Sheffield Hallam University | College of Business, Technology and Engineering

MSc Dissertation Report

"Advancing Robotic Self-Awareness through Multi-Modal Learning with Barlow Twins Architecture"

A dissertation submitted in partial fulfilment of the requirements of Sheffield Hallam University for the degree of Master of Science in MSc Artificial Intelligence

Student Name	Nguyen Quang Anh
Student ID	C3005993
Supervisor	Dr Alejandro Jimenez Rodriguez
Date of Submission	16/01/2025

This dissertation does NOT contain confidential material and thus can be made available to staff and students via the library.

Abstract

This dissertation investigates the development of robotic self-awareness through self-supervised learning, leveraging the Barlow Twins architecture for multi-modal signal processing. The research focuses on integrating interoceptive (joint states) and exteroceptive (image) data to create a unified latent space that facilitates self-perception in autonomous systems. Two primary downstream tasks are explored: self-recognition via binary classification and reconstructing missing sensor data using an encoder-decoder framework. Experimental results highlight the efficacy of pre-trained networks like ResNet-50 in improving classification accuracy, achieving approximately 85%. The study also demonstrates the potential of multi-modality models in reconstructing missing data, laying a foundation for advanced robotic cognition. However, challenges such as limited dataset size and underperforming initial models underscore the need for further refinement. This work contributes to bridging theoretical insights and practical applications in cognitive robotics, advancing the discourse on robotic self-awareness and its implications for human-robot interaction.

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my supervisor, Dr Alejandro Jimenez Rodriguez, for his invaluable guidance, insightful feedback, and unwavering support throughout this project. His expertise and encouragement have been instrumental in shaping the direction and outcome of this research.

I extend my appreciation to the faculty members at the College of Business, Technology, and Engineering at Sheffield Hallam University for their foundational teachings and resources, which provided the essential groundwork for this dissertation. Special thanks are due to my colleagues and peers, whose constructive discussions and collaborative spirit enriched my understanding and approach.

Lastly, I am profoundly thankful to my family and friends for their constant encouragement, patience, and belief in my abilities during this challenging yet rewarding journey.

Sincere appreciation from the depth of my heart.

Table of Contents

List of Figure

CHAPTER 1: INTRODUCTION

Self-awareness in robotics represents a significant frontier in the field of artificial intelligence, with implications ranging from improved autonomous decision-making in contingencies (Lanillos et al., 2017) to enriched human-robot interaction (Loureiro et al., 2022). Self-awareness in robots has profound implications, from enhancing autonomous decision-making to enabling complex human-robot interactions. The idea of a sense of self in robotics is supported by several research (Prescott, et al., 2024) (Lanillos & Cheng, 2020) and is a crucial step towards creating truly self-aware robots.

However, implementing ideas related to the sense of self in robots remains a challenging endeavor. One significant obstacle lies in the integration of cognitive architecture. Developing a comprehensive framework that incorporates various components of self-awareness, such as perception, learning, decision-making, and social interaction, is still an ongoing process (Chatila, R. et al., 2018). This approach suggests the potential utility of a model capable of processing multiple input signals, thereby enabling robots to extract richer information about both themselves and their surroundings. Additionally, there is a pressing need to bridge the gap between theoretical models of self-awareness and their practical implementation in robotic systems (Chella, A. et al., 2020). Closing this gap could facilitate more extensive experimentation on robots, including simulations and tests designed to evaluate human-like self-awareness, such as the mirror test. Moreover, the scarcity of robot-relevant training data poses another significant challenge to the development of advanced self-awareness models (Firoozi, R. et al., 2024). This limitation underscores the necessity for innovative solutions to enhance the data resources available for such endeavors.

Self-supervised learning (SSL), a subset of unsupervised learning, aims to learn discriminative features from unlabeled data without relying on human-annotated labels (Gui et al., 2024). This approach shows significant promise in advancing self-awareness in robotics, as it enables robots to learn from their interactions with the environment autonomously, thereby reducing obstacles associated with data collection and alleviating the burden of labeling large datasets. Furthermore, SSL has demonstrated its capability to support multi-modality (Duan, J. et al., 2022; Sükei, E. et al, 2024). This integration could be pivotal in the development of truly self-aware robots capable of functioning effectively in dynamic, real-world environments. Such advancements may also enhance a robot's ability to pass self-awareness assessments, including popular tests designed to evaluate this capability.

This research seeks to explore how contrastive learning—a self-supervised learning paradigm—can be utilized to cultivate a robotic sense of self. Specifically, the project focuses on leveraging the Barlow Twins architecture to process interoceptive and exteroceptive signals, creating a framework for self-perception. Section 2 will present the literature review, summarizing the Theory of Mind and the experiments currently employed in robotics for self-recognition testing. From this foundation, the section will propose the use of contrastive learning with Barlow Twins loss to address the challenges of multi-modality. Section 3 will explain simple network architectures utilized to construct the comprehensive network structure, the dataset design, and the methods for its augmentation. Section 4 will showcase the experimental results obtained through the implemented experiments. Section 5 will consolidate the findings, discuss their implications, and propose directions for further work in the future.

CHAPTER 2: LITERATURE REVIEW

Cognitive robotics is a field that addresses the knowledge representation and reasoning challenges encountered by autonomous robots (or agents) operating in dynamic and incompletely understood environments (Levesque et al., 2008). This approach integrates artificial intelligence with robotics to develop machines capable of perceiving, learning, adapting, and making decisions autonomously, closely mirroring human cognitive processes (Singh et al., 2022). Recent advancements in cognitive robotics have introduced more sophisticated approaches, incorporating multi-channel information processing, including both internal and external signals (Li et al., 2019). This approach has been applied across various domains, such as Natural Language Processing (Taniguchi et al., 2019) and Computer Vision (Leitner et al., 2013). As a result, it has led to enhanced human-robot interaction and improved efficiency and flexibility (Aly et al., 2017). In light of these developments, cognitive robotics is pushing the boundaries of what is achievable in robotics, enabling machines to operate with greater autonomy, adapt to complex environments, and interact more naturally with humans. This progress is opening new avenues for applications and markets within robotics (Makedon et al., 2021), particularly in areas that require advanced decision-making capabilities and adaptability (Lebiere et al., 2013).

2.1 Theory of self

The theory of self in psychology encompasses a variety of perspectives on how individuals perceive, evaluate, and understand their own identities. Michael Lewis' Self-Concept Theory (1995) focuses on how individuals develop an understanding of themselves through self-awareness and social interactions. He underscores the significance of both emotional and cognitive components in forming a coherent self-concept. In the context of robotics, a robot's self-concept could evolve based on its interactions with humans and the environment. For instance, a robot may develop a sense of "competence" after successfully performing tasks such as cleaning or assisting people, which could, in turn, shape its future actions and self-perception as an efficient assistant, thereby influencing its approach to new challenges. John Turner's Self-Categorization Theory (2011) posits that individuals categorize themselves into social groups, which then influences their identity, behavior, and perceptions of others. This theory highlights that self-concept is shaped by group membership, with the possibility of change depending on the context. Building on this idea, a robot may categorize itself within a specific function or role, such as a "service robot" (Gonzalez-Aguirre et al., 2021) or a "companion robot" (Ruggiero, A. et al., 2022). Consequently, its behavior and interactions would be driven by this self-categorization, prompting it to act in accordance with the expectations of the group or role it identifies with, such as being more helpful in caregiving contexts or more efficient in service settings. In this regard, the theory of self continues to evolve, drawing insights from various disciplines to offer a comprehensive understanding of human self-perception and identity. In line with William James' Theory (Woźniak, 2018), the self is divided into two categories: the ‘me’ (empirical self), which represents the object of self-knowledge, and the ‘I’ (pure ego), the thinking self, linked to consciousness and continuity. In the context of robotics, the "me" could be understood as the robot's stored data and past experiences—its learned behaviors, actions, and interactions with the environment. On the other hand, the "I" would refer to the robot's real-time processing and self-awareness, encapsulating the algorithmic "thinking" that enables it to make decisions and react to new stimuli, thereby maintaining a sense of continuity and purpose in its actions. Thus, the interplay between these elements creates a dynamic and evolving self-concept in robots. According to Dr. Tony Prescott, the conflict between the intuition of the self as an indivisible entity and the self as a construct linked to the body has been reflected through centuries of debate in science and philosophy concerning the human condition. In this research, I will not delve into the details of the theory of self; however, it remains a potential avenue for robotics.

In this context, robots can leverage theories of self to guide experimental designs and learning processes (Allan et al., 2022) (Lu, Y. et al. 2023), while humans, conversely, can utilize robots as subjects to test and refine hypotheses about self-perception and identity (Halilovic & Krivic, 2024), for example, research on autism (Scassellati et al., 2012) (Pennisi et al., 2016). Specific, the concepts of Minimal Self and Extended Self in self-awareness (Prescott, et al., 2024) provide valuable frameworks for both understanding and simulating self-related cognitive functions in robots. The minimal self encompasses two core elements: the sense of body ownership (SoO) and the sense of agency (SoA), both considered as the grounding of the sense of Self (Moore, 2016; Braun et al., 2018; Legaspi et al., 2019). These elements enable the distinction between "self" and "other," which is fundamental for any agent interacting with its environment. For robotics, this translates into the ability to recognize the robot’s own physical boundaries and actions. Implementing SoO involves sensory and motor integration, allowing robots to identify their own structural components, while SoA enables them to perceive the consequences of their actions, or to conceive or be aware of itself as persisting in time (Gallagher, S., 2000). On the other hand, the extended self builds upon the minimal self by incorporating more complex temporal and social dimensions. Drawing on developmental psychology, the extended self involves autobiographical memory, "mental time travel" into the past or future (Moore et al., 2016), and the capacity for social cognition through theory of mind (ToM) (Baron-Cohen, 1997). In robotics, this can be achieved through mechanisms for long-term memory storage and retrieval, predictive modeling, and social interaction frameworks. For example, a robot with autobiographical memory can adapt its behavior based on past experiences, while predictive planning supports goal-directed actions (Prescott, et al., 2024).

Self-awareness is the capacity to recognize oneself as a distinct entity with unique thoughts, emotions, and actions, encompassing both immediate and reflective understanding. It involves perceiving the self in relation to the environment and others, often emerging as a result of complex cognitive and perceptual processes (Gallagher, 2000; Rochat, 2003). In humans, self-awareness develops through processes such as self-recognition and self-reflection, allowing individuals to perceive their thoughts, actions, and emotions as uniquely their own (Rochat, 2003). In robotics, self-awareness is increasingly explored as a foundational capability for enabling autonomous systems to interact dynamically with their environments and adapt effectively. For instance, the distinction between internal processes and external stimuli has been identified as essential for self-evaluation and adaptive decision-making in robots (Chatila et al., 2018) . This mirrors theories of minimal self, which focus on immediate, embodied experiences, and extended self, which integrates temporal and social dimensions, forming the basis for advanced cognitive functions (Gallagher, 2000; Mentzou & Ross, 2024). Efforts to engineer self-awareness in robotics not only contribute to more autonomous systems but also provide a cross-disciplinary lens to refine theoretical understandings of human self-awareness and its emergence (Prescott & Camillieri, 2019). These advancements underscore the potential for robotics to both simulate and illuminate the complex architecture of self-awareness. This exploration naturally leads to the issue of self-awareness in robotics—a topic that holds promise for offering innovative approaches to understanding and advancing cognitive robotics.

2.2 Experiments with robot

By equipping robots with mechanisms to differentiate between self and other, recognize their own capabilities and limitations, and adapt based on contextual feedback, researchers can push the boundaries of what is achievable in autonomous systems. These developments not only enhance robot functionality but also provide a mirror to human cognitive processes, potentially deepening our understanding of self-awareness and its role in intelligent behavior. Building on this foundation, self-awareness in robotics represents a convergence of theoretical insights and practical applications. A variety of robots have been used to investigate different aspects of self and related behavioral phenomena (Asano et al., 2017; Moulin-Frier et al., 2017; Bongard et al., 2006; Koga et al., 2021; Saegusa et al., 2009; Roncone et al., 2016). By embedding concepts like the minimal and extended selves into robotic systems, researchers are not only advancing the field of robotics but also enriching the broader discourse on the nature of self. It opens new pathways for robots to interact more intuitively with humans, operate more effectively in dynamic environments, and contributes to interdisciplinary research that bridges artificial intelligence, psychology, and philosophy.

One of the most widely recognized and foundational experiments on self-awareness is the Mirror Test (Gallup Jr., 1982). Originally developed by Gordon Gallup Jr., this behavioral experiment evaluates self-recognition and self-awareness in both animals and humans. The test involves placing a mark on the subject in a location visible only through a mirror and observing whether the subject recognizes the mark as being on their own body. Success in this test is often considered evidence of a basic form of self-awareness. Beyond animals (Huttunen et al., 2017; Kohda et al., 2019; DeGrazia, 2009), robots have also demonstrated the ability to participate in such experiments, contributing to advancements in theories of self (Mentzou et al., 2024; Gorbenko et al., 2012). As noted by Mentzou et al. (2024) in their work *Self-Recognition as the Milestone of Explicit Self-Awareness*, numerous experiments employing diverse methodologies have been conducted to enable robots to pass the Mirror Test, including approaches using neural networks (Lanillos et al., 2020) and inner speech (Pipitone & Chella, 2021). Beyond the Mirror Test, other experiments explore different dimensions of self-awareness, such as the Sally-Anne Test (Demirel et al.,2021). This psychological experiment is designed to assess Theory of Mind, particularly the ability to recognize that others can hold beliefs different from one’s own. In the test, a character (Sally) places an object in a specific location and leaves the scene, after which another character (Anne) moves the object to a new location. Participants are then asked where Sally will look for the object upon her return. Success in this task reflects the ability to attribute false beliefs to others, a critical component of understanding perspectives distinct from one’s own. This capability is closely tied to self-awareness, as it requires differentiating and reasoning about one’s own mental states in relation to others.

Another experiment on robot self-perception, conducted through sensorimotor contingencies, was presented by Pablo Lanillos et al. (2016). Paper introduces a novel mechanism for self-perception in humanoid robots, emphasizing the integration of visual, proprioceptive, and tactile cues through a hierarchical Bayesian model. This system enables robots to distinguish between internal and external sensory stimuli by analyzing sensorimotor contingencies, allowing them to identify objects in their environment and interpret causality during interactions. Experimental validation demonstrated the robot's ability to discover usable objects and disambiguate visual artifacts, achieving a 92% success rate. Similar approaches in the literature include methods that use temporal coherence between motor commands and visual motion (Michel et al., 2004) or integrate visual and proprioceptive cues for body representation (Hikita et al., 2008). Other works employed probabilistic reasoning, such as Gold et al. (2009), who analyzed motor-visual correlations, or used temporal contingencies for self-detection (Stoytchev, 2011). More recent studies, like Rolf and Asada (2014), applied reinforcement learning for self-detection and goal-directed actions.

2.3 Multi-modality model

A common characteristic of the aforementioned experiments is their reliance on the integration of multiple modalities, effectively leveraging both internal and external signals.

By combining these modalities, meaningful information can be extracted, enabling experiments such as the Mirror Test to validate theories of self-awareness in robotics. Consequently, multi-modality models represent a critical area of research that warrants thorough exploration. These models have gained significant attention due to their ability to enhance human-robot interaction (HRI) and improve robotic capabilities across a wide range of tasks. By integrating diverse sensory inputs and output modalities, multi-modality models contribute to the creation of more robust and versatile robotic systems. To effectively implement these experiments, it is essential to establish robust methods for extracting and processing multi-modal signals, particularly image data, which is widely utilized in robotics (Kurka & Salazar, 2019; Cebollada et al., 2021). Image data plays a central role in enabling robots to perceive, understand, and interact with their environments effectively, owing to its versatility and the extensive information it provides for various applications (Liu et al., 2014). Beyond images, robotic systems can process a range of input signals, such as sound (Rascon & Meza, 2017), tactile feedback (Argall & Billard, 2010), and natural language (Tellex et al., 2020), offering diverse approaches to signal processing and enhancing functionality. In this research, the focus is placed on the integration of joint data with image data. Joint data offers advantages such as real-time monitoring (Faisal et al., 2019) and kinematic analysis (Raza et al., 2018), making it a complementary modality for image-based systems. While image data enables perception of the external environment, joint states offer critical insights into the robot's internal kinematics and real-time motion. Combining these modalities enhances the robot's ability to perceive itself and its surroundings effectively, creating a promising pathway for downstream tasks such as reconstructed missing data or classification of self and non-self. In conclusion, this study highlights the potential of combining image and joint data, emphasizing their integration as an effective and practical approach to advancing downstream experiments and improving human-robot interaction.

