into each SP view.

Figure 6.9: Examples of "Approximation Standard Planes" for 4CH

features further away from ID samples. Fig. 6.6 presents the 3D PCA plots, which contain both class representatives (as a solid line) and ID testing data (as a circle), with class-wise 3D PCA visualisation shown in Fig. 6.7 and 6.8. Each coloured line represents a class-representative vector and the coloured dots represent ICD data for different classes. Hence, the spread of a single-coloured cluster represents the compactness of ICD testing data for a given class. Here, I include class-wise visualisations for 13 standard plane classes on both U1D and our proposed model. Hence, it verifies that the first singular vector is a good class summary as ID testing feature is spread around the vector for each class and illustrates that ID features become more compact intra-class in the feature space under the proposed model.

Table. 6.3 demonstrates "Approximation Standard Planes" selected from the background (i.e. real-time ultrasound scan video) of 8 out of 13 standard plane views. These frames are identified as ICD in OCD detection, and then further classified into each SP view. Here δ_T at 95% TPR is used as threshold to determine the boundary of ICD. The selected approximations closely resemble the standard plane view for most classes by capturing non-standard anatomical information specific to each class. However, for cardiac views—such as 4CH—the "Approximation Standard Planes" approach proves less effective compared to other classes. This is due to the

inherent difficulty of distinguishing between cardiac views, which even experienced sonographers find challenging, as these views share similar global structures with only subtle local differences. Consequently, identifying accurate approximations for these classes is a more complex task. Fig. 6.9 presents additional approximations for the 4CH view. Some of these examples contain anatomical features associated with other cardiac views, such as 3VT, which shares a similar global pattern. Others resemble abdominal views due to the presence of rounded edges, or include frames with noise or incomplete anatomical structures that hinder further diagnostic interpretation. Such ambiguity is rarely observed in non-cardiac classes, further highlighting the complexity of approximating cardiac views. This behaviour aligns with the challenges faced by human experts, and is therefore a reasonable outcome. As a conclusion, our methods could effectively select "Approximation Standard Planes" from the background, which is valuable for the ultrasound search and selection process.

6.8 Ablation Study

Ablation experiments were conducted to investigate the impact of using a VAE as a regulariser, with different choices of class information encoded, on the model's performance. The following choices of models are considered:

- Baseline: A standard classification-based model U1D without any regulariser.
- + AE: An autoencoder is used as a regulariser, but without incorporating any class-specific information.
- + Vanilla CC-VAE: A class-conditioned VAE regulariser that use a standard multivariate Gausian as prior.
- + CC-VAE: A VAE regulariser that incorporates learnable class-conditioned information.
- + SC-VAE (ours): A VAE regulariser that incorporates softmax score as class information to guide the latent space.

Table 6.4: Evaluation Results of Global Metrics on OCD Detection

Metric	Baseline	+AE	+ vanilla CC-VAE	+CC-VAE	+SC-VAE (ours)
FPR at 95% TPR (%) ↓	12.70	12.84	11.34	12.17	**9.17**
δ_T at 95% TPR ($\times 10^{-2}$) ↑	50.25	54.87	56.93	55.13	**57.45**
Detection Error (%) ↓	7.86	8.23	7.72	7.72	**6.86**
AUROC ↑	97.19	96.84	97.03	97.07	**97.53**

Table 6.4 demonstrates the evaluating results of OCD detection with global metrics. Here, OCD detection with classification model only is used as a baseline. Different types of reconstruction models (AE, vanilla CC-VAE, CC-VAE, and SC-VAE) are added to the input features. Vanilla CC-VAE and CC-VAE are defined as using a standard multivariate Gaussian as prior and using a learnable class-conditioned Gaussian as prior, respectively. The proposed SC-VAE method outperforms the other method, which validate our design. An increase in δ_T at 95% TPR can be observed in different types of reconstruction models compared to the baseline model, which indicates that OCD features lie further away from ID feature space, hence OCD and ID features are more separable. This observation is reasonable since OCD features might be skewed from ID features during joint training due to its relatively high reconstruction error.

Table 6.5 demonstrates class-wise FPR at 95% TPR with class-wise means and standard deviations. Our proposed model achieves the lowest mean, standard deviation, and lowest FPRs among 8 of 13 classes. Compared to the baseline model, the proposed model tends to provide a bigger drop in FPR in classes with high FPRs. In the case of cardiac views—such as 3VT, 4CH, LVOT, and RVOT—high FRPs could translate to cardiac view classes having high similarity and being near neighbours in the feature space. These observations suggest that the proposed method might be a potential solution to the "near neighbour" issue.

6.9 Conclusion

In this work, I introduced a new concept - "out-of-clinical distribution" detection - for medical image analysis, with ICD data is defined as images that contain important

Table 6.5: Classwise FPR at 95% TPR on OCD Detection (%)

Anatomy	Baseline	+ AE	+ vanilla CC-VAE	+ CC-VAE	+ SC-VAE (ours)
3VT	19.05	20.63	23.80	17.46	**15.87**
4CH	19.44	21.29	15.74	18.52	**11.11**
LVOT	22.54	18.31	21.12	25.35	**5.64**
RVOT	22.58	25.89	25.80	**14.51**	16.13
BrainTv.	4.27	5.13	4.27	5.13	**3.42**
BrainTc.	6.73	3.85	**2.88**	3.85	3.85
SpineCor	20.69	13.79	13.79	20.69	**12.07**
SpineSag	11.11	13.04	10.63	12.56	**8.21**
Abdominal	6.98	5.42	**3.88**	8.53	9.30
Femur	10.40	12.0	9.61	11.20	**9.60**
Kidney	22.50	27.50	27.50	**17.50**	20.00
Lips	18.75	17.88	13.39	13.39	**11.61**
Profile	5.20	5.93	**5.19**	6.67	6.67
Mean	14.63	14.66	13.67	13.49	**10.26**
Std	6.97	7.71	8.31	6.15	**4.76**

and clinically interesting anatomical information, and OCD data is defined as images that do not contain anatomical information used for clinical decision making. A new type of OCD detection method was proposed by adding a novel softmax-conditioned variational autoencoder on input features. The softmax-conditioned variational autoencoder incorporates prediction information in both inference and generative models, which use a class-conditioned mixture Gaussian as prior, and enforce the ICD feature compactness. Experiments on fetal ultrasound data demonstrated that the proposed method improves detection performance in both classification-based and generative-based approaches, and remains effective even within classes that have high similarity and are near neighbors in the feature space. It achieved comparable results when adding DDPM as dual heads, suggesting that increasing model complexity does not always lead to equivalent gains in accuracy.

We cut and clear the fields, mowing the thorny undergrowth. The great fields are full of crops, already sown and well tended. Now we reap the harvest, and next year the millet will again be plentiful.

— *The Book of Songs, Hymns of Zhou, Cutting the Weeds.*

7 Discussion and Conclusion

Contents

7.1 Summary of Contributions

This thesis has presented three research studies aimed at addressing key challenges in adopting deep learning applications in clinical practice. Self-supervised learning was explored as an alternative framework to mitigate the need for large annotated datasets. Meanwhile, model robustness, explainability, and reliability were investigated in separate works to tackle the broader issue of model trustworthiness. Below, I summarise the main contributions of this thesis.

1. **Chapter 4: Fetal Ultrasound Video Representation Learning using Contrastive Rubik's Cube Recovery** In this chapter, I developed a novel self-supervised learning approach for robust local representation learning from raw fetal US scan data. The key contributions are as follows: First, I proposed feature-level representation learning to enhance robustness, enabling the model to discard US-specific noise and learn more stable features. Second,

I introduced a local representation learning strategy tailored to the unique characteristics of US data. This involved a novel local contrastive pair generation method, along with Rubik's Cube recovery and cube reconstruction tasks, to generate contrastive sub-cube pairs from US videos and effectively capture local spatio-temporal features. Third, I employed 3D Swin Transformers as a stronger feature extractor to further strengthen the feature learning process. The effectiveness and generalisability of the proposed method were validated through both in-domain and cross-domain clinical classification tasks. I further validate the design of our method through two ablation studies.

2. **Chapter 5: Explainability of Self-Supervised Representation Learning for Fetal Ultrasound Video** In this chapter, I explored the explainability of self-supervised learning models in the context of fetal US by leveraging video data and gaze-tracking information to generate clinically meaningful explanations. The key contributions are as follows: First, I defined explainability in a clinical setting as the model's ability to capture anatomy-aware knowledge. Based on this definition, I introduced a novel set of quantitative metrics that incorporate clinical-driven insights into the explaination process. Specifically, I utilised visually salient landmarks—regions that naturally attract sonographers' attention—and assessed how well the focus of the learned representations aligns with sonographers' gaze patterns during image interpretation. The effectiveness of the proposed metrics was validated across various SSL tasks, demonstrating their ability to provide more consistent and comparable explainability assessments than traditional qualitative methods, which are often coarse and difficult to interpret in the context of fetal US.

