Artificial Self-Awareness for Robots

Ali H H A H AlQallaf

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Computing Science
College of Science and Engineering
University of Glasgow

August 2022

Abstract

Robots are evolving and entering into various sectors and aspects of life. While humans are aware of their bodies and capabilities, which help them work on a task in different environments, robots are not. This thesis is about defining and developing a robotic artificial self-awareness framework. The aim is to allow robots to adapt to their environment and better manage their task. The robot's artificial self-aware knowledge is captured based on levels where each level helps a robot acquire higher self-awareness competence. These levels are inspired by Rochat [1] self-awareness development levels in humans, where each level is associated with a complexity of self-knowledge. Self-awareness in humans leads to distinguishing themselves from the environment, allowing humans to understand themselves and control their capabilities. This work focuses on the first and second levels of self-awareness through differentiation and situation (minimal self).

The artificial self-awareness level-1 proposes the first step towards a basic, minimal self-awareness in a robot. The artificial self-awareness level-2 proposes an increasing capacity of self-awareness knowledge in the robot. That is, this thesis posits an experimental methodology to evaluate whether the robot can differentiate and situate itself from the environment and to test whether artificial self-awareness level-1 and level-2 increase a robot's self-certainty in an unseen environment.

The research utilises deep neural network techniques to allow a dual-arm robot to identify itself within different environments. The robot vision and proprioception are captured using a camera and robot sensors to build a model that allows a robot to differentiate itself from the environment. The level-1 results indicate that a robot can distinguish itself with an accuracy of 80.3% on average in different environmental settings and under confounding input signals. Also, the level-2 results show that a robot can situate itself in different environments with an accuracy of 86.01% yielding a higher artificial self-certainty of 5.71%. This thesis work helps a robot be aware of itself in different environments.

Contents

List of Tables

List of Figures

Acknowledgements

I truly enjoyed every aspect of doing my PhD life. It was long and fun, with moments of a different trend. Whilst I am highly interested in my topic area, my family, friends, and the University of Glasgow members made it a pleasant experience. I accomplished the work of my PhD journey with excellent and remarkable support from the people around me, which made me a better researcher and person.

Firstly, I thank God for blessing my family and me and giving me the patience and strength to finish my research journey.

I am grateful to my mother for consistently praying for me and for her guidance. I express my heartfelt gratitude to my beloved wife "Noor" for her patience, for taking care of the family, and for her support over the PhD years; she was driving the family boat to the safe land when I was busy with my PhD research. I deeply appreciate my gorgeous children, Mohammed, Hoor, Hussain, and Haneen, for their patience and encouragement during my study. They were also a source of inspiration for my research. I thank my sisters and brothers for their care and encouragement.

I will always be hugely grateful to my supervisor, Dr Gerardo Aragon-Camarasa, for his outstanding support throughout my PhD journey and his care in every detail. He motivates me in all situations, facilitating my PhD journey and making it brilliant. Thanks, Gerardo, for your consistent guidance, mentorship, and availability throughout my PhD years; you are both a supervisor and a brother. I thank my second supervisor, Dr Paul Paul Siebert, who gave me constructive feedback and guidance, which helped strengthen my research.

Special thanks to fellow PhD colleagues of Computer Vision and Autonomous Systems (CVAS) members, who over the PhD years, also became my friends: Amaya, Piotr, Ozan, Nikos, Li, Florent, and other members; thank you all for your consistent encouragement.

This thesis is dedicated to my father's soul, Hameed AlQallaf, who taught me to seek and be hungry for knowledge.

Declaration

I declare that this thesis was composed by myself, the work contained in this thesis is my own with the exception of chapters 2, which contains introductory material, all work in this thesis was carried out by the author unless otherwise explicitly stated in the text.