Contrastive learning is a machine learning technique that teaches models to distinguish between similar and dissimilar samples without requiring labeled data, and may be a potential method in the study of multi-modality models (Zhang, Q., Wang, Y., & Wang, Y., 2023). The origins of Contrastive Learning date as far back as the 1990s and its development has spanned across many fields and domains including Metric Learning and natural language processing (Le-Khac, P. H. et al., 2020). It works by maximizing the similarity between positive pairs of samples while minimizing the similarity between negative pairs in a learned embedding space and can offer significant advantages in dealing with multi-modality problems. For instance, contrastive learning enables models to learn powerful and generic visual representations that can capture semantic information across different modalities (Yuan, X. et al., 2021). By using contrastive losses, features from multiple modalities can be adjusted and aligned in a unified representation space. Barlow Twins is not a subfield of contrastive learning, but rather an alternative approach within the broader field of self-supervised learning. While it shares some similarities with contrastive learning methods, it differs from traditional contrastive learning approaches in several keyways: does not require negative samples or large batch sizes and focuses on redundancy reduction rather than explicit contrast between positive and negative pairs (Hansen & Martinetz, 2024) (Tsai & Salakhutdinov, 2021). Barlow Twins provides an efficient method for aligning features from different modalities by reducing redundancy in representations. This makes it particularly suited for multi-modality tasks involving image and joint states data in robotics. By leveraging Barlow Twins, this research aims to optimize the processing of interoceptive (joint states) and exteroceptive (image) cues, enabling the robot to perform downstream experiments.

2.4 Summary

In summary, the integration of Barlow Twins into processing multi-modal cues represents a promising approach for optimizing downstream tasks in robotics, particularly for enhancing the sense of self in autonomous systems. By leveraging self-supervised methods, this research explores the structured use of interoceptive and exteroceptive signals to advance robotic cognition. For instance, as described by Lanillos et al. (2020), a robot can achieve self-recognition by correlating its actions with sensory feedback: when a robot moves its arm, proprioceptive and visual sensors capture the resulting changes, allowing the robot to affirm, "This is my arm not only because I sent the command to move it, but also because I sense the consequences of moving it." This dynamic interaction between action and sensory perception exemplifies the potential of self-supervised techniques in reinforcing self-awareness frameworks in robotics. The integration of such models not only enhances the robot's ability to interact effectively with its environment but also provides a foundation for tackling complex cognitive tasks. In this research, the focus will be on designing and experimenting with self-supervised methods and multi-modality architectures to optimize the processing of interoceptive and exteroceptive cues, thereby advancing the understanding of self-awareness in robotics.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Overview

3.1.1 Research Question & Objectives

Question: How can contrastive learning with Barlow Twins architecture be applied to develop a sense of self in a robot, using both interoceptive and exteroceptive signals?

Objectives:

1. Develop and implement a self-supervised learning framework using Barlow Twins to align and integrate multi-modal sensory data. Train a model capable of processing and embedding interoceptive and exteroceptive signals into a unified latent space.
2. Evaluate the effectiveness of the learned representations in downstream tasks. Conduct experiments on self-recognition (binary classification) and missing sensor data reconstruction to assess model performance.
3. Analyze the results, benefits of pre-training with ResNet-50 and other state-of-the-art architectures in improving accuracy and robustness.
4. Identify limitations and propose enhancements for future applications in robotic self-awareness. Explore challenges such as small dataset sizes, model overfitting, and the complexity of multi-modal data alignment.

3.1.2 Research Philosophy & Approach

The research will adopt a **Pragmatic** philosophy, as pragmatic philosophy ensures a focus on tangible results that can be tested, iterated in experiments. It emphasizes the integration of theory (contrastive learning principles, Barlow Twin architecture) with practice (robotic application and testing). This research also employs a **Deductive** approach that builds knowledge from Barlow Twins architecture to a specific robotic context. The study focuses on developing and testing a model based on these theories to validate their applicability in robotic settings. The research strategy is **Experimental**, emphasizing continual refinement of the model and its components. This includes exploring alternative configurations, testing diverse augmentation techniques, and conducting validation experiments inspired by real-world scenarios, such as the mirror test. Through this combined philosophy and approach, the research aims to bridge the gap between abstract theories of self-awareness and their tangible applications in robotics. By adopting pragmatism and experimentation, the study ensures a robust and iterative process of model development and validation, paving the way for advancements in robotic cognition.

3.1.3 Research Overview

This research explores the development of self-awareness in robotic systems through advanced machine learning techniques, focusing on the integration of multi-modal data and feature extraction. The primary objective is to design and evaluate models that can process diverse types of inputs, such as images and joint signals, to create a unified latent space capable of capturing the robot's sense of self and its environment. To achieve this, the study investigates various neural network architectures, including Multi-Layer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), and ResNet-50, which are strategically chosen for their complementary strengths in feature extraction and representation learning.

In designing and improving the architecture for feature extraction in robotic self-awareness models, employing structures such as Multi-Layer Perceptron (MLP) and Fully Connected Layers is fundamental. These architectures form the backbone of many neural networks, enabling the creation of deeper models capable of learning and extracting more abstract and meaningful representations from data. MLPs, with their layered architecture, allow for progressive refinement of information, making them ideal for learning complex patterns across different modalities. Fully Connected Layers, on the other hand, serve as the final stages in many models, aggregating features from earlier layers to produce compact and informative embeddings, essential for tasks requiring accurate feature representation. While these foundational structures are completed, experimenting with more advanced architectures such as ResNet-50 offers significant advantages, particularly for image-based feature extraction. ResNet-50, a high-performance convolutional neural network, is designed with a residual learning framework that addresses the challenges of vanishing gradients in deep networks. Integrating ResNet-50 into the experimental setup capitalizes on its proven ability to handle complex image-processing tasks effectively, thus complementing the more generalized feature extraction capabilities of MLPs and Fully Connected Layers.

The experiments are designed to systematically assess the performance of different architectures and their combinations in extracting meaningful features from multi-modal inputs. Initial experiments focus on using MLPs and Fully Connected Layers to create a latent space, emphasizing their ability to build progressively deeper and more refined representations. Subsequent trials incorporate CNNs for image feature extraction, given their proven efficacy in capturing spatial hierarchies. The study then extends to experiments with ResNet-50, a state-of-the-art network renowned for its exceptional performance in image analysis, to test its suitability for this specific application. To evaluate the effectiveness of the created latent space, I will assess it through a downstream task in the form of a classification test, with details to be described in subsequent sections. Once a network architecture achieving satisfactory accuracy in the classification test is established, the same architecture will be utilized for a second downstream task, the reconstruction task. This task requires reconstructing information from the input data and will be tested under two conditions: using complete input data and using incomplete input data.

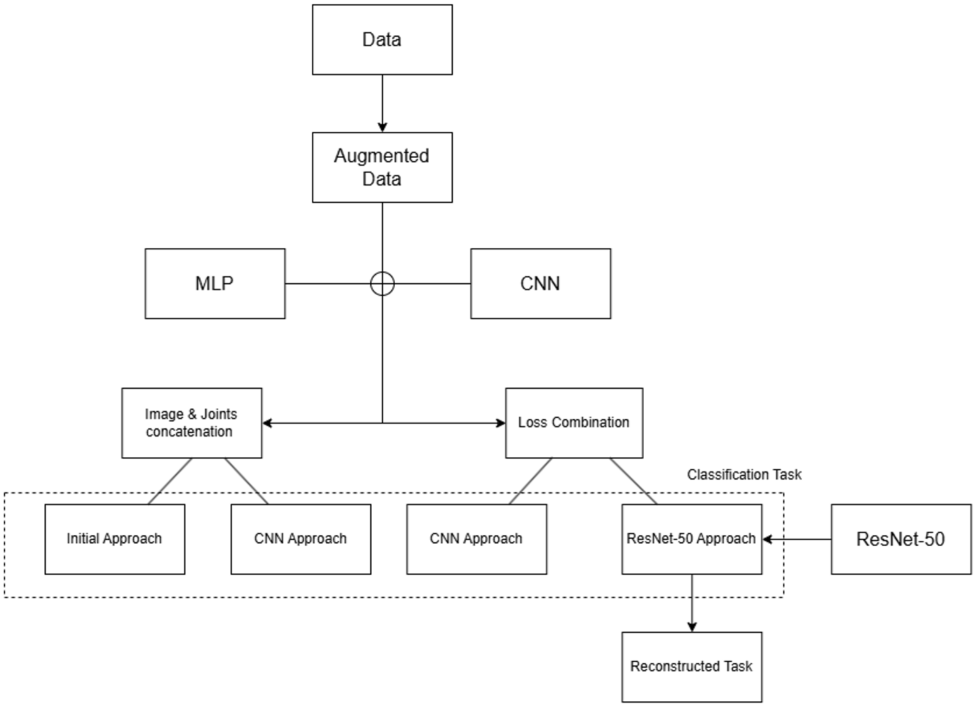


Figure 1 Overview of the approach

3.2 Dataset

3.2.1 Interoceptive and exteroceptive signals

Interoceptive and exteroceptive signals play crucial roles in robotic systems, mimicking biological sensory processes to enable robots to perceive their internal state and external environment. Exteroceptive signals are those that provide information about the robot's external environment. These signals are crucial for robots to navigate, avoid obstacles, and interact safely with their surroundings, thereby shaping the surrounding environment. Examples of exteroceptive sensors include vision sensors, proximity sensors, range sensors, tactile sensors, GPS receivers, etc. Interoceptive signals, in contrast, provide information about the robot's internal state. In robotics, these are often referred to as proprioceptive signals. They measure values internal to the robot system, such as: motor speed, joint angles, wheel load, etc.

In this study, the dataset used comprises two primary modalities: exteroceptive signals represented by images and interoceptive signals captured as joint states in '.npy' format. These signals are integral to achieving multi-modal self-supervised learning for robotic self-awareness, as they provide complementary information about the robot's internal state and its external environment. The exteroceptive signals are grayscale images that encode the robot's external perception, including the spatial configuration of its body relative to the surrounding environment. Each image is of a fixed resolution and represents a single frame captured by the robot's onboard camera. An example of such an image is shown in Figure 2.

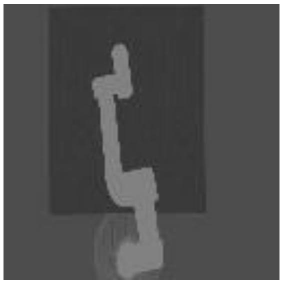


Figure 2 An instance of image data

These images serve as the robot's "vision," enabling it to learn spatial and environmental features critical for downstream tasks like self-recognition and reconstruction. The interoceptive signals are stored in '.npy' files, each corresponding to the same time step as the respective image in the dataset. This data serves as the robot's "proprioception," providing insight into its physical configuration and movement. By aligning these signals, the dataset enables the development of a unified latent space where internal and external perceptions coexist. Each data pair (image and '.npy' file) represents a snapshot of the robot's in environment, forming the basis for the training of self-supervised models. Through the careful processing of these signals, this work aims to enable the robot to develop a coherent self-representation that bridges its internal state with external observations.

3.2.2 Augmentation Method

One of the critical aspects of contrastive learning, including the Barlow Twin method, is the data augmentation step. At this stage, different versions of a pair of original data are generated while ensuring that key features are preserved and not lost during the transformation process. Cropping, flipping, rotation, and color transformations are commonly employed augmentation techniques. These methods generate diverse variations of the same data, enabling the model to learn robust representations that capture significant features regardless of input variability.

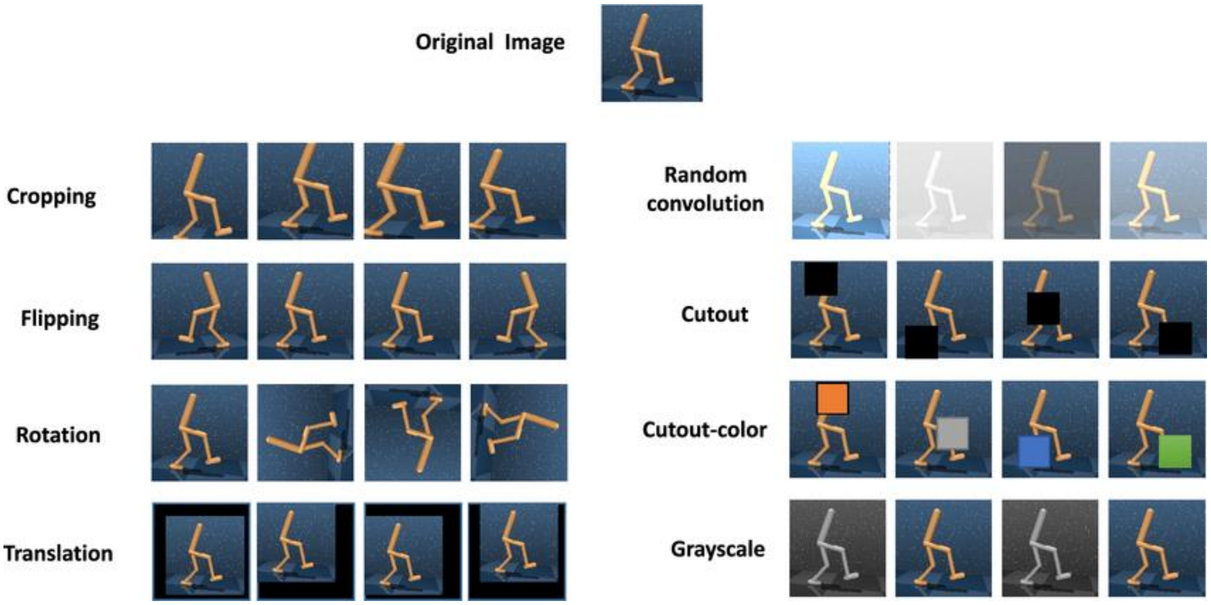


Figure 3 Popular Image Augmentation Methods [1]

The effectiveness of an augmentation method depends on several key considerations. In natural language processing datasets, for instance, augmentations must preserve semantic content while introducing diversity. Similarly, for time-series data, augmentations should retain trend and seasonality characteristics. The choice of augmentation techniques must also align with the specific properties of the dataset, such as trends, seasonality, and integration weights, as emphasized by Demirel and Holz (2024). Task relevance is another crucial factor. The selected augmentation techniques should correspond to the specific downstream task, as different augmentations can significantly impact performance, with variations in accuracy exceeding 30% in some cases (Liu et al., 2024). Moreover, combining multiple augmentation methods often yields better results than relying on a single approach, as demonstrated by Zhang and Ma (2022). The strength of augmentation also requires careful calibration. Overly weak augmentations might fail to cluster intra-class features effectively, while excessively strong augmentations risk collapsing inter-class features. For example, in my experiment, dataset where the images are in grayscale, applying overly intense brightness adjustments could obscure critical features of robotic arm images.