3. **Chapter 6: Out-of-Clinical-Distribution Detection with a Softmax-Conditioned Variational Autoencoder Boosted Classification Model** In this chapter, I introduced the novel concept of out-of-clinical-distribution detection to enable deep learning models to filter out irrelevant frames from large, heterogeneous real-time US scans and to assess model reliability in fetal

US analysis. The key contributions are as follows: First, I defined "in-clinical-distribution" data as frames containing clinically relevant regions with key anatomical information essential for diagnostic decision-making—for example, standard planes in fetal US. Second, I proposed the first OCD detection method for fetal US, based on a classification framework that incorporates a softmax-conditioned variational VAE. This SC-VAE model integrates class-dependent information into the reconstruction process while acting as a feature regulariser. I validated the effectiveness of our method on fetal US datasets, demonstrating superior performance compared to both classification-based and generative models, and achieving results comparable to state-of-the-art diffusion models—highlighting the simplicity and efficacy of our approach. Finally, I validate the design of our method through the ablation study.

7.2 Future Works

In the previous section, I discussed the technical contributions that has been introduced in this thesis. Beyond the technical advancements, these approaches hold potential for broader applications and could serve as a blueprint for future research. They may inspire further exploration and innovation in related areas, particularly in conjunction with work being developed concurrently by my peers. Below, I outline and discuss key areas for potential impact and further research, structured according to the contributions of each chapter.

1. **Chapter 4: Fetal Ultrasound Video Representation Learning using Contrastive Rubik's Cube Recovery** The proposed SSL approach is currently tested on fetal US scan, but also has the potential to be applied to other imaging modalities that exhibit significant local variations. During the development of this method, several other SSL approaches have also emerged for fetal US analysis. For instance, Fu et al. [124] incorporated anatomical information into the classic CL approach, and Gridach et al. [361] introduced a multi-modal contrastive learning method that combines US video

with sonographers' audio. This integration of additional information mirrors real-world diagnostic processes, where clinicians synthesise various types of data to make informed decisions. Since the ultimate goal of model training is to closely mimic how clinicians make diagnostic decisions, incorporating additional clinical information or leveraging multi-modal approaches could represent a promising direction for future research.

2. **Chapter 5: Explainability of Self-Supervised Representation Learning for Fetal Ultrasound Video** The concept of incorporating clinical understanding through sonographers' visual attention can be extended to various studies with different focus areas that require attention to specific clinical task. The proposed metric, designed based on visual attention, could serve as a standard performance measure for future research. However, in this field, there is currently no unified definition of key terms such as "explainability", "reliability", and "interpretability". Future efforts need to aim at establishing clear, unified definitions for these concepts, which would then allow for the introduction of standardised evaluation metrics across studies.

3. **Chapter 6: Out-of-Clinical-Distribution Detection with a Softmax-Conditioned Variational Autoencoder Boosted Classification Model** I introduced the concept of OCD detection within the context of clinical analysis, which can be applied to a wide range of clinical tasks beyond US. The proposed OCD detection approach can serve as a reference point for future research in this area. Additionally, our methods demonstrated competitive results compared to the diffusion-based DDPM models [342], which are currently a popular topic in the field. However, DDPM comes with significant computational costs. A key consideration moving forward is whether to follow the trend toward increasingly complex models like DDPM or to focus on further refining and advancing traditional, simpler models.

Balancing innovation with computational efficiency is worth considered not only for this research but also for advancements across the broader field of AI.

In the past when I departed, the willows swayed gently. Now as I return, the snow falls thickly

— *The Book of Songs, Minor Odes of the Kingdom, Gathering Vetch.*

References

[1] Timothy G Leighton. "What is ultrasound?" In: *Progress in Biophysics and Molecular Biology*. Vol. 93. 1-3. Elsevier, 2007, pp. 3–83.

[2] Jacqueline K Spencer and Ronald S Adler. "Utility of portable ultrasound in a community in Ghana". In: *Journal of Ultrasound in Medicine*. Vol. 27. 12. Wiley Online Library, 2008, pp. 1735–1743.

[3] J Bamber, N Miller, and M Tristam. "Diagnostic ultrasound". In: *Webb's Physics of Medical Imaging*. Vol. 2. CRC Press, 2012, pp. 351–486.

[4] Jennifer J Kurinczuk et al. "The contribution of congenital anomalies to infant mortality". In: *National Perinatal Epidemiology Unit, University of Oxford*. 2010.

[5] Lior Drukker et al. "Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video". In: *Scientific Reports*. Vol. 11. 1. Nature Publishing Group UK London, 2021, p. 14109.

[6] L Drukker et al. "Clinical workflow of sonographers performing fetal anomaly ultrasound scans: deep-learning-based analysis". In: *Ultrasound in Obstetrics & Gynecology*. Vol. 60. 6. Wiley Online Library, 2022, pp. 759–765.

[7] Harshita Sharma et al. "Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos". In: *Medical Image Analysis*. Vol. 69. Elsevier, 2021, p. 101973.

[8] Haoming Li et al. "CR-Unet: A composite network for ovary and follicle segmentation in ultrasound images". In: *IEEE Journal of Biomedical and Health Informatics*. Vol. 24. 4. IEEE, 2019, pp. 974–983.

[9] Kangning Zhang, Jianbo Jiao, and J Alison Noble. "Fetal Ultrasound Video Representation Learning Using Contrastive Rubik's Cube". In: *Simplifying Medical Ultrasound: 5th International Workshop, ASMUS 2024, Held in Conjunction with MICCAI 2024, Marrakesh, Morocco, October 6, 2024, Proceedings*. Springer Nature. 2024, pp. 187–197.

[10] Kangning Zhang, Jianbo Jiao, and J Alison Noble. "Out-of-Clinical-Distribution Detection with a Softmax-Conditioned Variational Autoencoder Regulariser: Application to Fetal Ultrasound". In: *International Conference on Medical Imaging and Computer-Aided Diagnosis*. Springer. 2024, pp. 558–568.

[11] Tom M Mitchell and Tom M Mitchell. *Machine Learning*. Vol. 1. 9. McGraw-hill New York, 1997.

[12] Christopher M Bishop and Nasser M Nasrabadi. *Pattern Recognition and Machine Learning*. Vol. 4. 4. Springer, 2006.

[13] Ian Goodfellow et al. *Deep Learning*. Vol. 1. 2. MIT press Cambridge, 2016.

[14] Jiajun Zhang, Chengqing Zong, et al. “Deep Neural Networks in Machine Translation: An Overview”. In: *IEEE Intelligent Systems*. Vol. 30. 5. 2015, pp. 16–25.

[15] Li Deng, Dong Yu, et al. “Deep learning: methods and applications”. In: *Foundations and Trends® in Signal Processing*. Vol. 7. 3–4. Now Publishers, Inc., 2014, pp. 197–387.

[16] Geoffrey Hinton et al. “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups”. In: *IEEE Signal Processing Magazine*. Vol. 29. 6. IEEE, 2012, pp. 82–97.

[17] Laura J Brattain et al. “Machine learning for medical ultrasound: status, methods, and future opportunities”. In: *Abdominal Radiology*. Vol. 43. 4. Springer, 2018, pp. 786–799.

[18] A Fourcade and RH Khonsari. “Deep learning in medical image analysis: A third eye for doctors”. In: *Journal of Stomatology, Oral and Maxillofacial Surgery*. Vol. 120. 4. Elsevier, 2019, pp. 279–288.

[19] Fouzia Altaf et al. “Going deep in medical image analysis: concepts, methods, challenges, and future directions”. In: *IEEE Access*. Vol. 7. IEEE, 2019, pp. 99540–99572.

[20] Syed Muhammad Anwar et al. “Medical image analysis using convolutional neural networks: a review”. In: *Journal of Medical Systems*. Vol. 42. Springer, 2018, pp. 1–13.

[21] Yanqi Xu et al. “Understanding differences in applying DETR to natural and medical images”. In: *arXiv preprint arXiv:2405.17677*. 2024.

[22] Phillip M Cheng and Harshawn S Malhi. “Transfer learning with convolutional neural networks for classification of abdominal ultrasound images”. In: *Journal of Digital Imaging*. Vol. 30. 2. Springer, 2017, pp. 234–243.

[23] J Alison Noble and Djamal Boukerroui. “Ultrasound image segmentation: a survey”. In: *IEEE Transactions on Medical Imaging*. Vol. 25. 8. IEEE, 2006, pp. 987–1010.

[24] Richard N Czerwinski, Douglas L Jones, and William D O’Brien. “Detection of lines and boundaries in speckle images-application to medical ultrasound”. In: *IEEE Transactions on Medical Imaging*. Vol. 18. 2. IEEE, 1999, pp. 126–136.

[25] Wei Wei et al. “A deep learning approach for 2D ultrasound and 3D CT/MR image registration in liver tumor ablation”. In: *Computer Methods and Programs in Biomedicine*. Elsevier, 2021, p. 106117.

[26] Lingyun Wu et al. “FUIQA: fetal ultrasound image quality assessment with deep convolutional networks”. In: *IEEE Transactions on Cybernetics*. Vol. 47. 5. IEEE, 2017, pp. 1336–1349.