Ali AlQallaf

Chapter 1

Introduction

This research aims to define and devise artificial self-awareness in a robot. Robotic artificial self-awareness is a novel research topic attracting increasing attention in recent years. This topic is relatively new and the literature is scarce. This is a challenging topic compared to traditional self-aware robots [2] [3] since more knowledge about the self-aware robot, multi-data sources integration, and different environments are required. This can be achieved by processing the integration of different robot senses in different environments to define the self. Furthermore, inspired by human self-awareness development levels, this thesis develops two levels of self-awareness and proposes a modular approach that implements the levels of artificial self-awareness in which each level process and decode an incremental complexity of self. Level-1 is employed to integrate the robot sensors inputs, achieving an initial self-recognition. The artificial self-awareness level-2 is employed to encode the temporal integration of level-1, allowing a robot to recognise itself over time.

1.1 Human Self-Awareness

Rochat [1] discusses that self-awareness is essential perceptual experience individuals acquire during their early life stages. When humans become self-aware, they can recognise themselves in any environment. This is possible because they can distinguish their body as separate entities from the world, allowing them to adapt to different situations and scenarios.

Rochat [1] has classified self-awareness into five levels, starting from sensing self as a separate entity in the world (Level 1) to self-consciousness (Level 5). Later, Rochat [4] proposed that self-unity (Level 0) is the primary phase of newborns which comprises the initial sensory experience during the first hours of life, and concluded that self-unity could endow machines to learn about their body within an environment. The ordering of the five levels of self-awareness is based on their relative complexity, and it is further divided into implicit (from zero to two), and explicit (from three to five) levels [1], [5]. Legrain et al. [5] have formulated that the implicit self-awareness levels are related to correlating the internal states with the body based on

the experience of the self within the environment. The explicit self-awareness levels are those that link the environment to how the environment influences the person. Gulick [6] emphasises that recognising self is a concept of self-awareness that allows a human to be adaptable and flexible in control. Rochat [7] mentioned that the self is what distinguishes humans from machines. He also commented that "it gives purpose and orientation to the actions performed by that organism" [7].

Self-awareness studies exist in the cognitive literature, which gives the current account of the understanding of human architecture and attempts to link self-awareness to digital computers. As Laure et al. [5] discussed, human self-awareness consists of different levels of complexity with different scales of development ability. At each self-awareness level, the acquisition of awareness and knowledge increases, and the individual becomes more self-aware. According to Rochat [1], these levels are:-

- **Level Zero – "Self-unity"**: The individual is born with basic multi-sensory and motor control capabilities, which they use to learn about themselves.
- **Level One –" Differentiation"**: The individual gets a sense of unique experience between what is out there and the felt movements, which initiate the sense of self.
- **Level Two –" Situation"**: Individual situates within its body by experiencing the relationship between seen movements and body stimulation over time.
- Level three - "Identification": The person develops self-existence in the environment, singles out and explicitly refers to self as " Me".
- Level four - "Permanence": The individual can identify himself by looking at a picture or video in different locations.
- Level five - "Self-consciousness": the person starts to care about people's perspectives towards oneself and starts to show emotions.

Rochat's levels of self-awareness inspire this research study. Therefore, the study presents a robotic artificial self-awareness by incorporating the implicit levels described above (bolded) in a robot. Mapping the implicit levels to a robotic artificial self-awareness will enable the robot to recognise and be aware of itself in the environment.

1.2 Artificial Self-Awareness

The main aim of this research is to define and devise self-awareness in robotics. Artificial self-awareness is a novel research topic and has attracted attention from cognitive, robotic, and AI researchers in recent years. Self-awareness is vital in humans, but it does not exist in robots. Despite the state-of-art in current robotic developments, robots still lack self-awareness. With

current advances in multi-sensory platforms, it is time to utilise these capabilities and construct a robot that can dynamically interact with humans and other robotic agents. Developing artificial self-awareness will allow a robot to characterise itself and distinguish itself as an entity in the world, making a potential path to comprehend its capabilities within its environment. Therefore, understanding self-awareness and adapting it to a robot represents a fundamental step in robotics. This research thus proposes to devise artificial self-aware modules which will allow the robot to distinguish itself and increase its self-certainty in different environments.

Implementing the first implicit levels of self-awareness mentioned in section 1.1 in a robot artificially allows it to get a degree of awareness and, consequently, acquire knowledge about itself.

The description for each level from the artificial perspective is as follows (direct mapping to Rochat's [1] internal self-aware levels):

- **Level-0**: corresponds to the robot's initialisation and ability to get inputs using the camera and proprioception. This level comprises the robot's kinematic structure, sensor definition and configuration, and motion planning.