In multimodal experiments, augmentation methods must consistently preserve the relationships between input signals. For example, one of my experiments that was tested and subsequently discarded involved using image flipping and rotation as augmentation methods for image data while adding noise to joint states. This mismatch altered the direction and axis of the robotic arm in the images without corresponding changes in joint

[1] https://www.researchgate.net/figure/The-augmentation-transformation-methods-we-investigated-cropping-flipping-rotation_fig2_364091402

states, disrupting the correlation between these two signal types. By carefully addressing these considerations, researchers and practitioners can identify the most appropriate augmentation strategies for specific contrastive learning tasks, leading to enhanced performance and more robust feature representations. Within the scope of this research, I will use simple augmentation methods, while the exploration of more complex methods is reserved for future work.

For image input, two augmentation methods will be applied to compute the Barlow Twin loss. These are central cropping (with a small ratio of approximately 15%) and brightness adjustment (increasing or decreasing pixel values by a small factor, around 10%). These modest adjustments ensure that the fundamental characteristics of the images are preserved. For joint states data, similarly, two augmentation methods are used. The first method involves adding noise to the values (with noise ratios also around 10%), and varying the noise levels can generate two augmentation variants. Before being input into the network, all data undergoes normalization using the Standard Scaler method, which scales the values to the range [0, 1] (This normalization process can also support future work, where missing values in the input could be assigned a placeholder value of -2).

3.3 Multi-Layer Perceptron (MLP)

Multi-Layer Perceptron (MLP) is a fundamental type of artificial neural network that has gained significant attention in the field of machine learning and deep learning. It consists of multiple layers of interconnected neurons, typically including an input layer, one or more hidden layers, and an output layer. This architecture allows MLPs to model complex relationships between inputs and outputs, making them powerful tools for various tasks such as classification, regression, and pattern recognition.

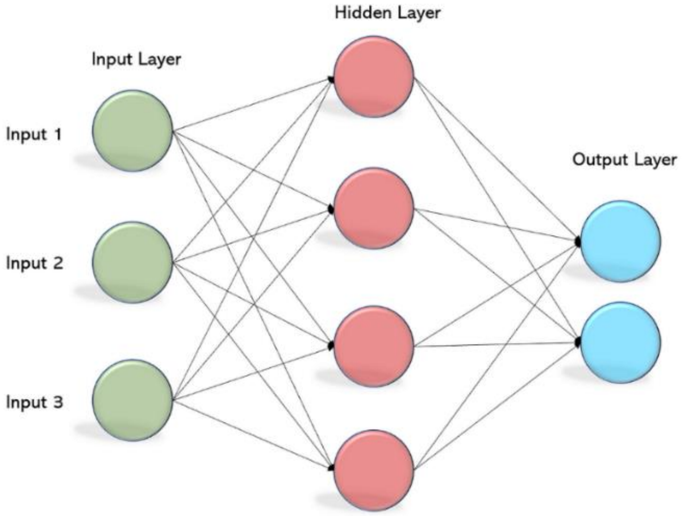


Figure 4 Multi-Layer Perceptron

In a multilayer perceptron (MLP), the architecture consists of three primary components: the input layer, hidden layers, and the output layer. The input layer is composed of neurons, each corresponding to an input feature. For example, if the dataset includes three input features, the input layer will contain three neurons. The hidden layers, which can vary in both number and size, process the information received from the input layer through complex transformations. Finally, the output layer produces the model's predictions or results. In cases where multiple outputs are required, the output layer will have a corresponding number of neurons to represent each output. In a multilayer perceptron (MLP), the layers are interconnected such that every neuron (or node) in one layer is connected to all neurons in the subsequent layer, continuing this pattern until the output layer. This type of connectivity is referred to as a Fully Connected Layer. Each connection in the network diagram represents the fully connected structure that characterizes an MLP, enabling comprehensive information flow and interaction between layers.

In a multilayer perceptron (MLP), key mechanisms that enable the model to function effectively include forward propagation, activation function, loss function, backpropagation, and optimization. During forward propagation, data flows sequentially from the input layer to the output layer, passing through the hidden layers. Each neuron in the hidden layers processes the input as follows:

$$z = \sum_{i=1}^{n} w_i x_i + b$$

With n representing the number of nodes in the preceding layer, x_i is the input feature, w_i is the corresponding weight and b is the bias term. However, with the computation of the weighted sum z as described above, it remains a linear function. Therefore, the weighted sum z is passed through an activation function to introduce non-linearity. Some common activation functions include Sigmoid, Tanh, and ReLU. In this research, I will focus on using ReLU (Rectified Linear Unit), defined by the following formula:

$$\mathrm{ReLU}(z) = max(0, z).$$

3.3.1 Loss function

One important step in the training process is the computation of the loss function. A loss function quantifies the difference between the predicted output of a model and the true target values. It serves as a measure of error, guiding the model during training by evaluating how well it performs on a given dataset. Minimizing the loss function is essential for improving the model's accuracy and generalization. In addition to the Barlow Twins loss used to train the model and create the latent space, I also employ several other loss functions for training models for downstream tasks, as detailed below.

For the first downstream task, the classification task, I used Sparse Categorical Cross entropy. Sparse Categorical Cross entropy is a loss function commonly used for multi-class classification problems, where each input is assigned to one of several categories. Unlike categorical cross entropy, sparse categorical cross entropy is used when the target labels are provided as integers rather than one-hot encoded vectors (As my classification experiment labels ‘1’ as similar or self, and ‘0’ as dissimilar or non-self). It calculates the cross-entropy loss between the true class labels and the predicted probabilities, penalizing incorrect predictions more heavily. The mathematical formula for Sparse Categorical Cross entropy is:

$$L = -\sum_{i=1}^{N} y_i \log(p_i)$$

Where: N is the number of classes, y_i is the true label for the i-th class (encoded as an integer), p_i is the predicted probability for the i-th class.

For the second downstream task, the reconstructed model, I used Mean Squared Error (MSE). Mean Squared Error is a loss function used to measure the average squared difference between the predicted values and the true target values. It penalizes larger errors more heavily due to the squaring, making it sensitive to outliers. The mathematical formula for MSE is:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2$$

Where: N is the number of data points, y_i is the true value of the i-th data point, $\widehat{y_i}$ is the predicted value for the i-th data point.

3.3.2 Backpropagation and Optimization

The goal of training an MLP is to minimize the loss function by adjusting the network’s weights and biases. This is achieved through backpropagation, a powerful algorithm in deep learning, primarily used to train artificial neural networks, particularly feed-forward networks. In each epoch, the model adapts these parameters, reducing loss by following the error gradient. Backpropagation often utilizes optimization algorithms such as gradient descent or stochastic gradient descent. The algorithm computes the gradient using the chain rule from calculus, allowing it to effectively navigate complex layers in the neural network to minimize the cost function. Once the gradient is computed, optimization methods are applied. Optimization is the process of adjusting a model's parameters to minimize the loss function and improve its performance. It involves iterative techniques, such as gradient descent, to find the parameter values that result in the best predictions. Effective optimization ensures that the model learns from the data efficiently and generalizes well to unseen data. In this research, I will employ and experiment with two widely used optimization methods: Stochastic Gradient Descent (SGD) and Adam Optimizer. Their respective formulas are as follows:

Stochastic Gradient Descent (SGD) parameter update rule is given by:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_\theta L(\theta)$$

where: θ represents the parameters (weights and biases), η is the learning rate and $\nabla_\theta L(\theta)$ denotes the gradient of the loss function with respect to θ.

Adam combines the benefits of momentum and adaptive learning rates, with the following update rules:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)\nabla_\theta L(\theta)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)\big(\nabla_\theta L(\theta)\big)^2$$

$$\widehat{m_t} = \frac{m_t}{1 - \beta_1^t}, \quad \widehat{v_t} = \frac{v_t}{1 - \beta_2^t}$$

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}} + \epsilon}$$

where: m_t and v_t are the first and second moment estimates, β_1 and β_2 are exponential decay rates for the moment estimates, ϵ is a small constant for numerical stability, η is the learning rate.

3.4 Convolutional neural network (CNN)

To extract signals from the image more efficiently in experiments, I will employ a Convolutional Neural network followed using the ResNet-50 network. Convolutional Neural Networks (CNNs) have emerged as a powerful class of deep learning algorithms, particularly well-suited for analyzing visual data or grid-like matrix and performing computer vision tasks. These networks are designed to automatically and adaptively learn spatial hierarchies of features from input images, making them highly effective for tasks such as image recognition, object detection, and image classification (Pak, M., & Kim, S., 2017) (Liu, B., Zhao, W., & Sun, Q., 2017)(Hussain et al., 2019). In a regular Neural Network, there are three types of layers: input layer, hidden layer, and output layer. However, a Convolutional Neural Network (CNN) expands the functionality of the hidden layers into more specific types, such as Convolutional layers, Pooling layers, and Fully Connected layers. The Convolutional layer applies filters to the input image to extract features, the Pooling layer downsamples the image to reduce computational complexity, and the Fully Connected layer performs the final extraction. The network learns the optimal filters through backpropagation and gradient descent. As the network processes the image through its layers, it constructs a hierarchical representation, capturing simple features in the early layers and more complex patterns in the deeper layers.

Convolution layers form the cornerstone of Convolutional Neural Networks (CNNs) and are instrumental in extracting spatial features from input data such as images. This process involves sliding a filter (or kernel) across the input image, computing dot products between the filter and local regions of the image. The result is a feature map that highlights the presence of specific features in the image. This operation exploits the principle of local connectivity, where each neuron in a layer is connected only to a small region of the previous layer, known as its receptive field (Araujo et al., 2019) (Luo, W. et al., 2016). Convolutions are often used for filtering images, such as smoothing, denoising, or detecting edges, by applying the Convolution operation. This operation involves sliding the filter matrix across the input feature map and computing the dot product at each spatial location, optionally followed by adding the bias term. To control the spatial dimensions of the output feature map, padding and stride are commonly used. Padding adds extra rows and columns, typically filled with zeros, around the input, which helps preserve the spatial dimensions or control the field of view. Stride, on the other hand, specifies the step size with which the filter slides across the input.

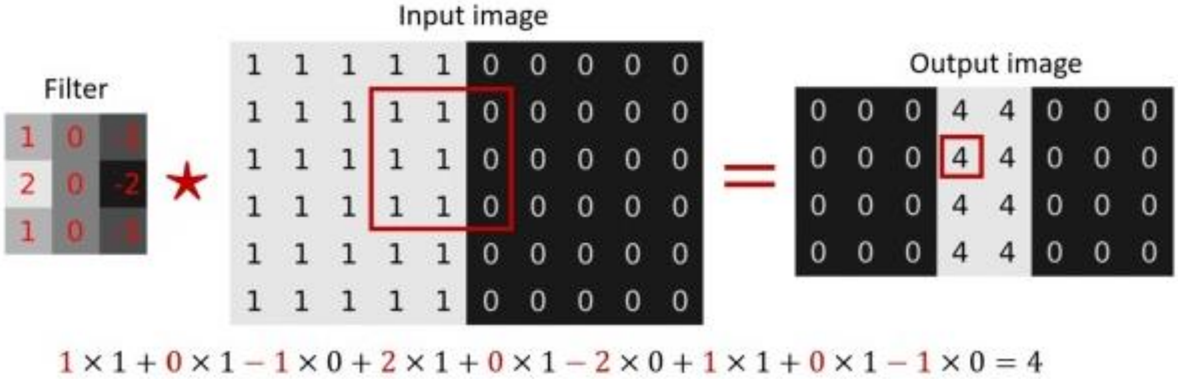


Figure 5 Convolution Operation [2]

Using various filter matrices, meaningful and trainable features can be extracted from input images. Deprez and Robinson (2024) describe the architecture of Convolutional Neural Networks (CNNs) as being based on three key principles. First, **local receptive fields** focus on extracting features from localized regions of the image rather than analyzing the entire image at once. This is achieved by applying small filters to specific areas, allowing the network to capture local patterns. Second, the concept of **shared weights** ensures that features are detected consistently across the image, regardless of

[2] Maria Deprez, Emma C. Robinson, in Chapter 11, Machine Learning for Biomedical Applications, 2024

their location. This property, known as translation invariance, is achieved by applying the same filter across all regions. Finally, **subsampling** reduces the size of the feature maps generated by convolutional layers. This downsampling step combines lower-level features into higher-level representations, making the exact positions of individual features less critical while preserving the overall patterns of significant features and their relative spatial relationships.

Regarding subsampling, this research employs max pooling layers after each convolutional layer. Max pooling is a widely used downsampling method in Convolutional Neural Networks (CNNs) to reduce the spatial dimensions of feature maps while retaining the most critical information. The process involves dividing the input feature map into non-overlapping regions, or "windows," and selecting the maximum value from each region. By doing so, max pooling emphasizes the most prominent features, making the network more robust to small spatial variations in the input, such as shifts or distortions. This reduction in dimensionality not only decreases the computational complexity of the network but also helps prevent overfitting by simplifying the learned representations.

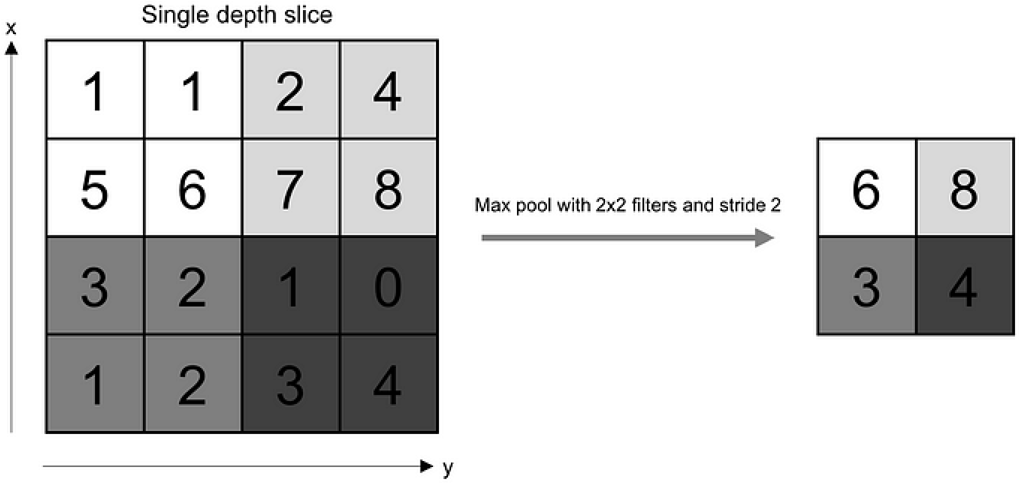


Figure 6 Pooling Operation [3]

3.5 ResNet-50 Architecture

ResNet-50 is a deep convolutional neural network (CNN) architecture that was developed by Microsoft Research in 2015. It is a variant of the popular ResNet architecture, which stands for “Residual Network.” The “50” in the name refers to the number of layers in the network, which is 50 layers deep[4]. This 50-layer deep convolutional neural network was designed to address the challenges of training very deep neural networks, particularly the

[3] https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939

[4] https://medium.com/@nitishkundu1993/exploring-resnet50-an-in-depth-look-at-the-model-architecture-and-code-implementation-d8d8fa67e46f

problem of vanishing gradients (Sharma, V., & Singh, N., 2021). The key innovation of ResNet-50 lies in its use of residual blocks, also known as skip connections or shortcut connections, where layer 1 output goes directly to layer N input. These connections allow the network to bypass one or more layers, enabling the direct flow of information from earlier layers to later ones. This architectural feature is crucial in mitigating the vanishing gradient problem, which had previously hindered the performance of very deep networks.