[27] Qinghua Huang, Fan Zhang, and Xuelong Li. “Machine learning in ultrasound computer-aided diagnostic systems: a survey”. In: *BioMed Research International*. Vol. 2018. Hindawi, 2018.

[28] J Alison Noble, Nassir Navab, and H Becher. “Ultrasonic image analysis and image-guided interventions”. In: *Interface Focus*. Vol. 1. 4. The Royal Society, 2011, pp. 673–685.

[29] Younsu Kim et al. "Low-cost ultrasound thermometry for HIFU therapy using CNN". In: *2018 IEEE International Ultrasonics Symposium (IUS)*. IEEE. 2018, pp. 1–9.

[30] Yuan Gao and J Alison Noble. "Detection and characterization of the fetal heartbeat in free-hand ultrasound sweeps with weakly-supervised two-streams convolutional networks". In: *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*. Springer. 2017, pp. 305–313.

[31] Abhilash R Hareendranathan et al. "Toward automatic diagnosis of hip dysplasia from 2D ultrasound". In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2017, pp. 982–985.

[32] Hariharan Ravishankar et al. "Learning and incorporating shape models for semantic segmentation". In: *Medical Image Computing and Computer Assisted Intervention MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I*. Springer. 2017, pp. 203–211.

[33] Jorden Hetherington et al. "SLIDE: automatic spine level identification system using a deep convolutional neural network". In: *International Journal of Computer Assisted Radiology and Surgery*. Vol. 12. Springer, 2017, pp. 1189–1198.

[34] Franklin Pereira et al. "Automated detection of coarctation of aorta in neonates from two-dimensional echocardiograms". In: *Journal of Medical Imaging*. Vol. 4. 1. Society of Photo-Optical Instrumentation Engineers, 2017, pp. 014502–014502.

[35] Kaizhi Wu, Xi Chen, and Mingyue Ding. "Deep learning based classification of focal liver lesions with contrast-enhanced ultrasound". In: *Optik*. Vol. 125. 15. Elsevier, 2014, pp. 4057–4063.

[36] Yuya Hiramatsu et al. "Automated detection of masses on whole breast volume ultrasound scanner: false positive reduction using deep convolutional neural network". In: *Medical Imaging 2017: Computer-aided Diagnosis*. Vol. 10134. Spie. 2017, pp. 717–722.

[37] Praotasna Sombune et al. "Automated embolic signal detection using deep convolutional neural network". In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2017, pp. 3365–3368.

[38] Shengfeng Liu et al. "Deep learning in medical ultrasound analysis: a review". In: *Engineering*. Vol. 5. 2. Elsevier, 2019, pp. 261–275.

[39] Dinggang Shen, Guorong Wu, and Heung-Il Suk. "Deep learning in medical image analysis". In: *Annual Review of Biomedical Engineering*. Vol. 19. Annual Reviews, 2017, pp. 221–248.

[40] Aiham Taleb et al. "3d self-supervised methods for medical imaging". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 18158–18172.

[41] Longlong Jing and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 2020.

[42] Ashish Jaiswal et al. “A survey on contrastive self-supervised learning”. In: *Technologies*. Vol. 9. 1. MDPI, 2020, p. 2.

[43] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. “Revisiting self-supervised visual representation learning”. In: *2019 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 1920–1929.

[44] Saeed Shurrab and Rehab Duwairi. “Self-supervised learning methods and applications in medical imaging analysis: A survey”. In: *PeerJ Computer Science*. Vol. 8. PeerJ Inc., 2022, e1045.

[45] Wei-Chien Wang et al. “A review of predictive and contrastive self-supervised learning for medical images”. In: *Machine Intelligence Research*. Vol. 20. 4. Springer, 2023, pp. 483–513.

[46] Carl Doersch, Abhinav Gupta, and Alexei A Efros. “Unsupervised visual representation learning by context prediction”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1422–1430.

[47] Mehdi Noroozi and Paolo Favaro. “Unsupervised learning of visual representations by solving jigsaw puzzles”. In: *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*. Springer. 2016, pp. 69–84.

[48] Pengyue Zhang, Fusheng Wang, and Yefeng Zheng. “Self supervised deep representation learning for fine-grained body part recognition”. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2017, pp. 578–582.

[49] Wenjia Bai et al. “Self-supervised learning for cardiac mr image segmentation by anatomical position prediction”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer. 2019, pp. 541–549.

[50] Xinrui Zhuang et al. “Self-supervised feature learning for 3d medical images by playing a rubik’s cube”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer. 2019, pp. 420–428.

[51] Jiuwen Zhu et al. “Rubik’s Cube+: A self-supervised feature learning framework for 3D medical image analysis”. In: *Medical Image Analysis*. Vol. 64. Elsevier, 2020, p. 101746.

[52] Siladittya Manna, Saumik Bhattacharya, and Umapada Pal. “Self-supervised representation learning for detection of ACL tear injury in knee MR videos”. In: *Pattern Recognition Letters*. Vol. 154. Elsevier, 2022, pp. 37–43.

[53] Aiham Taleb et al. “Multimodal self-supervised learning for medical image analysis”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2021, pp. 661–673.

[54] Gonzalo Mena et al. “Learning latent permutations with gumbel-sinkhorn networks”. In: *arXiv preprint arXiv:1802.08665*. 2018.

[55] Yuting Long. "Pneumonia Identification with Self-supervised Learning and Transfer Learning". In: *Application of Intelligent Systems in Multi-modal Information Analytics: 2021 International Conference on Multi-modal Information Analytics (MMIA 2021), Volume 1*. Springer. 2021, pp. 627–635.

[56] Anuja Vats, Marius Pedersen, and Ahmed Mohammed. "A preliminary analysis of self-supervision for wireless capsule endoscopy". In: *2021 9th European Workshop on Visual Information Processing (EUVIP)*. IEEE. 2021, pp. 1–6.

[57] Jianbo Jiao et al. "Self-supervised contrastive video-speech representation learning for ultrasound". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer. 2020, pp. 534–543.

[58] Ting Chen et al. "A simple framework for contrastive learning of visual representations". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1597–1607.

[59] Kaiming He et al. "Momentum contrast for unsupervised visual representation learning". In: *2020 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 9729–9738.

[60] Mathilde Caron et al. "Emerging properties in self-supervised vision transformers". In: *2021 IEEE CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 9650–9660.

[61] Jean-Bastien Grill et al. "Bootstrap your own latent-a new approach to self-supervised learning". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 21271–21284.

[62] Xinlei Chen and Kaiming He. "Exploring simple siamese representation learning". In: *2021 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 15750–15758.

[63] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748*. 2018.

[64] Shekoofeh Azizi et al. "Big self-supervised models advance medical image classification". In: *2021 IEEE CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 3478–3488.

[65] Ozan Ciga, Tony Xu, and Anne Louise Martel. "Self supervised contrastive learning for digital histopathology". In: *Machine Learning with Applications*. Vol. 7. Elsevier, 2022, p. 100198.

[66] Yoni Schirris et al. "DeepSMILE: self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images". In: *arXiv preprint arXiv:2107.09405*. Available, 2021.

[67] Maximilian Ilse, Jakub Tomczak, and Max Welling. "Attention-based deep multiple instance learning". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2127–2136.

[68] Francesca Inglese et al. "MRI-based classification of neuropsychiatric systemic lupus erythematosus patients with self-supervised contrastive learning". In: *Frontiers in Neuroscience*. Vol. 16. Frontiers Media SA, 2022, p. 695888.

[69] Anuroop Sriram et al. “Covid-19 prognosis via self-supervised representation learning and multi-image prediction”. In: *arXiv preprint arXiv:2101.04909*. 2021.

[70] Hari Sowrirajan et al. “Moco pretraining improves representation and transferability of chest x-ray models”. In: *Medical Imaging with Deep Learning*. PMLR. 2021, pp. 728–744.

[71] Yen Nhi Truong Vu et al. “Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation”. In: *Machine Learning for Healthcare Conference*. PMLR. 2021, pp. 755–769.

[72] Yutong Xie et al. “PGL: Prior-guided local self-supervised learning for 3D medical image segmentation”. In: *arXiv preprint arXiv:2011.12640*. 2020.

[73] Jiuwen Zhu et al. “Embedding task knowledge into 3D neural networks via self-supervised learning”. In: *arXiv preprint arXiv:2006.05798*. 2020.

[74] Ming Y Lu, Richard J Chen, and Faisal Mahmood. “Semi-supervised breast cancer histology classification using deep multiple instance learning and contrast predictive coding (conference presentation)”. In: *Medical Imaging 2020: Digital Pathology*. Vol. 11320. SPIE. 2020, 113200J.

[75] Krishna Chaitanya et al. “Contrastive learning of global and local features for medical image segmentation with limited annotations”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 12546–12558.