- **Level-1**: This level of artificial self-awareness is constructed by letting the robot learn to differentiate itself by seeing its arms, including hands and gripers, in association with the proprioception data. The robot gets a high-level description of its limbs' appearance and will be able to confirm if the observed hand belongs to it. At this level, the artificial self-awareness model is initiated with an initial sense of self.

- **Level-2**: The robot has figured out that both hands belong to it and constructed an idea about its body parts using the level-1 knowledge. After getting an initial sense of self, the robot situation model will be trained by moving its hands to relate its previous spatial and temporal state. Here, the robot will have confidence about its existence and relate seen moving parts to itself by predicting its next state over time.

This thesis proposed the methods and implementation of the above artificial self-awareness levels to increase robot self-certainty in a different environment.

1.3 Problems and Challenges

1.3.1 Associated With a Robotic Task

Robots are used in many sectors programmed to work on a certain task [8] [9]. Robotic researchers are trying to push robots to achieve human capabilities [10] [11] [12]. One of the missing pieces is self-awareness in the robots [13]. While humans are aware of their bodies and capabilities, which help them work on a task in different environments [1] [7], robots are not. Because of that, robots usually are static in their behaviour and mostly, they cannot handle

environmental changes [14]. Therefore, most robots operate in limited environments that are not complex to avoid environmental uncertainties that might affect the robots' tasks.

In addition, some robots interact with humans; they must be safe to avoid harming people around them [15] [16]. Robot companies equip their robots with sensors that sense the environment around them to avoid incidents [17]. Still, the uncertainty from humans or the environment might affect the sensors in a way that harms humans or the environment [18]. However, humans are more flexible and adaptable to different environments because they can recognise themselves within a different environment and use their self-knowledge to control their environments [1] [19].

1.3.2 The Artificial Self-awareness Challenges

The availability of a self-awareness dataset is one of the challenges associated with artificial self-awareness, as no dataset defined for self-awareness can be used to devise a robot with self-awareness artificially. Sensory fusion is important to achieve self-awareness in a robot artificially. However, fusing different modalities is challenging. Furthermore, creating a self-aware agent has no unified, straightforward technique. To achieve self-awareness for a robot artificially, many researchers have used different methods and sensors to attempt to devise the artificial self. Researchers of artificial self-awareness studies [20] [21] [2] [22] [23] [24] [25] are inspired by the cognitive studies that study human self-awareness, which also have different ways and theories of interpreting human awareness.

1.4 Research Motivation

Self-awareness is an essential perceptual experience, and individuals acquire self-awareness during their early life stages [1]. Self-awareness allows humans to control and understand their bodies, interact with others, and evaluate their potential for doing tasks according to their physical capabilities. The emergence of self-awareness in humans leads to distinguishing themselves from the environment and allows humans to comprehend themselves and control their physical capabilities [7]. Self-awareness is vital in humans, but it does not exist in robots. Robots still lack self-awareness despite the state of the art in current robotic developments. Therefore, understanding self-awareness and adapting it represents a fundamental step in robotics to extend the robot's physical capability and dynamicity. The human self-awareness inspiration has been devised into synthetic models that artificially allow a robot to be aware of itself in different environments. Therefore, this thesis proposes creating artificial self-aware modules which will allow the robot to distinguish itself and increase its autonomy.

Artificial self-awareness will let a robot know itself, guiding it to distinguish itself as an entity in the world, making a potential path to comprehend its physical capabilities within its

ecological self where it investigates its existence in different environments. Recognising and knowing its capabilities based on itself helps a robot perform better in different environments and interact with humans while perceiving itself and operating with other robots. This will give a robot better control over its known self in different environments and settings. The current interaction between the robots and humans is not triggered from within the robot itself; it is commonly static from recognising the environment and taking action based on the environmental stimuli. If a robot can dynamically evaluate its physical capabilities, it would add more self-control and make robots safer interacting with humans and machines. Moreover, if robots are supported by artificial self-awareness and interact with each other, engagement occurs with better control, which implies better robot cooperation. As a result, self-awareness will greatly impact different robotic fields, and their applications [26].