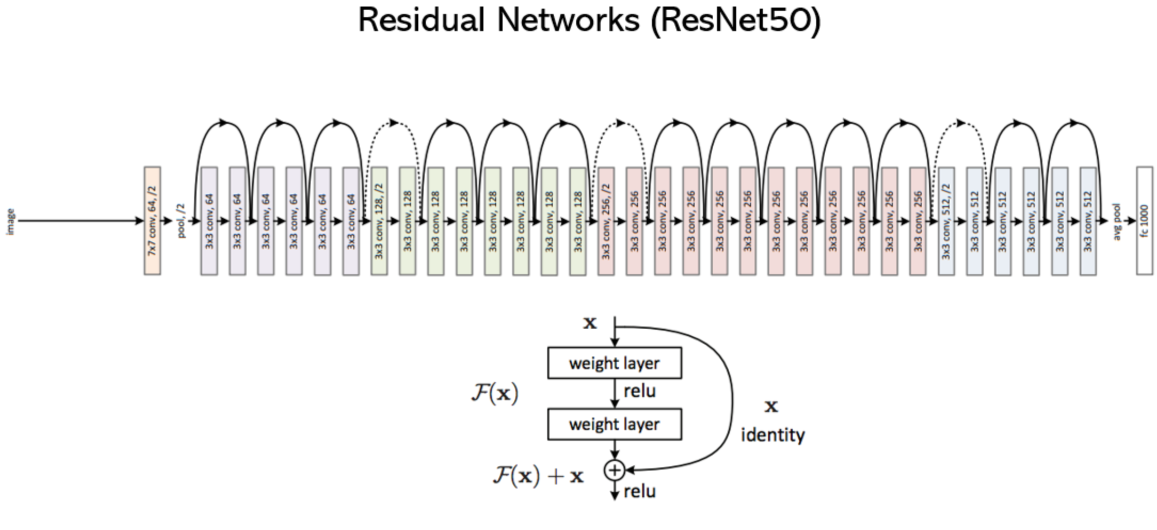


Figure 7 ResNet-50 using residual blocks (He, K. et al., 2016)

ResNet-50's architecture is composed of five main blocks, each containing a set of residual blocks. The network begins with a convolutional layer and a max-pooling layer, which preprocess the input image. Following these initial layers, the bulk of the network consists of stacked residual blocks. The residual blocks in ResNet-50 use a bottleneck design, which is a modification of the original two-layer blocks used in earlier ResNet variants. Each bottleneck block consists of three layers: a 1x1 convolution for dimensionality reduction, a 3x3 convolution for feature extraction, and another 1x1 convolution for restoring dimensions. This design allows for deeper networks with improved computational efficiency. One of the most significant aspects of ResNet-50 is its ability to train successfully despite its depth. The skip connections allow gradients to flow more easily through the network during backpropagation, enabling the network to learn effectively even with 50 layers. This is a substantial improvement over previous architectures that struggled with degradation in performance as depth increased.

The final layer of ResNet-50 is typically a fully connected layer with softmax activation, used for classification tasks. However, architecture is flexible and can be adapted for various computer vision tasks beyond classification, such as object detection and image segmentation. In my experience, I will not use the final layer but will instead utilize the preceding layers of ResNet-50 to extract features from the image. Subsequently, the following layers will be modified to integrate with the output of the joint or to create a latent space for downstream tasks.

3.6 Accuracy Evaluation Method

In the first downstream task experiment, I will test a classification task, and to evaluate the performance of the model, I will use the accuracy method based on the confusion matrix. The Confusion Matrix is a fundamental tool in machine learning and statistical classification, providing a comprehensive view of a model's performance. This square matrix presents a tabular summary of a classifier's predictions compared to the actual outcomes, offering insights into the types of errors and correct classifications made by the model.

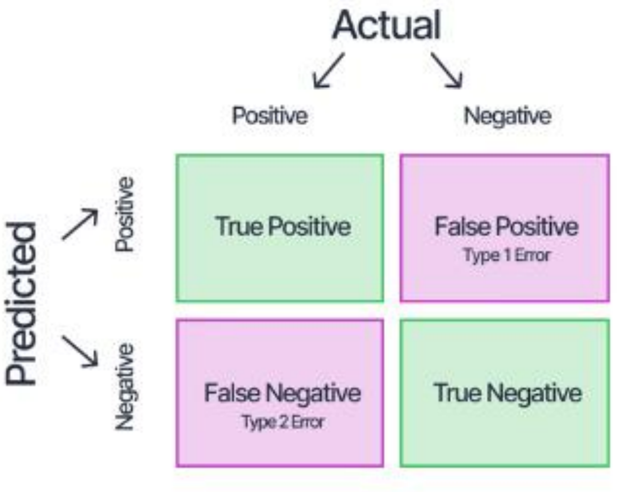


Figure 8 Confusion Matrix[5]

At its core, the Confusion Matrix is structured with rows representing the actual classes and columns representing the predicted classes, or vice versa. For a binary classification problem, the matrix consists of four key components: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These components form the basis for calculating various performance metrics. The matrix represents the number of instances generated by the model on the test data, providing insights into its predictive performance. A **True Positive (TP)** occurs when the model correctly predicts a positive outcome, aligning with the actual positive result. Similarly, a **True Negative (TN)** refers to instances where the model accurately predicts a negative outcome, matching the actual negative result. Conversely, a **False Positive (FP)** arises when the model incorrectly predicts a positive outcome, despite the actual outcome being negative; this is also known as a Type I error. Lastly, a **False Negative (FN)** happens when the model fails to predict

[5] https://www.v7labs.com/blog/confusion-matrix-guide

a positive outcome, incorrectly classifying it as negative, which corresponds to a Type II error. Accuracy is one of the most intuitive and commonly used metrics derived from the Confusion Matrix. It is calculated as the ratio of correct predictions (both true positives and true negatives) to the total number of cases examined. Mathematically, it is expressed as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

The accuracy metric provides an overall measure of the model's performance, indicating the proportion of correct predictions across all classes. It is particularly useful when the classes in the dataset are balanced and when all types of prediction errors are equally important (In the classification downstream task experiment, I have organized the data to be balanced between self and non-self labels).

3.7 Barlow Twin Method

One of the most effective methodologies for self-supervised learning in computer vision is visual representation learning (Bhattacharyya et al., 2022), which focuses on acquiring representations that remain invariant to various image distortions. However, these approaches frequently encounter a significant challenge in the form of trivial constant solutions, commonly referred to as collapse (Laurent et al.,2021; Jing et al.,2021). Embedding collapse is a critical phenomenon in representation learning that can adversely affect the performance and scalability of machine learning models (Guo et al., 2023). This issue manifests in two primary forms[6]. The first, **complete collapse**, occurs when representation vectors converge to a single point in the embedding space, resulting in constant features where the model generates identical embeddings for all inputs. The second, **dimensional collapse**, is a subtler and more challenging form of collapse, where representation vectors are confined to a low-dimensional subspace rather than fully utilizing the embedding space. This indicates that the collapse occurs along specific dimensions.

To mitigate embedding collapse, researchers have developed precise implementation strategies involving advanced architectural mechanisms such as momentum encoders, stop-gradient techniques, and more sophisticated models like Continuous-State Diffusion Models (CSDM) and Categorical Data in Diffusion Models (CATDM) (Zhang et al., 2023). Another widely adopted approach is contrastive learning, particularly via negative sampling, which prevents complete collapse by pulling positive pairs closer while pushing negative pairs apart in the embedding space (Nguyen et al., 2024). This method also encourages the embeddings to spread out, minimizing loss and benefiting from larger batch sizes to increase the availability of negative samples. However, as Nguyen et al.

[6] https://blog.reachsumit.com/posts/2024/11/embedding-collapse-recsys/

highlight, while contrastive learning effectively addresses complete collapse, it may not fully resolve dimensional collapse, necessitating further exploration of complementary strategies.

Barlow Twins is not a traditional contrastive learning approach, but it shares some similarities with contrastive methods while introducing a unique perspective on self-supervised learning. Barlow Twins has first appeared in the paper: Barlow Twins: Self-Supervised Learning via Redundancy Reduction (Zbontar et al., 2021). The method is called BARLOW TWINS, owing to neuroscientist H. Barlow's redundancy-reduction principle applied to a pair of identical networks. In this paper, the authors propose a novel way of doing visual representation learning by introducing a new objective function that naturally avoids collapse by calculating the cross-correlation matrix between the outputs of two identical networks fed with the distorted views of a sample, and making it as close to the identity matrix as possible. This causes the embedding vectors of distorted views of a sample to be close while minimizing the redundancy between the components of these vectors. As Zbontar mentioned, Barlow Twins does not require large batches nor asymmetry between the network twins such as a predictor network, gradient stopping, or a moving average (momentum encoders) on the weight updates. This is also an advantage for experiments in robotics as it can reduce the burden of collecting large amounts of data as well as large batches. The structure model utilized in the paper is designed as illustrated in the accompanying figure, with the components described as follows: Beginning with an image x from the dataset, two augmented versions of the image are generated, referred to as distorted images x1 and x2. These augmented images are then processed through an identical encoder-projector network, which outputs their respective representations.

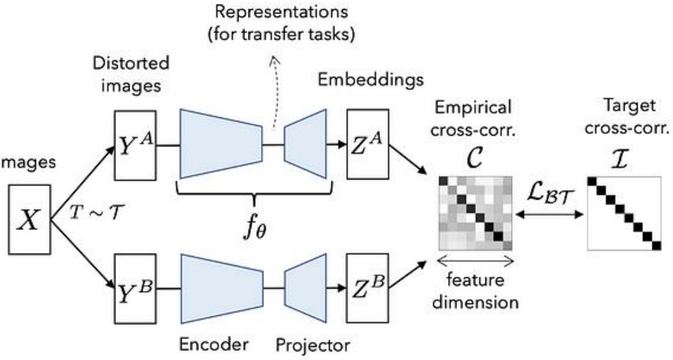


Figure 9 Barlow Twins Structure in Paper

In the paper, the encoder model comprises a ResNet-50 backbone followed by a projection head consisting of three linear layers, each with 8192 units. The first two layers are equipped with batch normalization and a ReLU activation function. Once the augmented images are passed through the network, two embedding vectors, z1 = *f(x1)* and *z2 = f(x2),* are obtained. These embeddings are normalized along the batch dimension before being used to compute the Barlow Twins loss. The Barlow Twins objection function is explained as follow[7]:

$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1-\mathcal{C}_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j\neq i} {\mathcal{C}_{ij}}^2}_{\text{redundancy reduction term}}$$

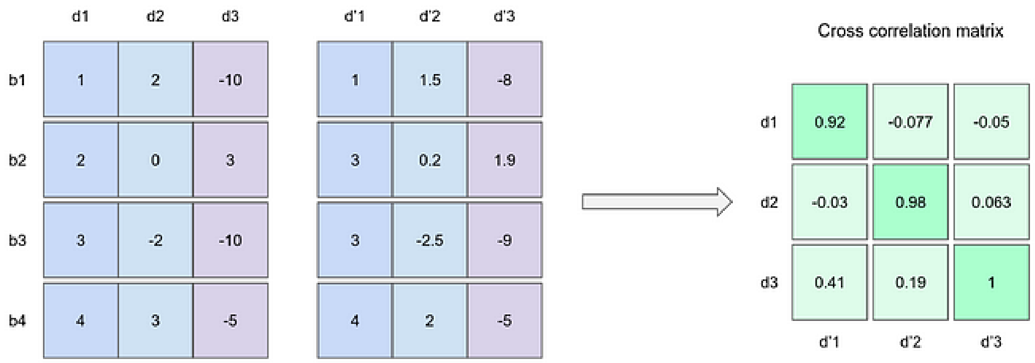
Figure 10 Barlow Twins Loss function

$$\mathcal{C}_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b \left(z^A_{b,i}\right)^2}\sqrt{\sum_b \left(z^B_{b,j}\right)^2}}$$

Figure 11 Pearson linear coefficient in the cross-correlation matrix between dimension i and j

In the equation, λ is a positive constant that balances the significance of the invariance term and the redundancy reduction term in the objective function, while C denotes the cross-correlation matrix computed between the normalized embeddings z1 and z2. The invariance term in the Barlow Twins objective function minimizes the loss by aligning the diagonal elements of the cross-correlation matrix with a value of one. This ensures that the embedding vectors of distorted views are highly correlated (correlation = 1), making them more similar to each other. Simultaneously, the redundancy reduction term reduces the loss by driving the off-diagonal elements of the cross-correlation matrix toward zero, decorrelating the components of the embedding vectors and reducing redundancy across their dimensions. The goal of the Barlow Twins objective function is to make the different dimensions of the embedding uncorrelated while keeping the same dimensions correlated.

[7] https://medium.com/@nazimbendib/paper-explained-barlow-twins-self-supervised-learning-via-redundancy-reduction-barlow-twins-92c90b49b21e



	d1	d2	d3
b1	1	2	-10
b2	2	0	3
b3	3	-2	-10
b4	4	3	-5

d'1	d'2	d'3
1	1.5	-8
3	0.2	1.9
3	-2.5	-9
4	2	-5

Cross correlation matrix

	d'1	d'2	d'3
d1	0.92	-0.077	-0.05
d2	-0.03	0.98	0.063
d3	0.41	0.19	1

Figure 12 An example for non-redundant embedding, with Cii ~ 1 and Cij ~ 0.[8]

The parameters of the Barlow Twins loss function can be tuned through the λ parameter, which adjusts the relative importance of the invariance term and the redundancy reduction term in the equation. A larger λ places greater emphasis on the redundancy reduction term, prioritizing the reduction of redundant embeddings during training. Conversely, a smaller λ gives more weight to the invariance term, ensuring stronger alignment between the embeddings of distorted views. Therefore, tuning the λ parameter is an essential consideration during the training and optimization of the model; however, in my experiments, I will use the λ parameter as suggested in the original paper (5×10^{-3}).

3.8 Latent Space

Latent space is a fundamental concept in machine learning and deep learning, representing an abstract, multidimensional space where complex data is encoded into a more compact and meaningful form. This space captures the essential features and underlying patterns of data, allowing for more efficient processing and analysis. The term "latent" refers to the hidden or unobservable nature of this space, as it represents features that are not directly present in the raw input data. In essence, latent space serves as a compressed representation of high-dimensional data, where similar data points are positioned closer together, facilitating various machine learning tasks.

[8] https://medium.com/@nazimbendib/paper-explained-barlow-twins-self-supervised-learning-via-redundancy-reduction-barlow-twins-92c90b49b21e

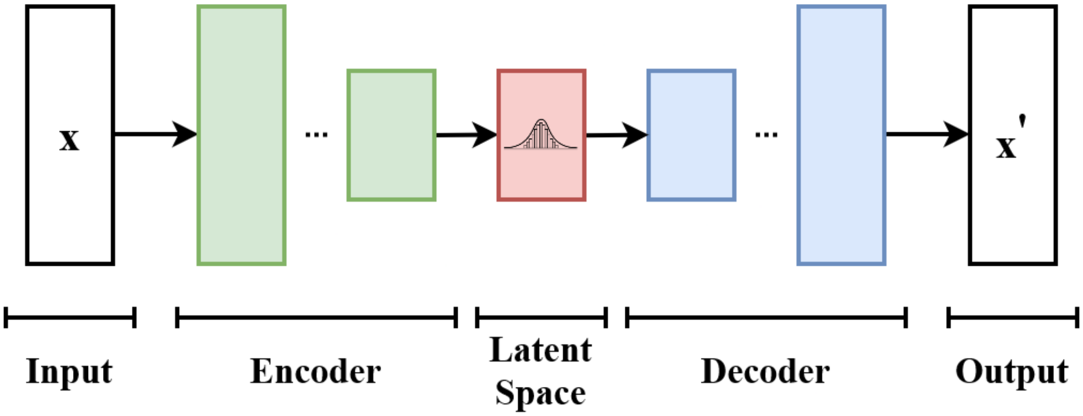


Figure 13 VAE with Latent Space [9]

Latent space is a widely adopted approach in numerous high-performance network architectures, including Autoencoders, Variational Autoencoders (VAEs) (Kingma, D. P. (2013), Generative Adversarial Networks (GANs), and recurrent models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs). Its advantages, such as dimensionality reduction, feature learning, and generative modeling, have demonstrated significant potential in various applications (Asperti & Tonelli, 2023; Kojima & Ikegami, 2022; Nemati et al., 2019). Mapping data into latent space typically involves dimensionality reduction, where high-dimensional data is transformed into a lower-dimensional representation. This transformation is achieved using neural networks like autoencoders, VAEs, or GANs, which encode the input data into the latent space and subsequently decode it back into its original form. This process enables the model to learn the most salient features of the data. Additionally, latent space excels in capturing complex relationships and similarities between data points, allowing similar data to cluster together in this abstract space. This clustering facilitates tasks such as data compression, feature extraction, and generative modeling by enabling more effective analysis, manipulation, and generation of data.