[76] Ke Yan et al. “SAM: Self-supervised learning of pixel-wise anatomical embeddings in radiological images”. In: *IEEE Transactions on Medical Imaging*. Vol. 41. 10. IEEE, 2022, pp. 2658–2669.

[77] Chenyu You et al. “Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV*. Springer. 2022, pp. 639–652.

[78] Jürgen Schmidhuber. “Deep learning in neural networks: An overview”. In: *Neural Networks*. Vol. 61. Elsevier, 2015, pp. 85–117.

[79] Diederik P Kingma. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114*. 2013.

[80] Ian J Goodfellow et al. “Generative adversarial networks”. In: *arXiv preprint arXiv:1406.2661*. 2014.

[81] Xiao Liu et al. “Self-supervised learning: Generative or contrastive”. In: *IEEE Transactions on Knowledge and Data Engineering*. Vol. 35. 1. IEEE, 2021, pp. 857–876.

[82] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. “Pixel recurrent neural networks”. In: *International Conference on Machine Learning*. PMLR. 2016, pp. 1747–1756.

[83] Aaron Van den Oord et al. “Conditional image generation with pixelcnn decoders”. In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016.

[84] Guanxiong Luo et al. “MRI reconstruction using deep Bayesian estimation”. In: *Magnetic Resonance in Medicine*. Vol. 84. 4. Wiley Online Library, 2020, pp. 2246–2261.

[85] Sucheng Ren et al. “Medical vision generalist: Unifying medical imaging tasks in context”. In: *arXiv preprint arXiv:2406.05565*. 2024.

[86] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. “Density estimation using real nvp”. In: *arXiv preprint arXiv:1605.08803*. 2016.

[87] Durk P Kingma and Prafulla Dhariwal. “Glow: Generative flow with invertible 1x1 convolutions”. In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[88] Vincent van de Schaft and Ruud JG van Sloun. “Ultrasound speckle suppression and denoising using MRI-derived normalizing flow priors”. In: *arXiv preprint arXiv:2112.13110*. 2021.

[89] Mustafa Hajij et al. “Normalizing flow for synthetic medical images generation”. In: *2022 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*. IEEE. 2022, pp. 46–49.

[90] Mangal Prakash et al. “Leveraging self-supervised denoising for image segmentation”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 428–432.

[91] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. “Noise2void-learning denoising from single noisy images”. In: *2019 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2129–2137.

[92] Liang Chen et al. “Self-supervised learning for medical image analysis using image context restoration”. In: *Medical Image Analysis*. Vol. 58. Elsevier, 2019, p. 101539.

[93] Zongwei Zhou et al. “Models genesis: Generic autodidactic models for 3d medical image analysis”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer. 2019, pp. 384–393.

[94] Franco Matzkin et al. “Self-supervised skull reconstruction in brain CT images with decompressive craniectomy”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*. Springer. 2020, pp. 390–399.

[95] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 2015, pp. 234–241.

[96] Pascal Vincent et al. “Extracting and composing robust features with denoising autoencoders”. In: *International Conference on Machine Learning*. 2008, pp. 1096–1103.

[97] Álvaro S Hervella et al. “Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction”. In: *Applied Soft Computing*. Vol. 91. Elsevier, 2020, p. 106210.

[98] Ling Yang et al. “Diffusion models: A comprehensive survey of methods and applications”. In: *ACM Computing Surveys*. Vol. 56. 4. ACM New York, NY, USA, 2023, pp. 1–39.

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

[100] Walter HL Pinaya et al. "Brain imaging generation with latent diffusion models". In: *MICCAI Workshop on Deep Generative Models*. Springer. 2022, pp. 117–126.

[101] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *2022 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10684–10695.

[102] Marina Domínguez et al. "Diffusion as sound propagation: Physics-inspired model for ultrasound image generation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference, Marrakesh, Morocco, October 6–10, 2024, Proceedings, Part IV*. Springer. 2024, pp. 613–623.

[103] Jiashu Xu. "A review of self-supervised learning methods in the field of medical image analysis". In: *International Journal of Image, Graphics and Signal Processing (IJIGSP)*. Vol. 13. 4. 2021, pp. 33–46.

[104] Shih-Cheng Huang et al. "Self-supervised learning for medical image classification: a systematic review and implementation guidelines". In: *NPJ Digital Medicine*. Vol. 6. 1. Nature Publishing Group UK London, 2023, p. 74.

[105] Karen Drukker et al. "Biomedical imaging and analysis through deep learning". In: *Artificial Intelligence in Medicine*. Elsevier, 2021, pp. 49–74.

[106] Feiyu Xu et al. "Explainable AI: A brief survey on history, research areas, approaches and challenges". In: *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, part II 8*. Springer. 2019, pp. 563–574.

[107] Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. "Self-supervised learning in medicine and healthcare". In: *Nature Biomedical Engineering*. Vol. 6. 12. Nature Publishing Group UK London, 2022, pp. 1346–1352.

[108] D Harrison McKnight and Norman L Chervany. "What is trust? A conceptual analysis and an interdisciplinary model". In: 2000.

[109] Keng Siau and Weiyu Wang. "Building trust in artificial intelligence, machine learning, and robotics". In: *Cutter Business Technology Journal*. Vol. 31. 2. 2018, p. 47.

[110] Caroline Jones, James Thornton, and Jeremy C Wyatt. "Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability". In: *Medical Law Review*. Vol. 31. 4. Oxford University Press, 2023, pp. 501–520.

[111] Onora O'neill. "Linking trust to trustworthiness". In: *International Journal of Philosophical Studies*. Vol. 26. 2. Taylor & Francis, 2018, pp. 293–300.

[112] David Archard et al. *Reading Onora O'Neill*. Routledge London/New York, NY, 2013.

[113] Luciano Floridi et al. "AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations". In: *Minds and Machines*. Vol. 28. Springer, 2018, pp. 689–707.

[114] Slavko Splichal. "In data we (don't) trust: The public adrift in data-driven public opinion models". In: *Big Data & Society*. Vol. 9. 1. SAGE Publications Sage UK: London, England, 2022, p. 20539517221097319.

[115] Karoline Reinhardt. "Trust and trustworthiness in AI ethics". In: *AI and Ethics*. Vol. 3. 3. Springer, 2023, pp. 735–744.

[116] Merriam-Webster Dictionary. "Merriam-webster". In: *On-line at http://www. mw. com/home. htm*. Vol. 8. 2. 2002, p. 23.

[117] Peter J Huber. "The 1972 wald lecture robust statistics: A review". In: *The Annals of Mathematical Statistics*. JSTOR, 1972, pp. 1041–1067.

[118] Houssem Ben Braiek and Foutse Khomh. "Machine learning robustness: A primer". In: *Trustworthy AI in Medical Imaging*. Elsevier, 2025, pp. 37–71.

[119] Timo Freiesleben and Thomas Grote. "Beyond generalization: a theory of robustness in machine learning". In: *Synthese*. Vol. 202. 4. Springer, 2023, p. 109.

[120] Qinkai Zheng et al. "Graph robustness benchmark: Benchmarking the adversarial robustness of graph machine learning". In: *arXiv preprint arXiv:2111.04314*. 2021.

[121] Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, et al. "Robustness and explainability of artificial intelligence". In: *Publications Office of the European Union*. Vol. 207. 2020, p. 2020.

[122] Alan Balendran et al. "A scoping review of robustness concepts for machine learning in healthcare". In: *NPJ Digital Medicine*. Vol. 8. 1. Nature Publishing Group UK London, 2025, p. 38.

[123] Samuel G Finlayson et al. "Adversarial attacks against medical deep learning systems". In: *arXiv preprint arXiv:1804.05296*. 2018.

[124] Zeyu Fu et al. "Anatomy-aware contrastive representation learning for fetal ultrasound". In: *Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*. Springer. 2022, pp. 422–436.

[125] Aritra Ghosh and Andrew Lan. "Contrastive learning improves model robustness under label noise". In: *2021 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 2703–2708.

[126] Ching-Yao Chuang et al. "Robust contrastive learning against noisy views". In: *2022 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 16670–16681.

[127] Zhuang Qian et al. "A survey of robust adversarial training in pattern recognition: Fundamental, theory, and methodologies". In: *Pattern Recognition*. Vol. 131. Elsevier, 2022, p. 108889.

[128] Joseph Y Halpern and Judea Pearl. "Causes and explanations: A structural-model approach. Part I: Causes". In: *The British Journal for the Philosophy of Science*. The University of Chicago Press, 2005.

[129] Adrienne Héritier. "Causal explanation". In: *Approaches and Methodologies in the Social Sciences*. Vol. 61. Cambridge University Press Cambridge, 2008.

[130] IEC ISO. *AWI TS 29119-11: Software and systems engineering-software testing-part 11: Testing of AI systems*. 2020.

[131] Sajid Ali et al. “Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence”. In: *Information Fusion*. Vol. 99. Elsevier, 2023, p. 101805.

[132] Arun Das and Paul Rad. “Opportunities and challenges in explainable artificial intelligence (xai): A survey”. In: *arXiv preprint arXiv:2006.11371*. 2020.