A robot has different layers that control its hardware and layers that interact with each other to exchange information about the robot's internal status. In Fig. 1.1 Self-awareness can support a layered robotic architecture by integrating a module specific to self-awareness. Combining different robot information with self-awareness information can form a useful context that benefits supporting a robot's tasks; for example, working on collaborative grasping, the robot will be able to know that it is the one doing this particular task. Accordingly, this robot's self-knowledge can be utilised to support different behaviours toward the intended task.

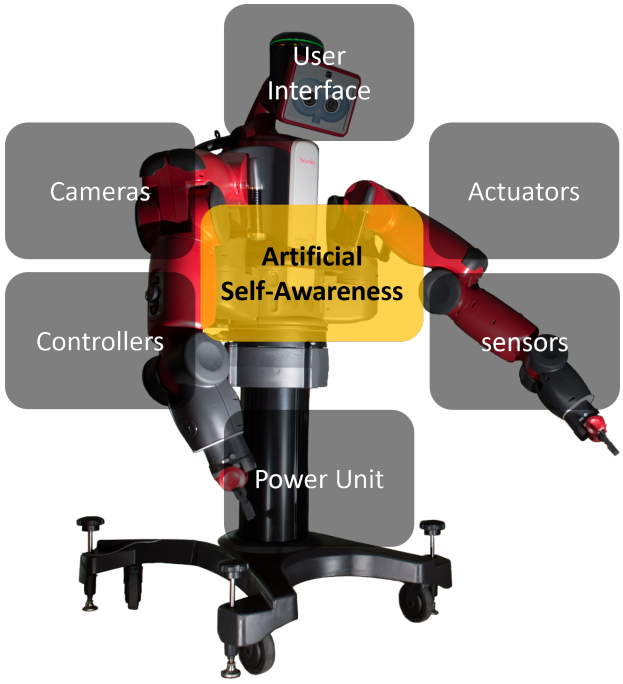

Figure 1.1: A robot with different hardware layers is used to accomplish a task. The self-awareness layer uses different layers and supports tasks.

1.5 Aims and Objectives

Inspired by human self-awareness development and the robot's sensory abilities, the artificial self-awareness scope is defined and integrated into a robot. Current research defined self-agents for theoretical purposes or as proof of concept, for example, including a mirror to state the robot existence [27] [28] [29]. This thesis aims to define artificial self-awareness and use it to support a wide range of robotic tasks. Robotic artificial self-awareness makes a robot recognise and be aware of itself to communicate with others [26], carry out tasks efficiently [30], and collaborate with others [26] [23]. Moreover, this research's secondary aim is to present a modular-based artificial self-awareness to motivate a further extension of artificial self-awareness in future work.

Research efforts attempted to create intelligent [12], autonomous [26], or human-level robots [11]. Artificial self-awareness can be an essential block that supports advancing robotic tasks to make a robot adaptable and flexible within its environment [26] [13] [31].

The objectives of the artificial self-awareness study are as follow:

- Understand self-awareness level-1 requirements and devise it artificially for a robot. Devise multisensory data to enable a robot to associate its movements with its physical body. Initiate a pilot model architecture of artificial self-awareness that enables the sense of self, using a deep neural network and simulated data. Develop an extended model architecture of artificial self-awareness that enables the sense of self-using real robotic data. Demonstrate the artificial self-awareness level-1 acquisition with a physical robot.

- Reconstruct a scalable and interpretable level-1 of artificial self-awareness architecture. Building architecture processes multimodal sensory data. Building an architecture that can fuse the vision and position data into a low-level data representation. Construct an output state of integrated vision and position vector to be processed by the next level of artificial self-awareness (level-2).

- Considering an advanced level of self-awareness such as level-2 gives a robot the ability to go beyond getting a sense of self by allowing it to utilise its higher self-confidence and integrate it to control its task. Devise level-2 of artificial self-awareness to adapt self over time and enhance a robotic task. Applied the framework of different environments with the confounding case and demonstrate that level-2 substantially improve robot self-recognition (higher self-certainty).