Given these advantages, leveraging latent space for downstream tasks with various input signals becomes a viable approach. This is particularly evident in the second experiment, which focuses on reconstruction. Due to its strengths in generative modeling, latent space can serve as a transfer station, extracting features from inputs using an encoder trained with the Barlow Twins framework. These features are then passed to decoders, enabling the reconstruction of meaningful information.

3.9 Downstream Task

[9] https://en.wikipedia.org/wiki/Variational_autoencoder

The use of latent space, combined with the aforementioned network architectures and loss functions, has enabled the model to preliminarily learn from input signals and extract features from two distinct inputs. The next step involves leveraging these features effectively.

The first downstream task is a classification task, designed to simulate a self-recognition system by integrating multiple types of signals. This system determines whether the state originates from "self" or is influenced externally, causing desynchronization among signals. The experimental setup is straightforward: the initial dataset is divided into 500 data points for training and the final 73 data points for testing. Since the data was originally labeled in pairs (images and joint states representing the robot arm at a specific time), these pairs are labeled as "self" with a label of "1." To create "non-self" data, the image at one time point is randomly paired with the joint state of a different time point, ensuring the data remains meaningful rather than entirely random. These mismatched pairs, termed "dissimilar," are assigned a label of "0." The same procedure is applied to construct the testing dataset. As a result, balanced datasets with equal representation of each label are created, transforming the task into a binary classification problem with labels “self” and “non-self,” or alternatively, “similar” and “dissimilar.”

The second downstream task is a reconstruction task, incorporating the idea of extracting features from multiple input signals. The Variational Autoencoder (VAE) framework, which typically consists of three components—Encoder, Decoder, and Sampling Layer—is designed to perform two key tasks: encoding input data into latent space and decoding latent space into output data. The VAE serves as a foundational structure capable of evolving into more advanced models, such as the Multimodal Variational Autoencoder (Multimodal VAE). This architecture, based on the VAE concept, supports the integration of multiple modalities as inputs and exhibits distinctive features. It shares a latent space across modalities, enabling the model to learn inter-modal relationships. Each modality is processed with its own Encoder and Decoder, while the shared latent space facilitates cross-modal learning. By leveraging this shared latent space, the model is trained using denoising strategies to reconstruct missing information. For instance, when data from certain sensors (e.g., tactile or auditory information) is unavailable, the model infers the missing data using inputs from other sensors. Consequently, this model not only reconstructs data but also aids in prediction and robotic control for complex tasks, such as recovering missing sensor information or predicting future states.
Based on this concept, I will develop the reconstruction task by designing an encoder-decoder network. The encoder will encode signals from various modalities (e.g., visual, tactile, auditory, and motor sensors) into a shared latent space, while the decoder will reconstruct each modality from this latent space. The encoder will utilize models pre-trained with the Barlow Twins loss function to create the latent space. The dataset will be partitioned similarly, with 500 pairs allocated for training and the remaining 79 pairs

reserved for testing; and only similar pairs will be utilized, excluding any "dissimilar pairs" from the dataset.

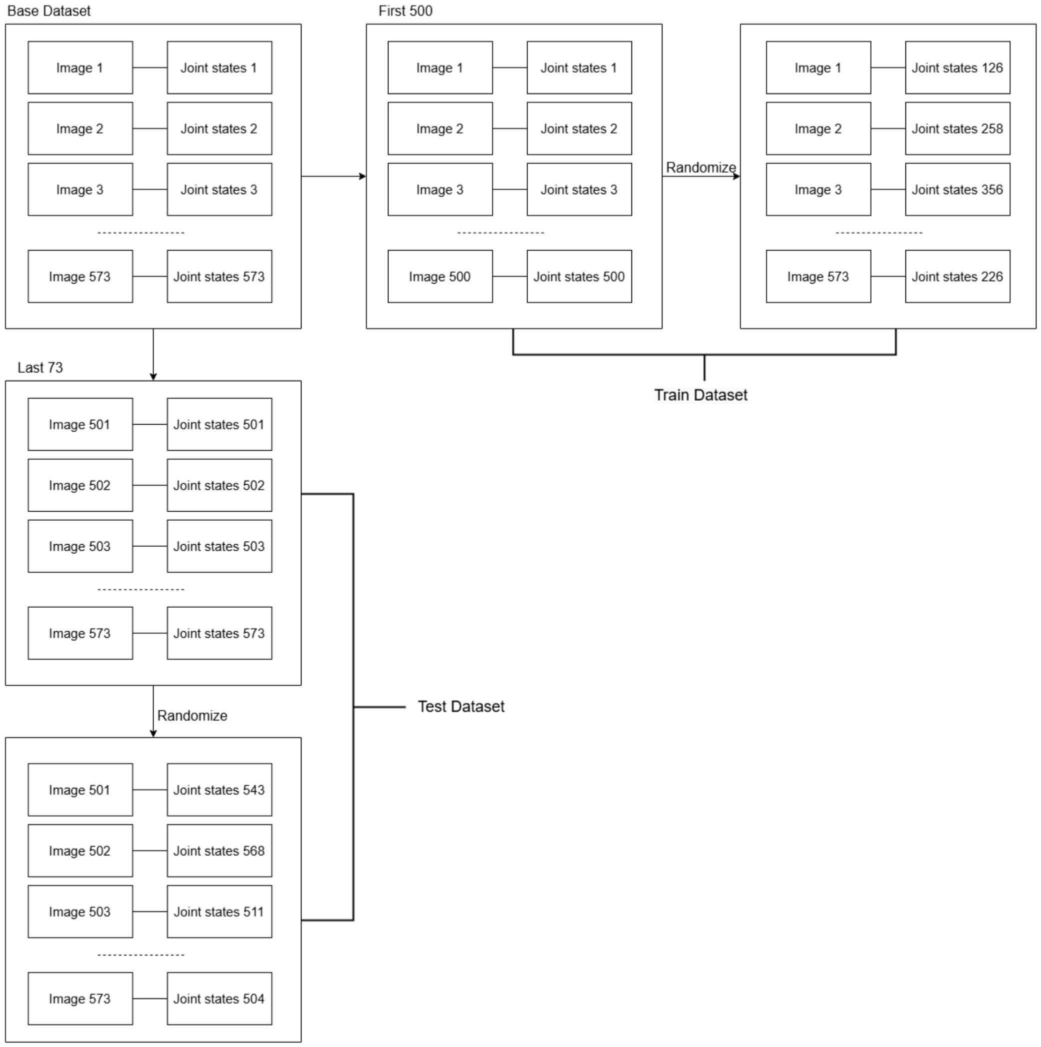


Figure 14 Creating classification dataset

As noted in the CVAE paper, learning from different sensor modalities poses challenges, as the learning model must handle diverse signal types and construct coherent representations even when some sensor inputs are missing. Thus, the second downstream task will focus on recovering missing data from the remaining signals (e.g., reconstructing images using joint states and vice versa) through the encoder-decoder structure. With a robust reconstruction network, further experiments can be conducted on real-world robots.

CHAPTER 4: EXPERIMENT RESULTS

4.1 Image & Joints concatenation

In this initial experiment, I will test the integration of Images and Joints, following the augmentation step. The first integration method involves a simple concatenation of the two. Subsequently, I will experiment with an approach where the joints are combined with the image after being processed through a CNN to extract features. For both integration methods, the downstream task involves adding a Fully Connected Layer of size 2 to perform a classification task on the aforementioned dataset.

4.1.1 Initial approach

For the initial experiment using the direct combination of Image and Joint, I will flatten the 128x128 image into a 16,384-dimensional vector and concatenate it with the 8-dimensional Joint states vector, resulting in a 16,392-dimensional (16384+8) input vector for the network. The first network employed is a simple model consisting of Fully Connected Layers to extract features from this input. The network architecture is described as follows:

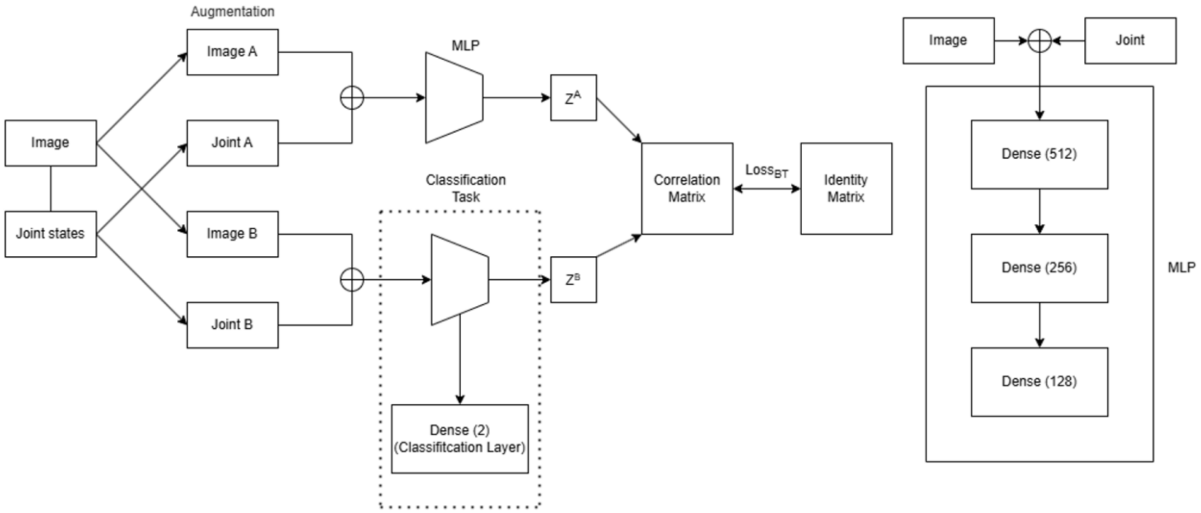


Figure 15 Simple Approach

With the above network architecture, the MLP will be trained using the embedding vector from the final layer (128 dimensions) to compute the Barlow Twins loss. Once the network has been trained, I will add a Fully Connected Layer of size 2 as a Classification layer to classify *self* and *non-self* categories from the dataset created for the classification downstream task. Model performance will be evaluated using accuracy as a metric.

Results: After training, the model predicted only the label '1' (i.e., *similar*) for all input data, indicating that the model failed to learn any meaningful features during training. This issue could arise due to the model's insufficient capacity to capture the features of the images and the relationships between images and joints.

4.1.2 Convolutional neural network (CNN)

After the first experiment, the approach of directly concatenating images with joint states may be deemed unsuitable. This could stem from several reasons, including the significant dimensional disparity between the two data types (the image size of 128x128 is much larger than the mere 8 dimensions of joint states) and the insufficient learning capacity of the basic MLP model. Therefore, I will improve the model by employing a CNN to extract features from the images while also reducing the dimensionality of the image input, making it more compatible for integration with the joint states. The improved model architecture is described as follows:

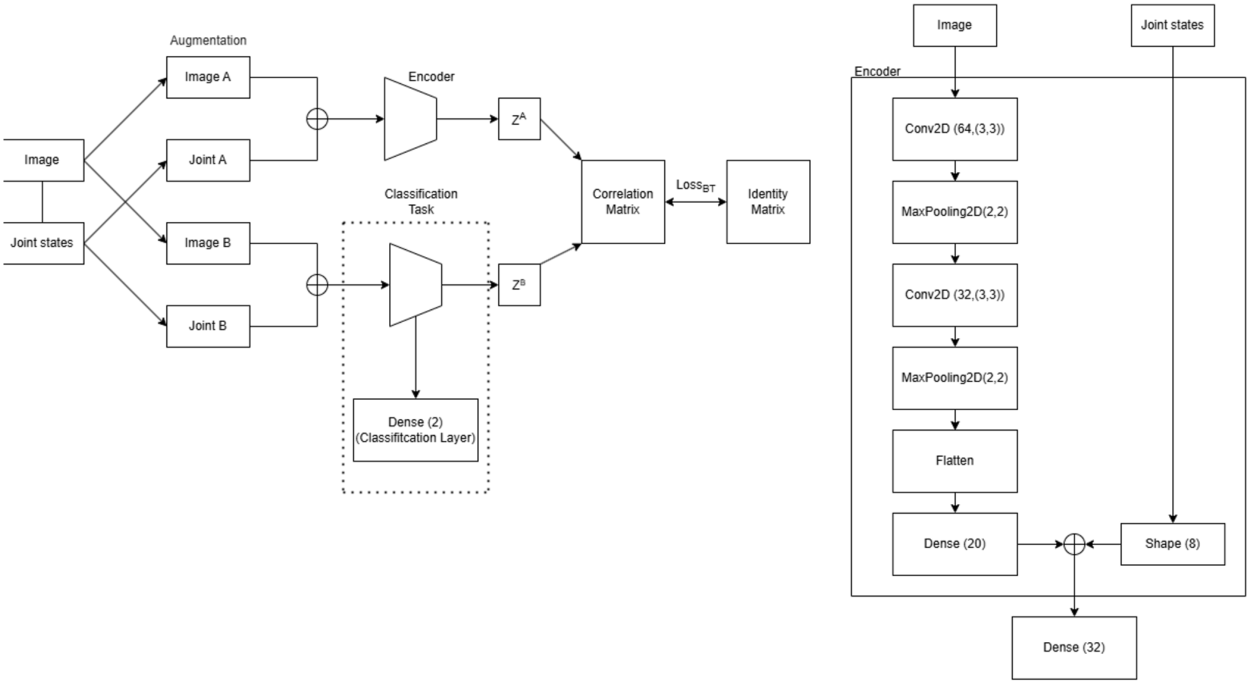


Figure 16 Image & Joints Concatenation with CNN

The key difference in this approach compared to the first is the use of a CNN to extract features from the images, followed by a Fully Connected Layer of size 20. This output is then directly combined with the Joint States, ensuring that the dimensional disparity between the two data types is minimized to prevent bias. However, the model's accuracy still shows no significant improvement, reaching only about 54%. Although the results no longer consist solely of predicting the label '1' as in the first approach, the performance remains weak. With an accuracy close to 50% in a binary classification task, this suggests the model is effectively performing no better than random guessing.

4.2 Loss Combination

Given the infeasible results from previous experiments, I conducted research into alternative approaches for handling multimodality. One suggestion from my supervisor, based on Zambelli et al. (2020) titled ‘Multimodal Representation Models for Prediction and Control from Partial Information’. The paper proposes an architecture based on the **Multimodal Variational Autoencoder (VAE)**, designed to learn and utilize multimodal sensory data from robots. The model demonstrates capabilities such as reconstructing missing data, predicting the robot’s own sensorimotor data and visual trajectories from other data sources, and controlling the robot in an online control loop.

For my research, I will focus on the first two functionalities as the basis for experimental design. The paper utilizes five types of signals: in addition to image and joint states, it incorporates tactile signals, sound signals, and a vector of velocity commands, all captured at the same time. The model consists of independent **encoders** and **decoders** for each sensory modality, as well as a **shared latent representation layer** that combines the outputs from the encoders of each modality. This architecture facilitates learning the relationships between modalities and generating a unified representation in the latent space.