[133] Mengnan Du, Ninghao Liu, and Xia Hu. “Techniques for interpretable machine learning”. In: *Communications of the ACM*. Vol. 63. 1. ACM New York, NY, USA, 2019, pp. 68–77.

[134] Grégoire Montavon et al. “Explaining nonlinear classification decisions with deep taylor decomposition”. In: *Pattern Recognition*. Vol. 65. Elsevier, 2017, pp. 211–222.

[135] Been Kim. “Interactive and interpretable machine learning models for human machine collaboration”. PhD thesis. Massachusetts Institute of Technology, 2015.

[136] Rich Caruana et al. “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission”. In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, pp. 1721–1730.

[137] Erico Tjoa and Cuntai Guan. “A survey on explainable artificial intelligence (xai): Toward medical xai”. In: *IEEE Transactions on Neural Networks and Learning Systems*. IEEE, 2020.

[138] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. “Machine learning interpretability: A survey on methods and metrics”. In: *Electronics*. Vol. 8. 8. Multidisciplinary Digital Publishing Institute, 2019, p. 832.

[139] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608*. 2017.

[140] Selim Temizer et al. “Collision avoidance for unmanned aircraft using Markov decision processes”. In: *AIAA Guidance, Navigation, and Control Conference*. 2010, p. 8040.

[141] William R Swartout. “XPLAIN: A system for creating and explaining expert consulting programs”. In: *Artificial Intelligence*. Vol. 21. 3. Elsevier, 1983, pp. 285–325.

[142] Bruce G Buchanan and Edward H Shortliffe. “Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project”. In: CUMINCAD, 1984.

[143] Johanna D Moore and William R Swartout. *Explanation in expert systemss: A survey*. Tech. rep. UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST, 1988.

[144] Robert Andrews, Joachim Diederich, and Alan B Tickle. “Survey and critique of techniques for extracting rules from trained artificial neural networks”. In: *Knowledge-based Systems*. Vol. 8. 6. Elsevier, 1995, pp. 373–389.

[145] Henriette Cramer et al. “The effects of transparency on trust in and acceptance of a content-based art recommender”. In: *User Modeling and User-adapted interaction*. Vol. 18. 5. Springer, 2008, p. 455.

[146] Gary M Olson and Judith S Olson. *Computer-supported Cooperative Work*. John Wiley & Sons, Inc., 2007.

[147] Michael Van Lent, William Fisher, and Michael Mancuso. "An explainable artificial intelligence system for small-unit tactical behavior". In: *Proceedings of the National Conference on Artificial Intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. 2004, pp. 900–907.

[148] Amina Adadi and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE Access*. Vol. 6. IEEE, 2018, pp. 52138–52160.

[149] Ashraf Abdul et al. "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018, pp. 1–18.

[150] Julia Angwin et al. "Machine bias". In: *ProPublica, May*. Vol. 23. 2016. 2016, pp. 139–159.

[151] Nicolas Papernot et al. "The limitations of deep learning in adversarial settings". In: *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE. 2016, pp. 372–387.

[152] Christian Szegedy et al. "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199*. 2013.

[153] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572*. 2014.

[154] Nicholas Carlini and David Wagner. "Adversarial examples are not easily detected: Bypassing ten detection methods". In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. 2017, pp. 3–14.

[155] Aleksander Madry et al. "Towards deep learning models resistant to adversarial attacks". In: *arXiv preprint arXiv:1706.06083*. 2017.

[156] Bryce Goodman and Seth Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"". In: *AI Magazine*. Vol. 38. 3. 2017, pp. 50–57.

[157] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. "The false hope of current approaches to explainable artificial intelligence in health care". In: *The Lancet Digital Health*. Vol. 3. 11. Elsevier, 2021, e745–e750.

[158] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence*. Vol. 1. 5. Nature Publishing Group UK London, 2019, pp. 206–215.

[159] Maxime Kayser et al. "Fool Me Once? Contrasting Textual and Visual Explanations in a Clinical Decision-Support Setting". In: *arXiv preprint arXiv:2410.12284*. 2024.

[160] Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *2014 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 580–587.

[161] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. 2012, pp. 1097–1105.

[162] David Bau et al. “Network dissection: Quantifying interpretability of deep visual representations”. In: *2017 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6541–6549.

[163] Matthew D Zeiler and Rob Fergus. “Visualizing and understanding convolutional networks”. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer. 2014, pp. 818–833.

[164] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034*. 2013.

[165] Ruigang Fu et al. “Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs”. In: *arXiv preprint arXiv:2008.02312*. 2020.

[166] Jost Tobias Springenberg et al. “Striving for simplicity: The all convolutional net”. In: *arXiv preprint arXiv:1412.6806*. 2014.

[167] Aravindh Mahendran and Andrea Vedaldi. “Salient deconvolutional networks”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer. 2016, pp. 120–135.

[168] Sebastian Bach et al. “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation”. In: *PloS One*. Vol. 10. 7. Public Library of Science, 2015, e0130140.

[169] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic attribution for deep networks”. In: *arXiv preprint arXiv:1703.01365*. 2017.

[170] Aditya Chattopadhay et al. “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks”. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018, pp. 839–847.

[171] Harish Guruprasad Ramaswamy et al. “Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization”. In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2020, pp. 983–991.

[172] Daniel Omeiza et al. “Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models”. In: *arXiv preprint arXiv:1908.01224*. 2019.

[173] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626.

[174] Bolei Zhou et al. “Learning deep features for discriminative localization”. In: *2016 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2921–2929.

[175] Daniel Smilkov et al. “Smoothgrad: removing noise by adding noise”. In: *arXiv preprint arXiv:1706.03825*. 2017.

[176] Been Kim et al. “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2668–2677.

[177] Jacob Pfau et al. “Robust Semantic Interpretability: Revisiting Concept Activation Vectors”. In: *arXiv preprint arXiv:2104.02768*. 2021.

[178] Tameem Adel, Zoubin Ghahramani, and Adrian Weller. “Discovering interpretable representations for both deep generative and discriminative models”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 50–59.

[179] Quanshi Zhang et al. “Interpreting cnns via decision trees”. In: *2019 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 6261–6270.

[180] Rishabh Agarwal et al. “Neural additive models: Interpretable machine learning with neural nets”. In: *arXiv preprint arXiv:2004.13912*. 2020.

[181] Cambridge Dictionary. “English dictionary”. In: *Cambridge University Press. Retrieved June*. Vol. 14. 2018, p. 2018.

[182] Oxford English Dictionary. “Oxford english dictionary”. In: *Simpson, JA & Weiner, ESC.–1989*. 1993.

[183] National Institute of Standards and Technology. *Reliability*. `https://csrc.nist.gov/glossary/term/reliability`. Accessed: April 4, 2025.

[184] Dario Amodei et al. “Concrete problems in AI safety”. In: *arXiv preprint arXiv:1606.06565*. 2016.

[185] Pranav Rajpurkar et al. “AI in health care: The hope, the hype, the promise, the peril”. In: *The Lancet Digital Health*. Vol. 4. 1. Elsevier, 2022, e2–e3.

[186] Shiyu Liang, Yixuan Li, and R Srikant. “Enhanced Out-of-Distribution Detection in Deep Neural Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 2022.

[187] Te Han, Yan-Fu Li, and Min Qian. “A hybrid generalization network for intelligent fault diagnosis of rotating machinery under unseen working conditions”. In: *IEEE Transactions on Instrumentation and Measurement*. Vol. 70. IEEE, 2021, pp. 1–11.

[188] Taotao Zhou et al. “An uncertainty-informed framework for trustworthy fault diagnosis in safety-critical applications”. In: *Reliability Engineering & System Safety*. Vol. 229. Elsevier, 2023, p. 108865.

[189] Anqi He and Xiaoning Jin. “Deep variational autoencoder classifier for intelligent fault diagnosis adaptive to unseen fault categories”. In: *IEEE Transactions on Reliability*. Vol. 70. 4. IEEE, 2021, pp. 1581–1595.

[190] Jingkang Yang et al. “Generalized out-of-distribution detection: A survey”. In: *International Journal of Computer Vision*. Springer, 2024, pp. 1–28.

[191] Michel Cannarsa. “Ethics guidelines for trustworthy AI”. In: *The Cambridge handbook of lawyering in the digital age*. Cambridge University Press Cambridge, UK, 2021, pp. 283–297.

[192] Eric Topol. “Deep medicine: how artificial intelligence can make healthcare human again”. In: *Basic Books*. 2019.

[193] Jingkang Yang et al. “Generalized out-of-distribution detection: A survey”. In: *arXiv preprint arXiv:2110.11334*. 2021.

[194] Dan Hendrycks and Kevin Gimpel. “A baseline for detecting misclassified and out-of-distribution examples in neural networks”. In: *arXiv preprint arXiv:1610.02136*. 2016.

[195] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. “Enhancing the reliability of out-of-distribution image detection in neural networks”. In: *arXiv preprint arXiv:1706.02690*. 2017.