1.6 Research Hypotheses and Questions

This thesis research hypotheses and research questions of the experiments undertaken to devise and develop artificial self-awareness in a robot are as follow:

Level-1 Pilot Model: Hypothesis

- The sense of self can be enabled in a robot using a Deep Neural Network (DNN) model architecture and the association of simulated multimodal data.

Level-1 Pilot Model: Research Questions

- RQ1: Is it possible to initiate a sense of self in a robot by processing simulated visual and internal velocity feedback?
- RQ2: Is it possible to associate simulated visual and internal velocity using DNN to enable a robot to acquire a sense of self?

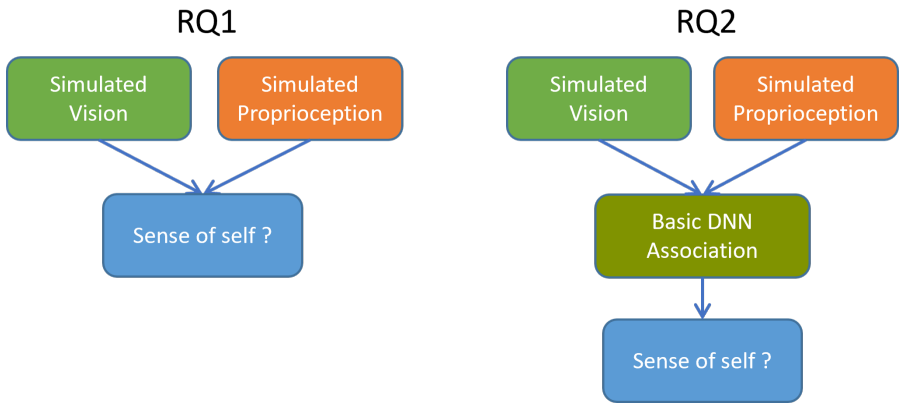

Figure 1.2: The research questions of the Level-1 Pilot Model of artificial self-awareness.

Level-1 Extended Model: Hypothesis

- Level-1 for artificial self-awareness in the robot increases its self-certainty in an unseen environment.

Level-1 Extended Model: Research Question

- RQ1: Does developing the sense of self allow a robot to know itself?
- RQ2: Is the robot able to associate the observed and the felt movements in different environments with high accuracy?

Artificial Self-Awareness Level-1: Hypothesis

- A robot's self-awareness sense of self can be enabled using an unsupervised learning method to associate and reconstruct the fused real multimodal data.

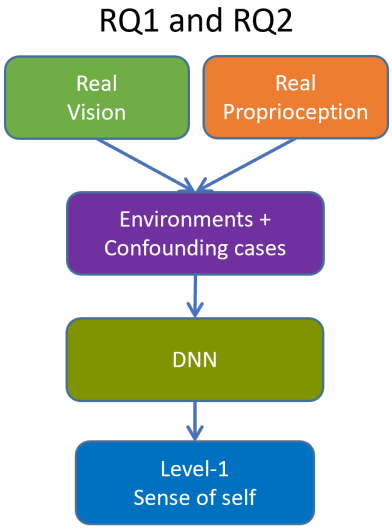

Figure 1.3: The research questions of the Level-1 Extended Model of artificial self-awareness.

Artificial Self-Awareness Level-1: Research Questions

- RQ1: Can a robot initiate a sense of self by associating its multimodal data using an unsupervised learning method?
- RQ2: Can a robot's sense of self enable it to interpret what it sees and sense by reconstructing its fused multimodal features?

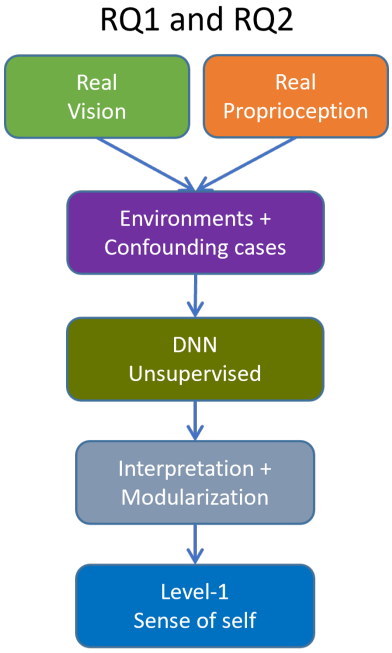

Figure 1.4: The research questions of the Artificial Self-Awareness Level-1.