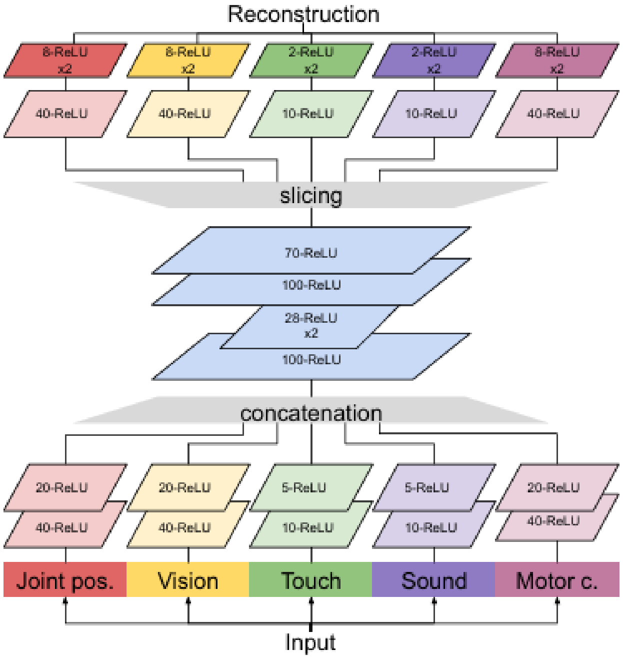


Figure 17 Multimodal Variational Autoencoder (Zambelli et al., 2020)

The loss function of the model is based on the Evidence Lower Bound (ELBO), which consists of two components: reconstruction loss, measuring how well the decoder reconstructs input data from the latent representation, and the Kullback-Leibler (KL) divergence, which regularizes the latent representation by enforcing its alignment with a predefined prior distribution, typically Gaussian. In this study, the focus is primarily on the reconstruction loss, weighted by modality-specific scaling factors to ensure balanced optimization across diverse sensor modalities. As Zambelli described, to balance the reconstruction losses across different sensor modalities, the model assigns weights to each modality based on its dimensionality. Modalities with fewer dimensions, such as tactile and sound, are given higher emphasis by scaling their loss inversely proportional to the number of dimensions ($1/D_m$). The final objective minimizes the sum of these weighted reconstruction losses, ensuring a fair contribution from all modalities.

Building on the suggestion from the above idea, I will employ this approach to compute a combination loss that ensures a balance between the two data types with significantly different sizes. The model architecture is as follows:

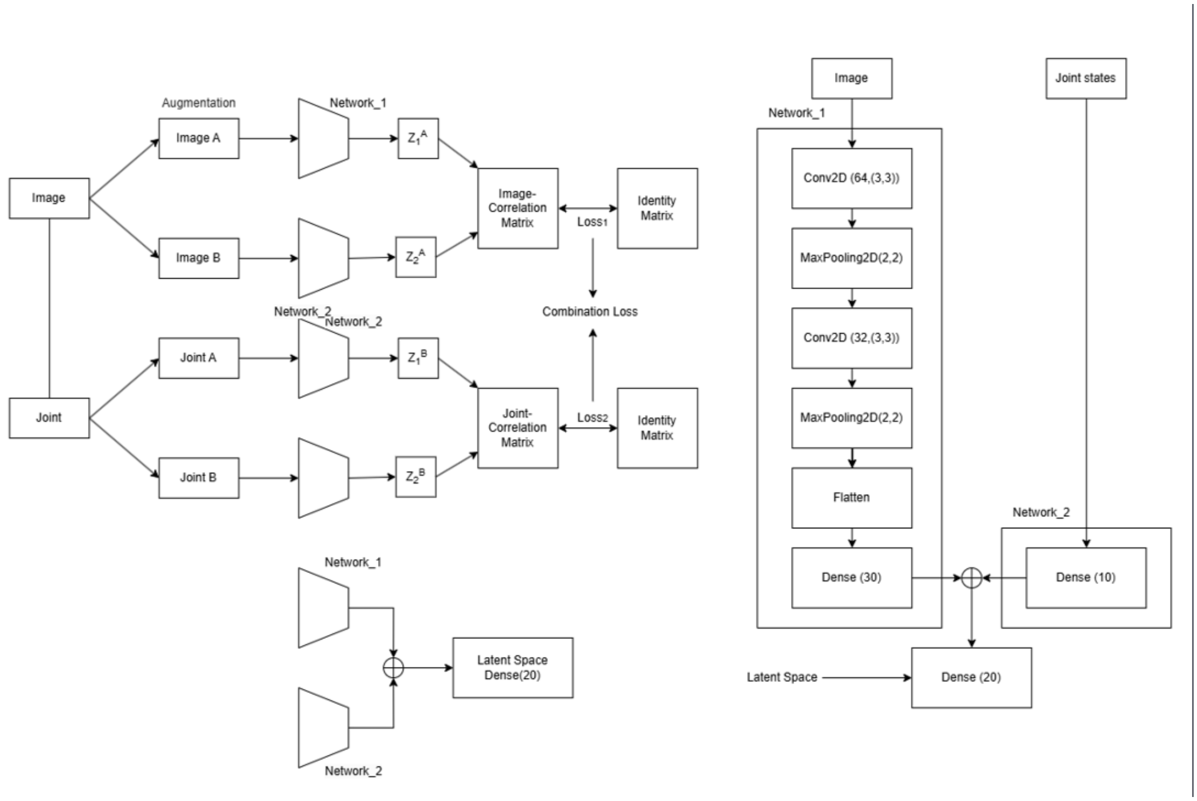


Figure 18 Combination loss approach. Left: the combination of two networks to compute the Barlow Twins loss for each type of data. These separate losses are then used to calculate a Combination Loss, which is subsequently employed to train the model and generate the latent space. Right: the details of the two networks. Network_1 extracts features from images using a CNN, while Network_2 extracts features from joint signals using a Fully Connected Layer. Finally, the embedding vectors from the two models are concatenated and passed through a Fully Connected Layer to produce the final latent space.

As shown in the diagram, I will still use the image and joint states from the same timestamp to maintain their correlation. However, unlike the previous approach where they were concatenated to create a combined embedding vector, I will now use two separate networks for the two data types. For the joint states, given the simple structure of the data (an array of size 8), I will employ an MLP with a layer of shape 10 to generate the embedding vector and then use Barlow Twin to compute a separate loss, Loss_2, for the joint states. For the image data, to extract better features, I will use a CNN with an output layer consisting of a Fully Connected Layer (FCL) of shape 30 to create the embedding vector. Similarly, I will apply Barlow Twin to compute a separate loss, Loss_1, for the image. Finally, the combination loss will be calculated using the following formula:

$$\text{Combination Loss} = \lambda * \text{Loss}_1 + \text{Loss}_2$$

With λ representing the ratio of the embedding dimensions of the image and joint states to balance the dimensions between the two data types, as suggested in the paper, initially, λ was set to 10/30 = 1/3. However, it was later adjusted and reduced to 1/16. Despite this, the results of the classification downstream task remained unchanged, at approximately 50%. As a result, I will experiment with a better feature extraction network for images, replacing CNN with ResNet-50.

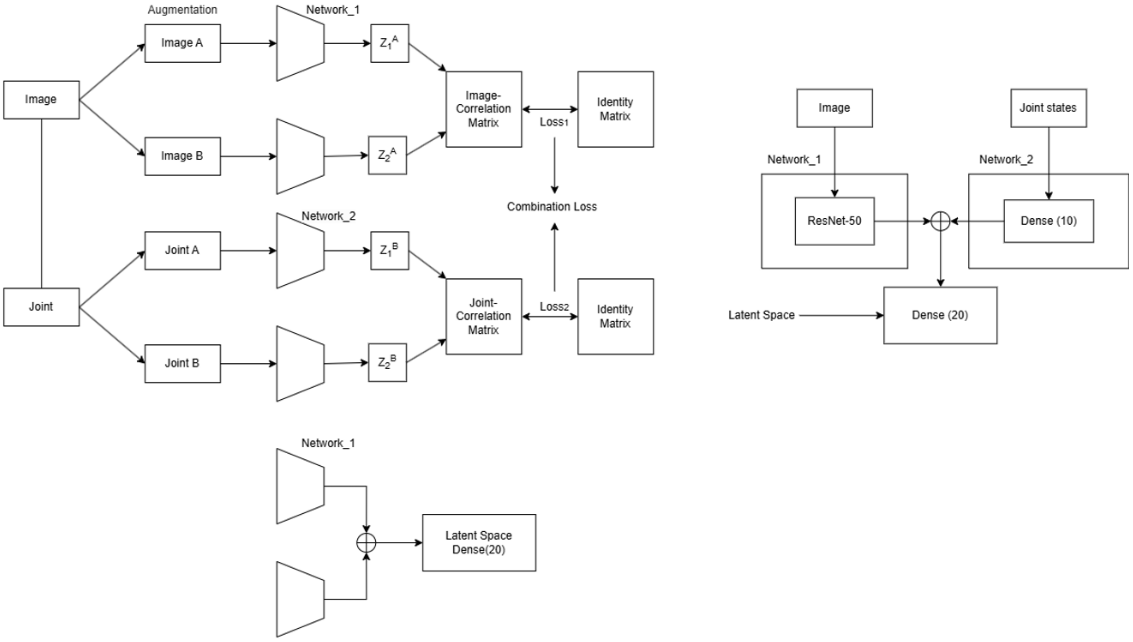
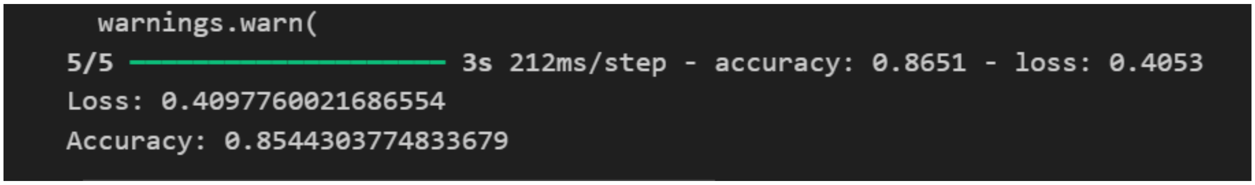


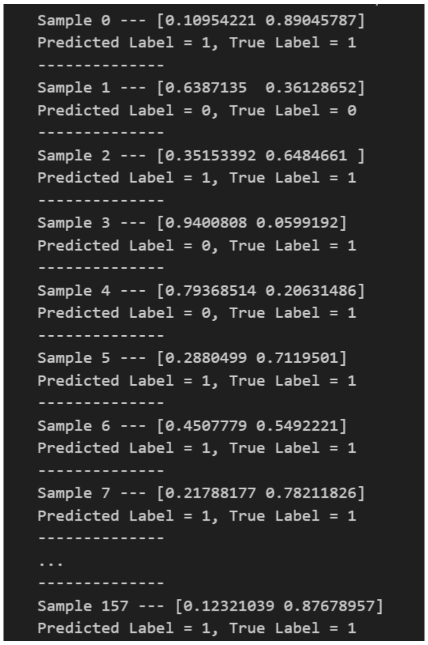
Figure 19 Combination Loss with ResNet-50

This approach, which incorporates pretrained weights for the ResNet-50 network, yielded improved results in the classification task. Therefore, for the subsequent downstream task — the reconstructed model — I will focus on using this network architecture.



```
  warnings.warn(
5/5 ━━━━━━━━━━━━━━━━━━━━ 3s 212ms/step - accuracy: 0.8651 - loss: 0.4053
Loss: 0.4097760021686554
Accuracy: 0.8544303774833679
```

Figure 20 Classification Task with 5 epoch training



```
Sample 0 --- [0.10954221 0.89045787]
Predicted Label = 1, True Label = 1
--------------
Sample 1 --- [0.6387135  0.36128652]
Predicted Label = 0, True Label = 0
--------------
Sample 2 --- [0.35153392 0.6484661 ]
Predicted Label = 1, True Label = 1
--------------
Sample 3 --- [0.9400808 0.0599192]
Predicted Label = 0, True Label = 1
--------------
Sample 4 --- [0.79368514 0.20631486]
Predicted Label = 0, True Label = 1
--------------
Sample 5 --- [0.2880499 0.7119501]
Predicted Label = 1, True Label = 1
--------------
Sample 6 --- [0.4507779 0.5492221]
Predicted Label = 1, True Label = 1
--------------
Sample 7 --- [0.21788177 0.78211826]
Predicted Label = 1, True Label = 1
--------------
...
--------------
Sample 157 --- [0.12321039 0.87678957]
Predicted Label = 1, True Label = 1
```

Figure 21 Classification prediction result with ResNet-50+pretrained weights

4.3 Reconstructed model

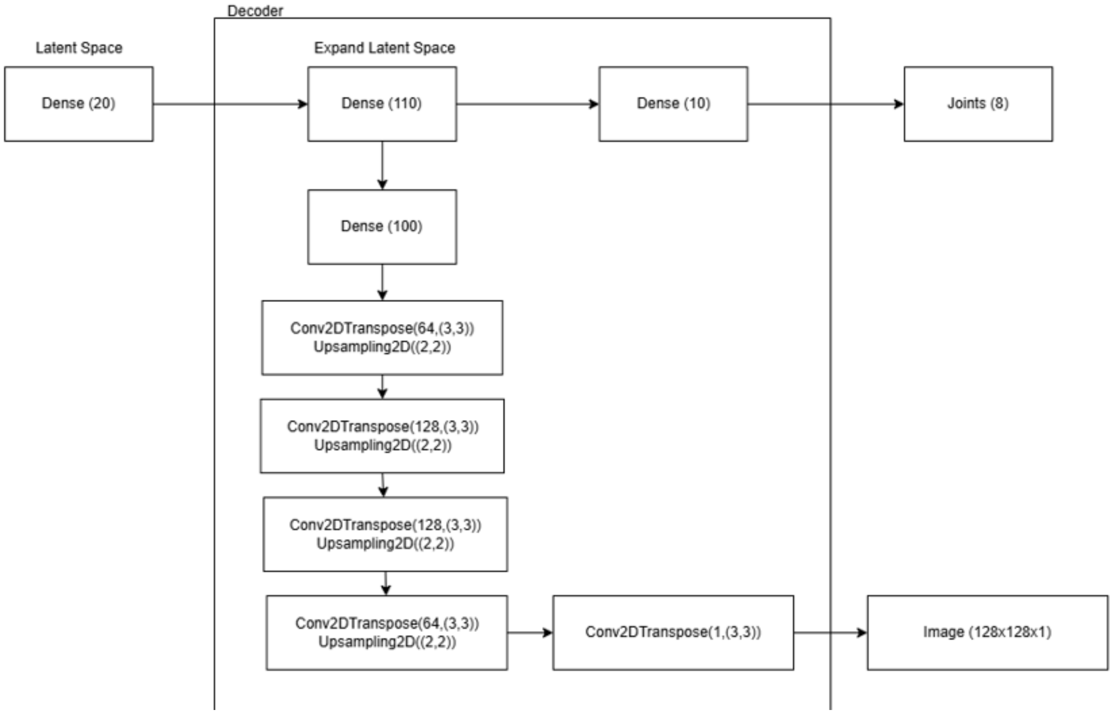


Figure 22 Reconstructed model

In the second downstream task — the Reconstructed model — I will aim to build an encoder-decoder model capable of reconstructing missing data from the latent space created by the encoder in the previous stage. The encoder network will consist of an output of size 20, trained from the combination loss between the two data types, as described earlier. Subsequently, I will construct a decoder network, designed to be the inverse of the encoder network (inspired by the architecture in the Zambelli paper), with the goal of recovering the missing input data. Some experimental results on the validation set are as follows:

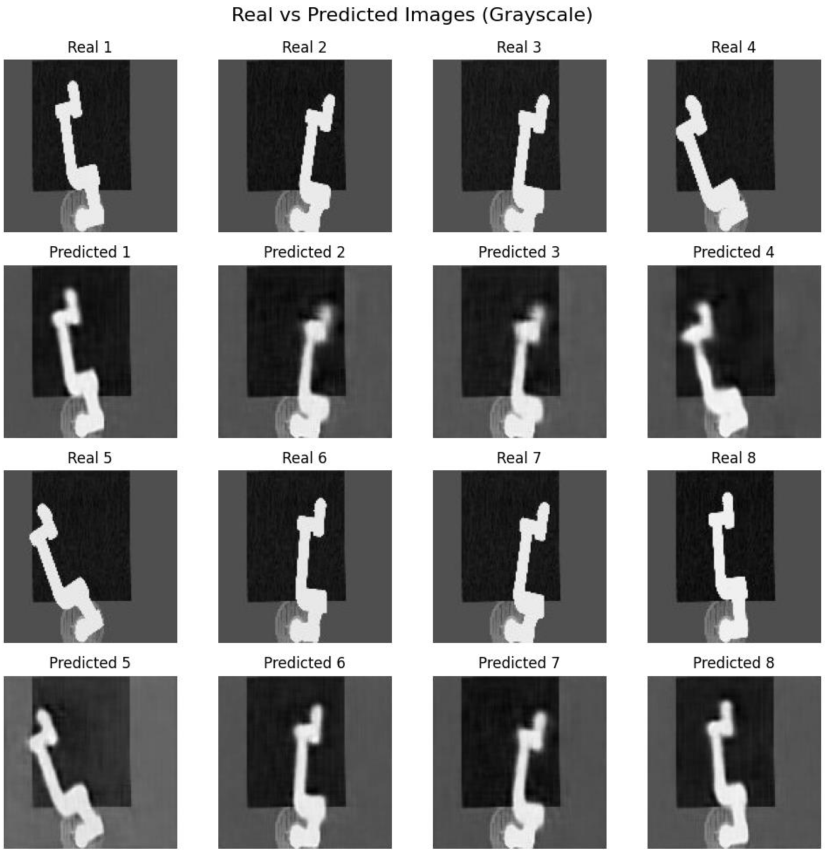


Figure 23 Real Image vs Reconstructed Image (with both image and joint states input)

Figure 23 shows a comparison between the reconstructed image and its actual counterpart, with both the complete image and joint states as input.