[196] Yiyou Sun, Chuan Guo, and Yixuan Li. “React: Out-of-distribution detection with rectified activations”. In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 144–157.

[197] Xin Dong et al. “Neural mean discrepancy for efficient out-of-distribution detection”. In: *2022 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 19217–19227.

[198] Yiyou Sun and Yixuan Li. “DICE: Leveraging sparsification for out-of-distribution detection”. In: *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*. Springer. 2022, pp. 691–708.

[199] Terrance DeVries and Graham W Taylor. “Learning confidence for out-of-distribution detection in neural networks”. In: *arXiv preprint arXiv:1802.04865*. 2018.

[200] Yezhen Wang et al. “Energy-based open-world uncertainty modeling for confidence calibration”. In: *2021 IEEE CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 9302–9311.

[201] Julian Bitterwolf, Alexander Meinke, and Matthias Hein. “Certifiably adversarially robust detection of out-of-distribution data”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 16085–16095.

[202] Sven Gowal et al. “On the effectiveness of interval bound propagation for training verifiably robust models”. In: *arXiv preprint arXiv:1810.12715*. 2018.

[203] Jiefeng Chen et al. “Robust out-of-distribution detection for neural networks”. In: *arXiv preprint arXiv:2003.09711*. 2020.

[204] Jiefeng Chen et al. “Atom: Robustifying out-of-distribution detection using outlier mining”. In: *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*. Springer. 2021, pp. 430–445.

[205] Sunil Thulasidasan et al. “On mixup training: Improved calibration and predictive uncertainty for deep neural networks”. In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[206] Hongyi Zhang. “mixup: Beyond empirical risk minimization”. In: *arXiv preprint arXiv:1710.09412*. 2017.

[207] Dan Hendrycks et al. “Augmix: A simple data processing method to improve robustness and uncertainty”. In: *arXiv preprint arXiv:1912.02781*. 2019.

[208] Dan Hendrycks et al. “Pixmix: Dreamlike pictures comprehensively improve safety measures”. In: *2022 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 16783–16792.

[209] Jinsol Lee and Ghassan AlRegib. "Gradients as a measure of uncertainty in neural networks". In: *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2020, pp. 2416–2420.

[210] Rui Huang, Andrew Geng, and Yixuan Li. "On the importance of gradients for detecting distributional shifts in the wild". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 677–689.

[211] MJ Desforges, PJ Jacob, and JE Cooper. "Applications of probability density estimation to the detection of abnormal conditions in engineering". In: *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*. Vol. 212. 8. SAGE Publications Sage UK: London, England, 1998, pp. 687–703.

[212] Kimin Lee et al. "A simple unified framework for detecting out-of-distribution samples and adversarial attacks". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[213] Mohammad Sabokrou et al. "Adversarially learned one-class classifier for novelty detection". In: *2018 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 3379–3388.

[214] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. "Generative probabilistic novelty detection with adversarial autoencoders". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[215] Lucas Deecke et al. "Image anomaly detection with generative adversarial networks". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*. Springer. 2019, pp. 3–17.

[216] Ev Zisselman and Aviv Tamar. "Deep residual flow for out of distribution detection". In: *2020 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 13994–14003.

[217] Weitang Liu et al. "Energy-based out-of-distribution detection". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 21464–21475.

[218] Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. "Mood: Multi-level out-of-distribution detection". In: *2021 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 15313–15323.

[219] Engkarat Techapanurak, Masanori Suganuma, and Takayuki Okatani. "Hyperparameter-free out-of-distribution detection using cosine similarity". In: *Proceedings of the Asian Conference on Computer Vision*. 2020.

[220] Xingyu Chen et al. "A boundary based out-of-distribution classifier for generalized zero-shot learning". In: *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV*. Springer. 2020, pp. 572–588.

[221] Alireza Zaeemzadeh et al. "Out-of-distribution detection using union of 1-dimensional subspaces". In: *2021 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 9452–9461.

[222] Jie Ren et al. "A simple fix to mahalanobis distance for improving near-ood detection". In: *arXiv preprint arXiv:2106.09022*. 2021.

[223] Haiwen Huang et al. “Feature space singularity for out-of-distribution detection”. In: *arXiv preprint arXiv:2011.14654*. 2020.

[224] Yiyou Sun et al. “Out-of-distribution detection with deep nearest neighbors”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 20827–20840.

[225] Eduardo Dadalto Camara Gomes et al. “Igeood: An information geometry approach to out-of-distribution detection”. In: *arXiv preprint arXiv:2203.07798*. 2022.

[226] João E Strapasson, Julianna Pinele, and Sueli IR Costa. “Clustering using the Fisher-Rao distance”. In: *2016 IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*. IEEE. 2016, pp. 1–5.

[227] Taylor Denouden et al. “Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance”. In: *arXiv preprint arXiv:1812.02765*. 2018.

[228] Yibo Zhou. “Rethinking reconstruction autoencoder-based out-of-distribution detection”. In: *2022 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 7379–7387.

[229] Yijun Yang, Ruiyuan Gao, and Qiang Xu. “Out-of-distribution detection with semantic mismatch under masking”. In: *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*. Springer. 2022, pp. 373–390.

[230] Wenyu Jiang et al. “Read: Aggregating reconstruction error into out-of-distribution detection”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 12. 2023, pp. 14910–14918.

[231] Donna Kirwan. “NHS Fetal Anomaly Screening Programme”. In: *National Standards and Guidance for England*. Vol. 18. 0. 2010.

[232] Delwyn Nicholls, Linda Sweet, and Jon Hyett. “Psychomotor skills in medical ultrasound imaging: an analysis of the core skill set”. In: *Journal of Ultrasound in Medicine*. Vol. 33. 8. Wiley Online Library, 2014, pp. 1349–1352.

[233] Mohammad Ali Maraci et al. “Searching for structures of interest in an ultrasound video sequence”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2014, pp. 133–140.

[234] Lisa M Gangarosa. “The practice of ultrasound: A step-by-step guide to abdominal scanning”. In: *Gastroenterology*. Vol. 129. 4. Elsevier, 2005, p. 1357.

[235] Nigel Thomson and Audrey Paterson. “Sonographer registration in the United Kingdom–a review of the current situation”. In: *Ultrasound*. Vol. 22. 1. SAGE Publications Sage UK: London, England, 2014, pp. 52–56.

[236] Joet al Moriarty, Mary Baginsky, and Jill Manthorpe. “Literature review of roles and issues within the social work profession in England”. In: *Professional Standards Authority, London*. 2015.

[237] Laurent Julien Salomon et al. “Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan”. In: *Ultrasound in Obstetrics & Gynecology*. Vol. 37. 1. Wiley Online Library, 2011, pp. 116–126.

[238] Gary L Darmstadt et al. “60 million non-facility births: who can deliver in community settings to reduce intrapartum-related deaths?” In: *International Journal of Gynecology & Obstetrics*. Vol. 107. Elsevier, 2009, S89–S112.

[239] Ling Zhang et al. "Intelligent scanning: Automated standard plane selection and biometric measurement of early gestational sac in routine ultrasound examination". In: *Medical Physics*. Vol. 39. 8. Wiley Online Library, 2012, pp. 5015–5027.

[240] Dong Ni et al. "Standard plane localization in ultrasound by radial component model and selective search". In: *Ultrasound in Medicine & Biology*. Vol. 40. 11. Elsevier, 2014, pp. 2728–2742.

[241] A Abuhamad et al. "Automated retrieval of standard diagnostic fetal cardiac ultrasound planes in the second trimester of pregnancy: a prospective evaluation of software". In: *Ultrasound in Obstetrics & Gynecology*. Vol. 31. 1. Wiley Online Library, 2008, pp. 30–36.

[242] Bahbibi Rahmatullah, Aris T Papageorghiou, and J Alison Noble. "Integration of local and global features for anatomical object detection in ultrasound". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part III 15*. Springer. 2012, pp. 402–409.

[243] Dong Ni et al. "Selective search and sequential detection for standard plane localization in ultrasound". In: *Abdominal Imaging. Computation and Clinical Applications: 5th International Workshop, Held in Conjunction with MICCAI 2013, Nagoya, Japan, September 22, 2013. Proceedings 5*. Springer. 2013, pp. 203–211.

[244] Michal Sofka et al. "Automatic detection and measurement of structures in fetal head ultrasound volumes using sequential estimation and integrated detection network (IDN)". In: *IEEE Transactions on Medical Imaging*. Vol. 33. 5. IEEE, 2014, pp. 1054–1070.

[245] Hao Chen et al. "Fetal abdominal standard plane localization through representation learning with knowledge transfer". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2014, pp. 125–132.

[246] J Alison Noble. *Reflections on ultrasound image analysis*. 2016.

[247] I Sarris et al. "Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21st Project". In: *BJOG: An International Journal of Obstetrics & Gynaecology*. Vol. 120. Wiley Online Library, 2013, pp. 33–37.

[248] Richard Droste et al. "Ultrasound image representation learning by modeling sonographer visual attention". In: *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*. Springer. 2019, pp. 592–604.