Artificial Self-Awareness Level-2: Hypothesis

- A temporal perception of the robot's dynamic movement increases the self-certainty of the robot in a different environment if the predicted reconstructed signals for vision and proprioception are statistically significant and classification accuracy is above level-1 for the same dataset.

Artificial Self-Awareness Level-2: Research Question

- RQ1: Is a robot able to relate its arm movements with its body?
- RQ2: Is a robot able to build an inter-modal link between what is seen and felt?

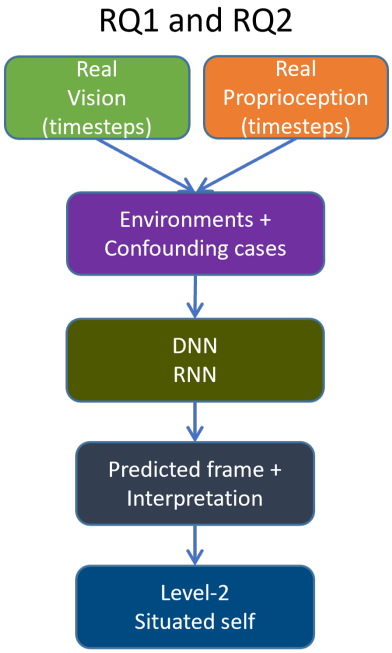

Figure 1.5: The research questions of the Artificial Self-Awareness Level-2.

1.7 Development Process

A modular approach is implemented as a development process for this research of artificial self-awareness, where modularity is based on the levels of artificial self-awareness, which is suitable for verifying incremental increases in a robot's self-awareness acquisition based on discrete models.

Self-awareness is presented by levels where each level shows an advanced competence of self-awareness acquisition, this motivates to select a modular approach to independently implement each implicit level of the artificial self-awareness. The modularity approach's advantage allows the implementation to separate into different independent modules [32], where each

model concentrates on solving a different problem. Also, it motivates easier troubleshooting of the modules in any development stage, allowing easy integration and upgrade as they are decoupled. In addition, the successor model output is fed to the next predecessor model for further processing execution. However, the disadvantage of the modular approach is that the errors associated with the output of a model will disrupt and affect the result of the next models.

The implementation explained in Fig 1.6 shows that the level-1 artificial self-awareness module outputs a sense of self features. These output features are processed as input by the level-2 module, whose output different features express the situated self.

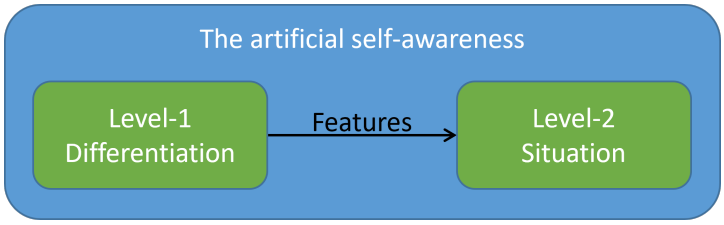

Figure 1.6: Level-1 of artificial self-awareness (differentiation) outputs features that get further processing by level-2 (situation).

1.8 The Proposed Self-aware Research Approach

This thesis approach focuses on defining the self before interacting with the environment, i.e. agents must accumulate self-knowledge to familiarise themselves before dealing with the environment. Also, this thesis proposes a modular approach to devise artificial self-awareness in a robot, in association with the inspiration from levels of human self-awareness by Rochat's research [1].

1.9 Research Contribution / The Impact of This Research

The thesis scientific contribution that participates in advancing the artificial self-awareness and the overall robotics domain:

- Proposed an experimental methodology that allows testing whether a neural network and a robot can learn (supervised) the sense of self by recognising itself from the environment through differentiation.
- A real robot demonstration shows the robot being able to differentiate itself.
- Proposed an experimental methodology that allows testing whether a neural network and a robot can learn (unsupervised) the temporal self through situations.