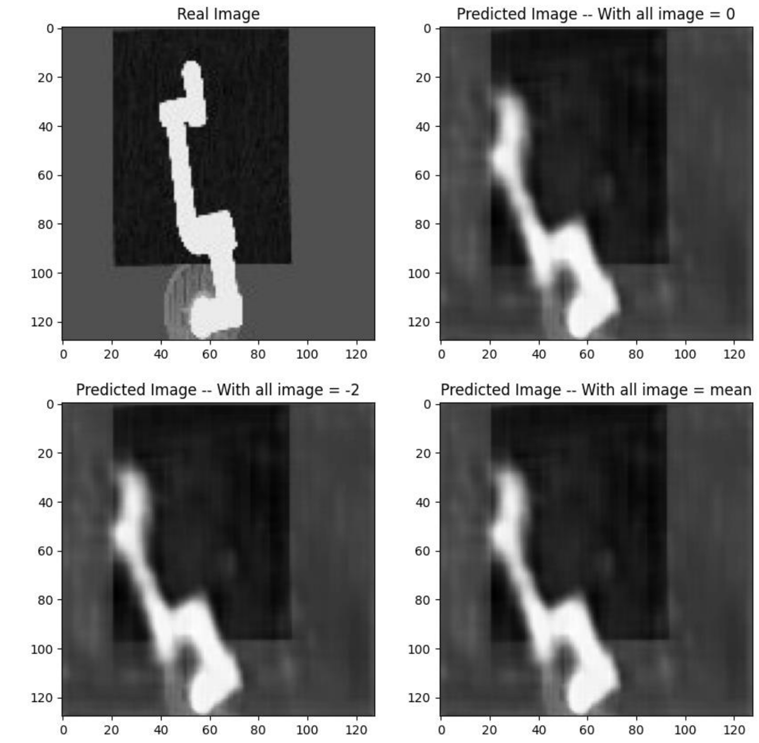


Figure 24 Real Image vs Reconstructed Image without Image Information

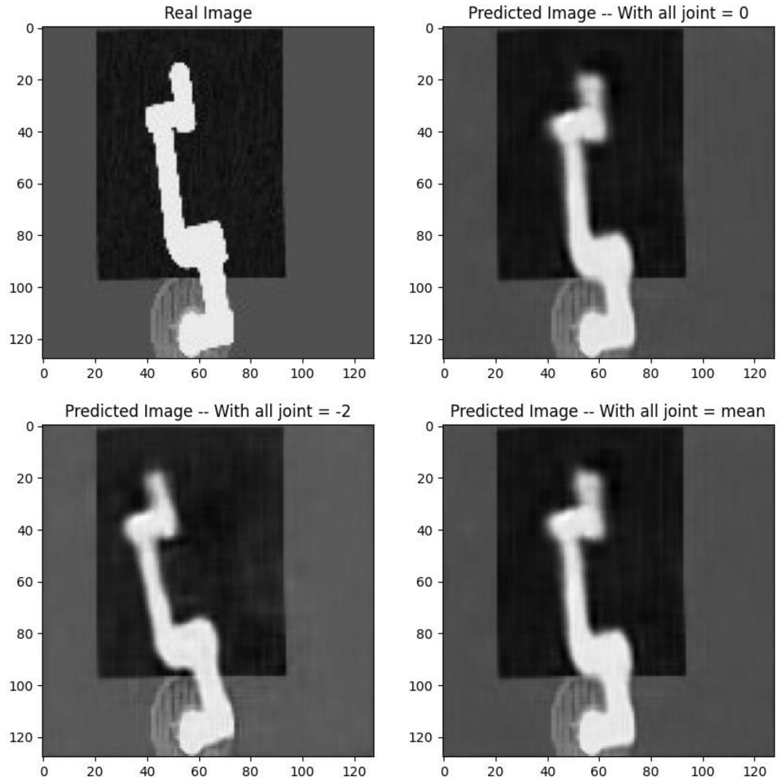


Figure 25 Real Image vs Reconstructed Image without joint information

Meanwhile, Figures 24 and 25 demonstrate experiments on the reconstructed image data under specific conditions. In Figure 24, the joint states data is kept unchanged, but the image data is modified under several conditions: all image pixels set to 0, all image pixels set to -2 (the out-of-range value for the normalized scale [-1,1]), and all image pixels set to the average value of the image. Similarly, for Figure 25, the image data remains unchanged, but the joint states are modified according to the same conditions.

It can be observed that the reconstruction quality is better with image data. The model's performance can potentially be improved through parameter tuning or by incorporating the missing data during the training process. For example, training the dataset with cases where only image data or only joint states are available. More details on this will be discussed in the Discussion section.

CHAPTER 5: DISCUSSION

The experiments have yielded concrete results, such as the use of pre-trained weight models demonstrating superior performance compared to conventional approaches in the classification downstream task, achieving an accuracy of approximately 85%. Additionally, the encoder-decoder model employed in the reconstructed downstream task experiments has shown promising performance, paving the way for further investigations involving diverse input signals, such as audio or tactile data, or even combining multiple signal types simultaneously (e.g., up to five signal types in the MVAE model). These experiments have demonstrated the feasibility of integrating multiple signals by utilizing a single modality model to extract features from various input signals into a shared latent space while applying the Barlow Twins loss method.

However, several limitations and research questions remain. The first issue pertains to the limited dataset, comprising only 597 instances of paired input signals, which may affect the model's effectiveness and accuracy. Furthermore, given the range of experiments conducted, additional time is required for parameter tuning and in-depth evaluation of the models. For instance, in the first approach—Image & Joints concatenation—there was insufficient time to experiment with more advanced networks, such as ResNet50 with pre-trained weights, preventing a comprehensive assessment of this approach. Another experimental constraint was the relatively low number of epochs used during training, which may have contributed to the models not achieving optimal performance. Another limitation involves addressing the question of why the initial models were constrained to an accuracy of approximately 50%. Beyond the possibility of underfitting, other potential causes should be considered. For example, training processes might encounter issues such as the loss function converging to a local minimum or gradient vanishing problems. As depicted in Figure 25, the training loss curve for the Image & Joints concatenation approach, as well as for other underperforming models, exhibits a similar shape. Despite achieving low training loss values, these models underperformed. Further investigation into training parameters could provide insights to mitigate underperformance in these models.

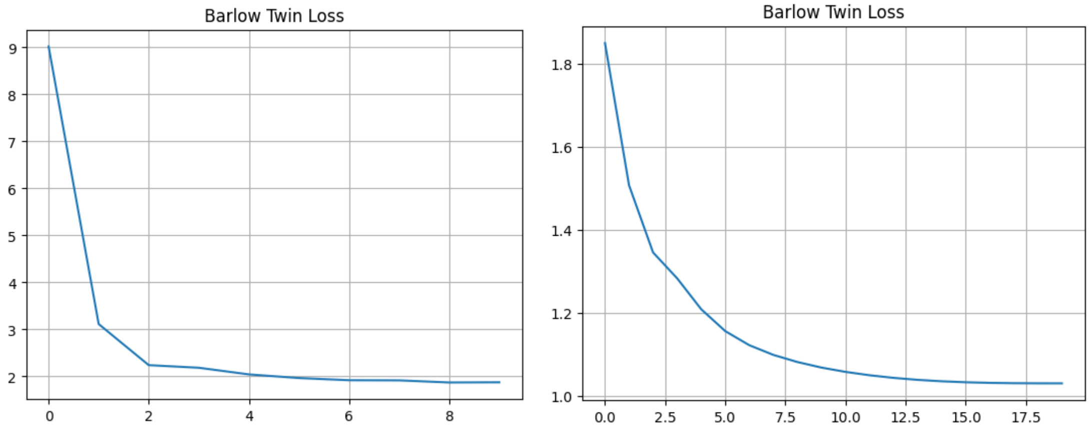


Figure 26 Training Loss in Image & Joints concatenation approach (Left: Initial Approach; Right: CNN Approach)

Future research can explore the utilization of ResNet50 with pretrained weights as a feature extractor for the initial approaches to image representation. This investigation aims to assess whether incorporating ResNet50 can enhance the performance of these methods. Additionally, underperforming approaches should be further refined through hyperparameter tuning and extended training epochs. For the reconstructed model developed, it is essential to test its effectiveness under more challenging datasets. As demonstrated in the MVAE study, training on datasets with missing inputs allowed the model to predict absent data based on remaining signals more effectively. Conducting similar experiments with incomplete datasets could provide valuable insights into the robustness and versatility of the proposed model. Finally, advanced augmentation techniques, such as flipping and rotation, should be applied to the models that have already demonstrated strong performance. This will enable an evaluation of the impact of these augmentation methods on training processes using the Barlow Twins loss function, thereby contributing to a deeper understanding of their role in enhancing model generalization.

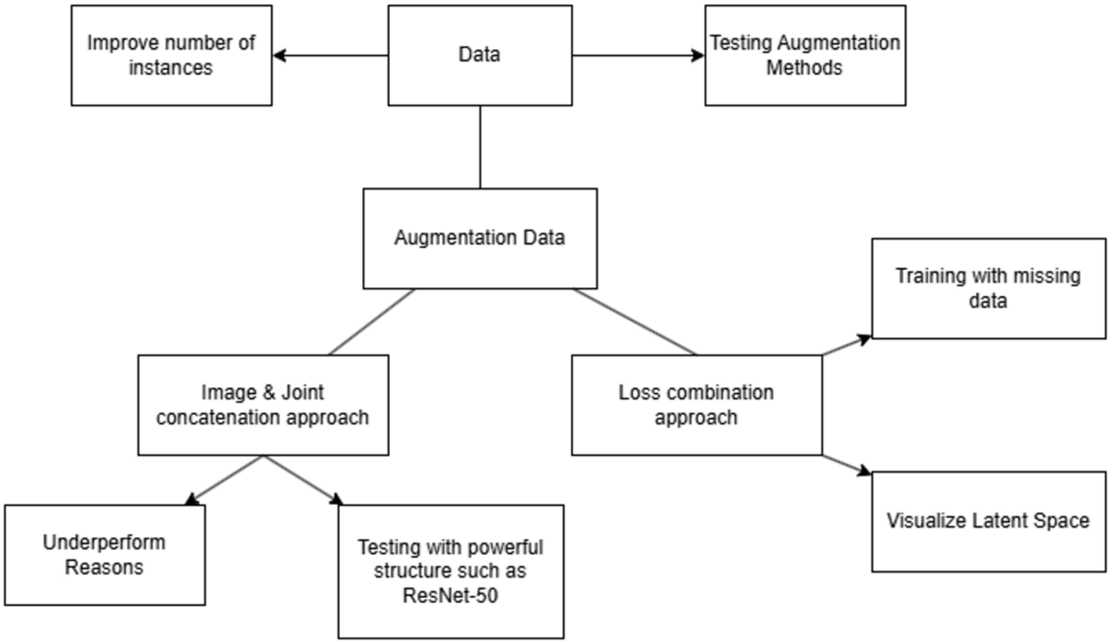

Improve number of instances
Data
Testing Augmentation Methods
Augmentation Data
Training with missing data
Image & Joint concatenation approach
Loss combination approach
Underperform Reasons
Testing with powerful structure such as ResNet-50
Visualize Latent Space

Figure 27 Future work Scope

CHAPTER 6: CONCLUSION

This research has explored the integration of self-supervised learning techniques, particularly the Barlow Twins framework, to advance robotic self-awareness through multi-modal data processing. By combining interoceptive (joint states) and exteroceptive (image) signals, the study aimed to enhance a robot's ability to perceive itself in relation to its environment. The results from the classification and reconstruction experiments demonstrate the feasibility of utilizing latent space representations for downstream tasks. Specifically, the use of ResNet-50 with pre-trained weights significantly improved the model's performance in classification tasks, achieving an accuracy of approximately 85%. Similarly, the encoder-decoder framework for reconstruction tasks successfully demonstrated the potential to recover missing sensor data, even under challenging conditions.

One of the key takeaways is the importance of balancing and integrating diverse data modalities to create a cohesive and meaningful representation. The innovative application of Barlow Twins in reducing redundancy and aligning representations highlights the potential of this approach for robotics applications. Additionally, the study has contributed insights into effective augmentation techniques, parameter tuning, and the design of robust neural architectures for multi-modal tasks.

However, this research is not without its limitations. The relatively small dataset size posed challenges for model generalization, and some early-stage models underperformed due to issues such as local minima and insufficient training epochs. These challenges underscore the need for larger, more diverse datasets and further experimentation with advanced architectures and optimization techniques. Future work could explore the incorporation of additional sensory modalities, such as audio and tactile data, to further enhance the robot's self-perception and its ability to interact with complex environments.

In conclusion, this study represents a significant step forward in bridging theoretical concepts of self-awareness with practical implementations in robotics. By leveraging self-supervised learning and multi-modal integration, this work not only advances robotic cognition but also lays the groundwork for future research in autonomous systems capable of more adaptive and human-like interactions. While the journey towards fully self-aware robots remains long, the findings here contribute valuable knowledge to the field, paving the way for more intelligent, autonomous, and responsive robotic systems.

The project may not have achieved its maximum potential, it has provided me with invaluable insights into applying machine learning and AI techniques to address real-world challenges. I am deeply appreciative of the opportunity to explore and gain knowledge from University's MSc Artificial Intelligence course.

REFERENCES

Allan, D. D., Vonasch, A. J., & Bartneck, C. (2022). The doors of social robot perception: The influence of implicit self-theories. International Journal of Social Robotics, 14(1), 127-140.

Aly, A., Griffiths, S., & Stramandinoli, F. (2017). Metrics and benchmarks in human-robot interaction: Recent advances in cognitive robotics. Cognitive Systems Research, 43, 313-323.

Araujo, A., Norris, W., & Sim, J. (2019). Computing receptive fields of convolutional neural networks. Distill, 4(11), e21.

Argall, B. D., & Billard, A. G. (2010). A survey of tactile human–robot interactions. Robotics and autonomous systems, 58(10), 1159-1176.

Asano, Y., Okada, K., & Inaba, M. (2017). Design principles of a human mimetic humanoid: Humanoid platform to study human intelligence and internal body system. Science Robotics, 2(13), eaaq0899.

Asperti, A., & Tonelli, V. (2023). Comparing the latent space of generative models. Neural Computing and Applications, 35(4), 3155-3172.