[249] Harshita Sharma et al. "Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2019, pp. 987–990.

[250] Yuan Gao and J Alison Noble. "Learning and understanding deep spatio-temporal representations from free-hand fetal ultrasound sweeps". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V*. Springer. 2019, pp. 299–308.

[251] Lok Hin Lee, Yuan Gao, and J Alison Noble. "Principled ultrasound data augmentation for classification of standard planes". In: *International Conference on Information Processing in Medical Imaging*. Springer. 2021, pp. 729–741.

[252] Yifan Cai et al. "SonoEyeNet: Standardized fetal ultrasound plane detection informed by eye tracking". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2018, pp. 1475–1478.

[253] Yifan Cai et al. "Spatio-temporal visual attention modelling of standard biometry plane-finding navigation". In: *Medical Image Analysis*. Vol. 65. Elsevier, 2020, p. 101762.

[254] Yipei Wang et al. "Task model-specific operator skill assessment in routine fetal ultrasound scanning". In: *International Journal of Computer Assisted Radiology and Surgery*. Vol. 17. 8. Springer, 2022, pp. 1437–1444.

[255] Mohammad Alsharid et al. "Captioning ultrasound images automatically". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer. 2019, pp. 338–346.

[256] Robail Yasrab et al. "End-to-end first trimester fetal ultrasound video automated crl and nt segmentation". In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2022, pp. 1–5.

[257] Elizaveta Savochkina et al. "First Trimester Video Saliency Prediction Using Clstmu-Net with Stochastic Augmentation". In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2022, pp. 1–4.

[258] Hsin-Ying Lee et al. "Unsupervised representation learning by sorting sequences". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017.

[259] Demetrios L Economides and Jeffrey M Braithwaite. "First trimester ultrasonographic diagnosis of fetal structural abnormalities in a low risk population". In: *BJOG: An International Journal of Obstetrics & Gynaecology*. Vol. 105. 1. Wiley Online Library, 1998, pp. 53–57.

[260] Robail Yasrab et al. "A machine learning method for automated description and workflow analysis of first trimester ultrasound scans". In: *IEEE Transactions on Medical Imaging*. Vol. 42. 5. IEEE, 2022, pp. 1301–1313.

[261] NJ Dudley and E Chapman. "The importance of quality management in fetal measurement". In: *Ultrasound in Obstetrics & Gynecology*. Vol. 19. 2. Wiley Online Library, 2002, pp. 190–196.

[262] Madeline C Schiappa, Yogesh S Rawat, and Mubarak Shah. "Self-supervised learning for videos: A survey". In: *ACM Computing Surveys*. Vol. 55. 13s. ACM New York, NY, 2023, pp. 1–37.

[263] Longlong Jing et al. "Self-supervised spatiotemporal feature learning via video rotation prediction". In: *arXiv preprint arXiv:1811.11387*. 2018.

[264] Yixiong Chen et al. "USCL: pretraining deep ultrasound image diagnosis model through video contrastive representation learning". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*. Springer. 2021, pp. 627–637.

[265] Chunhui Zhang et al. “HiCo: hierarchical contrastive learning for ultrasound video model pretraining”. In: *Proceedings of the Asian Conference on Computer Vision*. 2022, pp. 229–246.

[266] Veenu Rani et al. “Self-supervised learning: A succinct review”. In: *Archives of Computational Methods in Engineering*. Vol. 30. 4. Springer, 2023, pp. 2761–2775.

[267] Yutong Bai et al. “Can temporal information help with contrastive self-supervised learning?” In: *arXiv preprint arXiv:2011.13046*. 2020.

[268] Rui Qian et al. “Enhancing self-supervised video representation learning via multi-level feature optimization”. In: *2021 IEEE CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 7990–8001.

[269] Fahad Shamshad et al. “Transformers in medical imaging: A survey”. In: *Medical Image Analysis*. Elsevier, 2023, p. 102802.

[270] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929*. 2020.

[271] Jun Li et al. “Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives”. In: *Medical Image Analysis*. Vol. 85. Elsevier, 2023, p. 102762.

[272] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *2021 IEEE CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 10012–10022.

[273] Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[274] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[275] Wang Wenxuan et al. “Transbts: Multimodal brain tumor segmentation using transformer”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I*. 2021, pp. 109–119.

[276] Yading Yuan. “Automatic skin lesion segmentation with fully convolutional-deconvolutional networks”. In: *arXiv preprint arXiv:1703.05165*. 2017.

[277] Arnab Kumar Mondal et al. “xViTCOS: explainable vision transformer based COVID-19 screening using radiography”. In: *IEEE Journal of Translational Engineering in Health and Medicine*. Vol. 10. IEEE, 2021, pp. 1–10.

[278] Shuang Yu et al. “Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*. Springer. 2021, pp. 45–54.

[279] Shijie Liu et al. “Transformer for polyp detection”. In: *arXiv preprint arXiv:2111.07918*. 2021.

[280] Rong Tao and Guoyan Zheng. "Spine-transformers: Vertebra detection and localization in arbitrary field-of-view spine ct with transformers". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer. 2021, pp. 93–103.

[281] Chun-Mei Feng et al. "Accelerated multi-modal MR imaging with transformers". In: *arXiv preprint arXiv:2106.14248*. Vol. 3. 2021.

[282] Ce Wang et al. "DuDoTrans: dual-domain transformer provides more attention for sinogram restoration in sparse-view CT reconstruction". In: *arXiv preprint arXiv:2111.10790*. 2021.

[283] Yucheng Tang et al. "Self-supervised pre-training of swin transformers for 3d medical image analysis". In: *2022 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 20730–20740.

[284] Yu-Qi Yang et al. "Swin3d: A pretrained transformer backbone for 3d indoor scene understanding". In: *arXiv preprint arXiv:2304.06906*. 2023.

[285] Bill Waggener and William N Waggener. *Pulse Code Modulation Techniques*. Springer Science & Business Media, 1995.

[286] Yuning You et al. "Graph contrastive learning with augmentations". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 5812–5823.

[287] Kaiming He et al. "Deep residual learning for image recognition". In: *2016 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[288] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101*. 2017.

[289] Fatemeh Haghighi et al. "DiRA: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis". In: *2022 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 20824–20834.

[290] Christian F Baumgartner et al. "SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound". In: *IEEE Transactions on Medical Imaging*. Vol. 36. 11. IEEE, 2017, pp. 2204–2215.

[291] Robert F Woolson. "Wilcoxon signed-rank test". In: *Encyclopedia of Biostatistics*. Vol. 8. Wiley Online Library, 2005.

[292] Fei Jiang et al. "Artificial intelligence in healthcare: past, present and future". In: *Stroke and Vascular Neurology*. Vol. 2. 4. BMJ Specialist Journals, 2017.

[293] World Health Organization et al. "Working for health and growth: investing in the health workforce". In: World Health Organization, 2016.

[294] Junaid Bajwa et al. "Artificial intelligence in healthcare: transforming the practice of medicine". In: *Future Healthcare Journal*. Vol. 8. 2. Royal College of Physicians, 2021, e188.

[295] Rahul C Deo. "Machine learning in medicine". In: *Circulation*. Vol. 132. 20. Am Heart Assoc, 2015, pp. 1920–1930.

[296] Alejandro Barredo Arrieta et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. In: *Information Fusion*. Vol. 58. Elsevier, 2020, pp. 82–115.

[297] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. “Why a right to explanation of automated decision-making does not exist in the general data protection regulation”. In: *International Data Privacy Law*. Vol. 7. 2. Oxford University Press, 2017, pp. 76–99.

[298] Quan-shi Zhang and Song-Chun Zhu. “Visual interpretability for deep learning: a survey”. In: *Frontiers of Information Technology & Electronic Engineering*. Vol. 19. 1. Springer, 2018, pp. 27–39.

[299] Quanshi Zhang et al. “Interpreting cnn knowledge via an explanatory graph”. In: *arXiv preprint arXiv:1708.01785*. 2017.

[300] Tim Miller. “Explanation in artificial intelligence: Insights from the social sciences”. In: *Artificial Intelligence*. Vol. 267. Elsevier, 2019, pp. 1–38.

[301] Zachary C Lipton. “The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.” In: *Queue*. Vol. 16. 3. ACM New York, NY, USA, 2018, pp. 31–57.

[302] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. “Methods for interpreting and understanding deep neural networks”. In: *Digital Signal Processing*. Vol. 73. Elsevier, 2018, pp. 1–15.

[303] Mohammed Hassanin et al. “Visual attention methods in deep learning: An in-depth survey”. In: *arXiv preprint arXiv:2204.07756*. 2022.

[304] Chia-Chien Wu, Farahnaz Ahmed Wick, and Marc Pomplun. “Guidance of visual attention by semantic information in real-world scenes”. In: *Frontiers in Psychology*. Vol. 5. Frontiers, 2014, p. 54.

[305] Fei Yan et al. “Review of visual saliency prediction: Development process from neurobiological basis to deep models”. In: *Applied Sciences*. Vol. 12. 1. MDPI, 2021, p. 309.

[306] Xun Huang et al. “Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks”. In: *2015 IEEE CVF International Conference on Computer Vision (ICCV)*. 2015, pp. 262–270.

[307] Ian L Dryden and Kanti V Mardia. *Statistical Shape Analysis: with Applications in R*. Vol. 995. John Wiley & Sons, 2016.

[308] Y Zhan et al. “Robust Multi-Landmark Detection Based on Information Theoretic Scheduling”. In: *Medical Image Recognition, Segmentation and Parsing*. Elsevier, 2016, pp. 45–70.

[309] Patrick Brass et al. “Ultrasound guidance versus anatomical landmarks for internal jugular vein catheterization”. In: *Cochrane Database of Systematic Reviews*. 1. John Wiley & Sons, Ltd, 2015.

[310] Jun Zhang et al. “Detecting anatomical landmarks for fast Alzheimer’s disease diagnosis”. In: *IEEE Transactions on Medical Imaging*. Vol. 35. 12. IEEE, 2016, pp. 2524–2533.

[311] Ching-Wei Wang et al. “A benchmark for comparison of dental radiography analysis algorithms”. In: *Medical Image Analysis*. Vol. 31. Elsevier, 2016, pp. 63–76.

[312] Mustafa Elattar et al. “Automatic aortic root landmark detection in CTA images for preprocedural planning of transcatheter aortic valve implantation”. In: *The International Journal of Cardiovascular Imaging*. Vol. 32. Springer, 2016, pp. 501–511.

[313] Florin C Ghesu et al. “An artificial agent for anatomical landmark detection in medical images”. In: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III 19*. Springer. 2016, pp. 229–237.

[314] Yefeng Zheng et al. “3D deep learning for efficient and robust landmark detection in volumetric data”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*. Springer. 2015, pp. 565–572.

[315] Sercan Ö Arık, Bulat Ibragimov, and Lei Xing. “Fully automated quantitative cephalometry using convolutional neural networks”. In: *Journal of Medical Imaging*. Vol. 4. 1. Society of Photo-Optical Instrumentation Engineers, 2017, pp. 014501–014501.

[316] Christian Payer et al. “Regressing heatmaps for multiple landmark localization using CNNs”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*. Springer. 2016, pp. 230–238.

[317] Christian Payer et al. “Integrating spatial configuration into heatmap regression based CNNs for landmark localization”. In: *Medical Image Analysis*. Vol. 54. Elsevier, 2019, pp. 207–219.

[318] Lior Rokach and Oded Maimon. “Clustering methods”. In: *Data Mining and Knowledge Discovery Handbook*. Springer, 2005, pp. 321–352.

[319] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. “BIRCH: an efficient data clustering method for very large databases”. In: *ACM Sigmod Record*. Vol. 25. 2. ACM New York, NY, USA, 1996, pp. 103–114.

[320] John A Hartigan and Manchek A Wong. “Algorithm AS 136: A k-means clustering algorithm”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)*. Vol. 28. 1. JSTOR, 1979, pp. 100–108.

[321] Douglas A Reynolds et al. “Gaussian mixture models.” In: *Encyclopedia of Biometrics*. Vol. 741. 659-663. Berlin, Springer, 2009.

[322] Kamran Khan et al. “DBSCAN: Past, present and future”. In: *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*. IEEE. 2014, pp. 232–238.

[323] Qing He et al. “Clustering in extreme learning machine feature space”. In: *Neurocomputing*. Vol. 128. Elsevier, 2014, pp. 88–95.

[324] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. "Self-supervised video representation learning by pace prediction". In: *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII*. Springer. 2020, pp. 504–521.

[325] Abraham Bookstein, Vladimir A Kulyukin, and Timo Raita. "Generalized hamming distance". In: *Information Retrieval*. Vol. 5. Springer, 2002, pp. 353–375.

[326] Richard Droste et al. "Discovering Salient Anatomical Landmarks by Predicting Human Gaze". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1711–1714.

[327] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics*. Vol. 20. Elsevier, 1987, pp. 53–65.

[328] Sergey Zagoruyko and Nikos Komodakis. "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer". In: *arXiv preprint arXiv:1612.03928*. 2016.

[329] Pramila P Shinde and Seema Shah. "A review of machine learning and deep learning applications". In: *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE. 2018, pp. 1–6.

[330] Justin Ker et al. "Deep learning applications in medical image analysis". In: *IEEE Access*. Vol. 6. IEEE, 2017, pp. 9375–9389.

[331] Raj Madhavan et al. "Toward trustworthy and responsible artificial intelligence policy development". In: *IEEE Intelligent Systems*. Vol. 35. 5. IEEE, 2020, pp. 103–108.

[332] Sara Narteni et al. "From explainable to reliable artificial intelligence". In: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer. 2021, pp. 255–273.

[333] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 295–305.

[334] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature*. Vol. 521. 7553. Nature Publishing Group UK London, 2015, pp. 436–444.

[335] Haoqi Wang et al. "Vim: Out-of-distribution with virtual-logit matching". In: *2022 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 4921–4930.

[336] Kate Highnam et al. "Beth dataset: Real cybersecurity data for unsupervised anomaly detection research". In: *CEUR Workshop Proceedings*. Vol. 3095. 2021, pp. 1–12.

[337] Daniel Bogdoll, Maximilian Nitsche, and J Marius Zöllner. "Anomaly detection in autonomous driving: A survey". In: *2022 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 4488–4499.

[338] Tianshi Cao et al. "A benchmark of medical out of distribution detection". In: *arXiv preprint arXiv:2007.04250*. 2020.

[339] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. "Deep learning for medical image processing: Overview, challenges and the future". In: *Classification in BioApps: Automation of Decision Making*. Springer, 2018, pp. 323–350.

[340] UK National Screening Committee et al. "Fetal anomaly screening: programme handbook". In: *London: Public Health England*. 2015.

[341] Dong Gong et al. "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection". In: *2019 IEEE CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 1705–1714.

[342] Mark S Graham et al. "Denoising diffusion models for out-of-distribution detection". In: *2023 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 2948–2957.

[343] Neal Mangaokar et al. "Jekyll: Attacking medical image diagnostics using deep generative models". In: *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE. 2020, pp. 139–157.

[344] Ziwei Liu et al. "Large-scale long-tailed recognition in an open world". In: *2019 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2537–2546.

[345] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. "Exploring the limits of out-of-distribution detection". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 7068–7081.

[346] Andre GC Pacheco et al. "On out-of-distribution detection algorithms with deep neural skin cancer classifiers". In: *2022 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2020, pp. 732–733.

[347] Chandramouli Shama Sastry and Sageev Oore. "Detecting out-of-distribution examples with gram matrices". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8491–8501.

[348] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. "Deep anomaly detection with outlier exposure". In: *arXiv preprint arXiv:1812.04606*. 2018.

[349] Christoph Berger et al. "Confidence-based out-of-distribution detection: a comparative study and analysis". In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer. 2021, pp. 122–132.

[350] Mark A Kramer. "Nonlinear principal component analysis using autoassociative neural networks". In: *AIChE Journal*. Vol. 37. 2. Wiley Online Library, 1991, pp. 233–243.

[351] Shuyu Lin et al. "Anomaly detection for time series using vae-lstm hybrid model". In: *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ieee. 2020, pp. 4322–4326.

[352] Jinwon An and Sungzoon Cho. "Variational autoencoder based anomaly detection using reconstruction probability". In: *Special Lecture on IE*. Vol. 2. 1. 2015, pp. 1–18.

[353] Xiao Li et al. “Latent space factorisation and manipulation via matrix subspace projection”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5916–5926.

[354] Seunghyoung Ryu, Minsoo Kim, and Hongseok Kim. “Denoising autoencoder-based missing value imputation for smart meters”. In: *IEEE Access*. Vol. 8. IEEE, 2020, pp. 40656–40666.

[355] Ricardo Cardoso Pereira, Pedro Henriques Abreu, and Pedro Pereira Rodrigues. “Vae-bridge: Variational autoencoder filter for bayesian ridge imputation of missing data”. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–7.

[356] Junwen Bai, Shufeng Kong, and Carla P Gomes. “Gaussian mixture variational autoencoder with contrastive learning for multi-label classification”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 1383–1398.

[357] Zissimos P Mourelatos and Jun Zhou. “A design optimization method using evidence theory”. In: 2006.

[358] Ilias Diakonikolas et al. “Being robust (in high dimensions) can be practical”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 999–1008.

[359] Alireza Zaeemzadeh et al. “Iterative projection and matching: Finding structure-preserving representatives and its application to computer vision”. In: *2019 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5414–5423.

[360] Zhaomin Chen et al. “Autoencoder-based network anomaly detection”. In: *2018 Wireless Telecommunications Symposium (WTS)*. IEEE. 2018, pp. 1–5.

[361] Mourad Gridach et al. “Dual Representation Learning From Fetal Ultrasound Video and Sonographer Audio”. In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2024, pp. 1–4.