Baron-Cohen, S. (1997). Mindblindness: An essay on autism and theory of mind. MIT press.

Bhattacharyya, P., Li, C., Zhao, X., Fehérvári, I., & Sun, J. (2022). Visual representation learning with self-supervised attention for low-label high-data regime. arXiv preprint arXiv:2201.08951.

Bongard, J., Zykov, V., & Lipson, H. (2006). Resilient machines through continuous self-modeling. Science, 314(5802), 1118-1121.

Braun, N., Debener, S., Spychala, N., Bongartz, E., Sörös, P., Müller, H. H., & Philipsen, A. (2018). The senses of agency and ownership: a review. Frontiers in psychology, 9, 535.

Cebollada, S., Payá, L., Flores, M., Peidró, A., & Reinoso, O. (2021). A state-of-the-art review on mobile robotics tasks using artificial intelligence and visual data. Expert Systems with Applications, 167, 114195.

Chatila, R., Renaudo, E., Andries, M., Chavez-Garcia, R. O., Luce-Vayrac, P., Gottstein, R., ... & Khamassi, M. (2018). Toward self-aware robots. Frontiers in Robotics and AI, 5, 88.

DeGrazia, D. (2009). Self-awareness in animals (pp. 201-217). The philosophy of animal minds. Cambridge, England: Cambridge University Press.

Demirel, B. U., & Holz, C. (2024). Finding order in chaos: A novel data augmentation method for time series in contrastive learning. Advances in Neural Information Processing Systems, 36.

Demirel, B., Moulin-Frier, C., Arsiwalla, X. D., Verschure, P. F., & Sánchez-Fibla, M. (2021). Distinguishing self, other, and autonomy from visual feedback: A combined correlation and acceleration transfer analysis. Frontiers in Human Neuroscience, 15, 560657.

Deprez, M., & Robinson, E. C. (2024). Chapter 11 - Convolutional neural networks. In M. Deprez & E. C. Robinson (Eds.), Machine learning for biomedical applications (pp. 233–270). Academic Press. https://doi.org/10.1016/B978-0-12-822904-0.00016-9

Eurich, T. (2018). What self-awareness really is (and how to cultivate it). Harvard Business Review, 4(4), 1-9.

Faisal, A. I., Majumder, S., Mondal, T., Cowan, D., Naseh, S., & Deen, M. J. (2019). Monitoring methods of human body joints: State-of-the-art and research challenges. Sensors, 19(11), 2629.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. Trends in cognitive sciences, 4(1), 14-21.

Gallup Jr, G. G. (1982). Self-awareness and the emergence of mind in primates. American Journal of Primatology, 2(3), 237-248.

Gold, K., & Scassellati, B. (2009). Using probabilistic reasoning over time to self-recognize. Robotics and autonomous systems, 57(4), 384-392.

Gonzalez-Aguirre, J. A., Osorio-Oliveros, R., Rodriguez-Hernandez, K. L., Lizárraga-Iturralde, J., Morales Menendez, R., Ramirez-Mendoza, R. A., ... & Lozoya-Santos, J. D. J. (2021). Service robots: Trends and technology. Applied Sciences, 11(22), 10702.

Gorbenko, A., Popov, V., & Sheka, A. (2012). Robot self-awareness: Exploration of internal states.

Gui, J., et al. (2024). A survey on self-supervised learning: Algorithms, applications, and future trends. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 9052-9071. https://doi.org/10.1109/TPAMI.2024.3415112

Guo, X., Pan, J., Wang, X., Chen, B., Jiang, J., & Long, M. (2023). On the Embedding Collapse when Scaling up Recommendation Models. arXiv preprint arXiv:2310.04400.

Halilovic, A., & Krivic, S. (2024). Robot Explanation Identity. arXiv preprint arXiv:2405.13841.

Hansen, H. O., Jahrens, M., & Martinetz, T. Why Barlow Twins Work: The Critical Role of Normalization and Its Link to Sample Contrastive Learning.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Hikita, M., Fuke, S., Ogino, M., Minato, T., & Asada, M. (2008, August). Visual attention by saliency leads cross-modal body representation. In 2008 7th IEEE International Conference on Development and Learning (pp. 157-162). IEEE.

Hussain, M., Bird, J. J., & Faria, D. R. (2019). A study on CNN transfer learning for image classification. In Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK (pp. 191-202). Springer International Publishing.

Huttunen, A. W., Adams, G. K., & Platt, M. L. (2017). Can self-awareness be taught? Monkeys pass the mirror test—again. Proceedings of the National Academy of Sciences, 114(13), 3281-3283.

Jing, L., Vincent, P., LeCun, Y., & Tian, Y. (2021). Understanding dimensional collapse in contrastive self-supervised learning. arXiv preprint arXiv:2110.09348.

Kingma, D. P. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Koga, Y., Kawaharazuka, K., Toshimitsu, Y., Nishiura, M., Omura, Y., Asano, Y., ... & Inaba, M. (2021). Self-body image acquisition and posture generation with redundancy using musculoskeletal humanoid shoulder complex for object manipulation. IEEE Robotics and Automation Letters, 6(4), 6686-6692.

Kohda, M., Hotta, T., Takeyama, T., Awata, S., Tanaka, H., Asai, J. Y., & Jordan, A. L. (2019). If a fish can pass the mark test, what are the implications for consciousness and self-awareness testing in animals?. PLoS biology, 17(2), e3000021.

Kojima, H., & Ikegami, T. (2022). Organization of a Latent Space structure in VAE/GAN trained by navigation data. Neural Networks, 152, 234-243.

Kurka, P. R. G., & Salazar, A. A. D. (2019). Applications of image processing in robotics and instrumentation. Mechanical Systems and Signal Processing, 124, 142-169.

Lanillos, P., & Cheng, G. (2020). Robot self/other distinction: Active inference meets neural networks learning in a mirror. In ECAI 2020 (pp. 2410-2416). IOS Press.

Lanillos, P., & Cheng, G. (2020). Robot self/other distinction: active inference meets neural networks learning in a mirror. In ECAI 2020 (pp. 2410-2416). IOS Press.

Lanillos, P., Dean-Leon, E., & Cheng, G. (2016). Yielding self-perception in robots through sensorimotor contingencies. IEEE Transactions on Cognitive and Developmental Systems, 9(2), 100-112.

Lanillos, P., Dean-Leon, E., & Cheng, G. (2017). Yielding self-perception in robots through sensorimotor contingencies. IEEE Transactions on Cognitive and Developmental Systems, 9(2), 100-112. https://doi.org/10.1109/TCDS.2016.2627820

Laurent, T., von Brecht, J. H., & Bresson, X. (2023). Feature Collapse. arXiv preprint arXiv:2305.16162.

Le-Khac, P. H., Healy, G., & Smeaton, A. F. (2020). Contrastive representation learning: A framework and review. Ieee Access, 8, 193907-193934.

Lebiere, C., Jentsch, F., & Ososky, S. (2013). Cognitive models of decision making processes for human-robot interaction. In Virtual Augmented and Mixed Reality. Designing and Developing Augmented and Virtual Environments: 5th International Conference, VAMR 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part I 5 (pp. 285-294). Springer Berlin Heidelberg.

Legaspi, R., He, Z., & Toyoizumi, T. (2019). Synthetic agency: sense of agency in artificial intelligence. Current Opinion in Behavioral Sciences, 29, 84-90.

Leitner, J., Harding, S., Frank, M., Förster, A., & Schmidhuber, J. (2013). An integrated, modular framework for computer vision and cognitive robotics research (icVision). In Biologically Inspired Cognitive Architectures 2012: Proceedings of the Third Annual Meeting of the BICA Society (pp. 205-210). Springer Berlin Heidelberg.

Levesque, H., & Lakemeyer, G. (2008). Cognitive robotics. Foundations of artificial intelligence, 3, 869-886.

Lewis, M. (1995). Shame: The exposed self. Simon and Schuster.

Li, J., Li, Z., Chen, F., Bicchi, A., Sun, Y., & Fukuda, T. (2019). Combined sensing, cognition, learning, and control for developing future neuro-robotics systems: a survey. IEEE Transactions on Cognitive and Developmental Systems, 11(2), 148-161.

Liu, B., Zhao, W., & Sun, Q. (2017, October). Study of object detection based on Faster R-CNN. In 2017 Chinese automation congress (CAC) (pp. 6233-6236). IEEE.

Liu, S., Cui, W., Wu, Y., & Liu, M. (2014). A survey on information visualization: recent advances and challenges. The Visual Computer, 30, 1373-1393.

Liu, Z., Alavi, A., Li, M., & Zhang, X. (2024). Guidelines for Augmentation Selection in Contrastive Learning for Time Series Classification. arXiv preprint arXiv:2407.09336.

Loureiro, F., Avelino, J., Moreno, P., & Bernardino, A. (2022). Self-perception of interaction errors through human non-verbal feedback and robot context. In F. Cavallo et al. (Eds.), Social Robotics. ICSR 2022. Lecture Notes in Computer Science (Vol. 13818, pp. 531-541). Springer. https://doi.org/10.1007/978-3-031-24670-8_42

Lu, Y., Chen, C., Chen, P., & Yu, S. (2023). Designing social robot for adults using self-determination theory and AI technologies. IEEE Transactions on Learning Technologies, 16(2), 206-218.

Luo, W., Li, Y., Urtasun, R., & Zemel, R. (2016). Understanding the effective receptive field in deep convolutional neural networks. Advances in neural information processing systems, 29.

Makedon, V., Mykhailenko, O., & Vazov, R. (2021). Dominants and Features of Growth of the World Market of Robotics. European Journal of Management Issues, 29(3), 133-141.

Mentzou, A., & Ross, J. (2024). The Emergence of Self-Awareness: Insights from Robotics. Human Development, 68(2), 90-100.

Michel, P., Gold, K., & Scassellati, B. (2004, September). Motion-based robotic self-recognition. In 2004 IEEE/RSJ international conference on intelligent robots and systems (IROS)(IEEE Cat. No. 04CH37566) (Vol. 3, pp. 2763-2768). IEEE.

Moore, C., Lemmon, K., & Skene, K. (Eds.). (2001). The self in time: Developmental perspectives. Psychology Press.

Moore, J. W. (2016). What is the sense of agency and why does it matter?. Frontiers in psychology, 7, 1272.

Moulin-Frier, C., Fischer, T., Petit, M., Pointeau, G., Puigbo, J. Y., Pattacini, U., ... & Verschure, P. F. (2017). DAC-h3: a proactive robot cognitive architecture to acquire and express knowledge about the world and the self. IEEE Transactions on Cognitive and Developmental Systems, 10(4), 1005-1022.

Nemati, S., Rohani, R., Basiri, M. E., Abdar, M., Yen, N. Y., & Makarenkov, V. (2019). A hybrid latent space data fusion method for multimodal emotion recognition. IEEE Access, 7, 172948-172964.

Nguyen, B., Takida, Y., Murata, N., Uesaka, T., Ermon, S., & Mitsufuji, Y. (2024). Mitigating Embedding Collapse in Diffusion Models for Categorical Data. arXiv preprint arXiv:2410.14758.

Pak, M., & Kim, S. (2017, August). A review of deep learning in image recognition. In 2017 4th international conference on computer applications and information processing technology (CAIPT) (pp. 1-3). IEEE.

Parmiggiani, A., Maggiali, M., Natale, L., Nori, F., Schmitz, A., Tsagarakis, N., ... & Metta, G. (2012). The design of the iCub humanoid robot. International journal of humanoid robotics, 9(04), 1250027.

Pennisi, P., Tonacci, A., Tartarisco, G., Billeci, L., Ruta, L., Gangemi, S., & Pioggia, G. (2016). Autism and social robotics: A systematic review. Autism Research, 9(2), 165-183.

Pipitone, A., & Chella, A. (2021). Robot passes the mirror test by inner speech. Robotics and Autonomous Systems, 144, 103838.

Prescott, T. J., & Dominey, P. F. (2024). Synthesizing the temporal self: robotic models of episodic and autobiographical memory. Philosophical Transactions B, 379(1913), 20230415.

Prescott, T. J., Vogeley, K., & Wykowska, A. (2024). Understanding the sense of self through robotics. Science Robotics, 9(95), eadn2733. https://doi.org/10.1126/scirobotics.eadn2733

Rascon, C., & Meza, I. (2017). Localization of sound sources in robotics: A review. Robotics and Autonomous Systems, 96, 184-210.

Raza, K., Khan, T. A., & Abbas, N. (2018). Kinematic analysis and geometrical improvement of an industrial robotic arm. Journal of King Saud University-Engineering Sciences, 30(3), 218-223.

Rolf, M., & Asada, M. (2014, October). Autonomous development of goals: From generic rewards to goal and self detection. In 4th International Conference on Development and Learning and on Epigenetic Robotics (pp. 187-194). IEEE.

Roncone, A., Hoffmann, M., Pattacini, U., Fadiga, L., & Metta, G. (2016). Peripersonal space and margin of safety around the body: learning visuo-tactile associations in a humanoid robot with artificial skin. PloS one, 11(10), e0163713

Ruggiero, A., Mahr, D., Odekerken-Schröder, G., Spena, T. R., & Mele, C. (2022). Companion robots for well-being: a review and relational framework. Research handbook on services management, 309-330.

Saegusa, R., Metta, G., Sandini, G., & Sakka, S. (2009, February). Active motor babbling for sensorimotor learning. In 2008 IEEE International Conference on Robotics and Biomimetics (pp. 794-799). IEEE.

Scassellati, B., Admoni, H., & Matarić, M. (2012). Robots for use in autism research. Annual review of biomedical engineering, 14(1), 275-294.

Sharma, V., & Singh, N. (2021, November). Deep convolutional neural network with ResNet-50 learning algorithm for copy-move forgery detection. In 2021 7th International conference on signal processing and communication (ICSC) (pp. 146-150). IEEE.

Singh, S., Chaudhary, D., Gupta, A. D., Lohani, B. P., Kushwaha, P. K., & Bibhu, V. (2022, April). Artificial intelligence, cognitive robotics and nature of consciousness. In 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM) (pp. 447-454). IEEE.

Stoytchev, A. (2011). Self-detection in robots: a method based on detecting temporal contingencies. Robotica, 29(1), 1-21.

Taniguchi, T., Mochihashi, D., Nagai, T., Uchida, S., Inoue, N., Kobayashi, I., ... & Inamura, T. (2019). Survey on frontiers of language and robotics. Advanced Robotics, 33(15-16), 700-730.

Tellex, S., Gopalan, N., Kress-Gazit, H., & Matuszek, C. (2020). Robots that use language. Annual Review of Control, Robotics, and Autonomous Systems, 3(1), 25-55.

Tsai, Y. H. H., Bai, S., Morency, L. P., & Salakhutdinov, R. (2021). A note on connecting barlow twins with negative-sample-free contrastive learning. arXiv preprint arXiv:2104.13712.

Turner, J. C., & Reynolds, K. J. (2011). Self-categorization theory. Handbook of theories in social psychology, 2(1), 399-417.

Woźniak, M. (2018). “I” and “Me”: the self in the context of consciousness. Frontiers in psychology, 9, 1656.

Yuan, X., Lin, Z., Kuen, J., Zhang, J., Wang, Y., Maire, M., ... & Faieta, B. (2021). Multimodal contrastive training for visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6995-7004).

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021, July). Barlow twins: Self-supervised learning via redundancy reduction. In International conference on machine learning (pp. 12310-12320). PMLR.

Zhang, J., & Ma, K. (2022). Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16650-16659).

Zhang, Q., Wang, Y., & Wang, Y. (2023, July). On the generalization of multi-modal contrastive learning. In International Conference on Machine Learning (pp. 41677-41693). PMLR.

Zhang, Y., Zhu, H., Song, Z., Koniusz, P., & King, I. (2023). Mitigating the popularity bias of graph collaborative filtering: A dimensional collapse perspective. Advances in Neural Information Processing Systems, 36, 67533-67550.