- Compiled a dataset comprising synchronised RGB images (egocentric view) and proprioception sensor readings of a dual-arm robot which is used to test this thesis hypothesis. This dataset can be utilised for other multimodal integration studies, which can be accessed at https://doi.org/10.5281/zenodo.7539147.

- Proposed different environment and confounding cases framework used to analyse the self-awareness levels acquisition.

1.9.1 Robotic Field

- Introduce a method focusing on letting a robot learn about itself before advancing and heading to deal with the environment.

- Introduce a method to motivate self-acquisition based on a modular approach that shows a higher competency of artificial self-awareness at advanced levels and is easy to scale up with the future integration of more artificial self-awareness levels.

- Propose using generative models to track and validate the features associated with the defined levels of artificial self-awareness for easy troubleshooting of any issues associated with the models.

- Present an experiment showing deep neural networks' ability to fuse different data modalities to achieve self-knowledge for a robot.

1.9.2 Robot Demonstration

- Level-1 self/environment recognition: https://youtu.be/woZUa2QWJxw

1.9.3 List of Publications

- AlQallaf, A. and Aragon-Camarasa, G. (2020) A Pilot Investigation into Robotic Self-Awareness. ICRA Workshop: Brain-PIL 2020: New advances in brain-inspired perception, interaction and learning, at 2020 International Conference on Robotics and Automation (ICRA), 31 May - 31 Aug 2020.

- AlQallaf and G. Aragon-Camarasa, "Enabling the Sense of Self in a Dual-Arm Robot," in 2021 IEEE International Conference on Development and Learning (ICDL), Aug. 2021, pp. 1–7 [33].

- Poster titled "Enabling the Sense of Self in a Dual-Arm Robot" by Ali AlQallaf and Gerardo Aragon-Camarasa:

  https://icdl-2021.org/program/poster/ICDL21_32_ALQALLAF_poster.pdf.

1.10 The Thesis Structure

The thesis is organised into six main chapters, as follows: **Chapter 2** presents the background required for this thesis and the prior work of self-awareness in humans and robots. Also, it reviews the research studies gap and discusses the ways to bridge that gap work. After that, the accomplishments of the experiment work of this thesis are presented in the following three chapters. The first level-1 sense of self-acquisition achievements by a pilot and extended models is presented in **Chapter 3**. The second level-1 sense of self robust acquisition results achieved using supervised and unsupervised methods are presented in **Chapter 4**. In **Chapter 5**, the level-2 situated self obtained over time provides a robot with higher self-certainty. The final Chapter **Chapter 6** reviews the conclusion of this thesis and the future work that could be carried out.

Chapter 2

Literature Review

This chapter provides a comprehensive review of earlier reported theories, methodologies and applications relevant to the proposed research. Defining and developing a novel artificial self-awareness framework and integrating it into a robotic system requires a wide range of research in different disciplines, such as hardware and software architectures, robotics skills, and cognitive abilities. This chapter has three main themes: background, artificial self-awareness literature review, and literature discussion. The background presents the related technologies and knowledge used in this study's development. The artificial self-awareness literature and classification contains studies relevant to self-aware systems and their classification from a different perspective. The literature discussion presents a critique, limitation, and study advancement.

This chapter's structure is as follows: In section 2.1, the background section presents the necessary knowledge to understand this thesis's methods and implementations. Section 2.2 reviews and maps each level of self-awareness from an artificial perspective. Section 2.3 classifies the relevant papers within the same scope by their aims, study approach, and technology. Finally, Section 2.4 discusses the limitation of current self-aware systems and the methods to advance the robotic artificial self-awareness field.

2.1 Background

2.1.1 Robotic Systems

A robotic system comprises different components formed into certain shapes that can be programmed to sense the environment, decide, and execute various tasks [34] [35]. The robotic system components composed can be electronics or different soft materials [36]. The robotic system's shapes differ based on its target. For example, there is a robot that represents an arm used for pick-and-place purposes, some robots cloning an animal form that is used for education, rescue missions or as a toy, and a humanoid robot that is used for research and other different purposes. There are different types of robotic systems for industrial and consumer, each has

different characteristics. The industrial robot is usually bulky and used for mass production; on the other hand, the consumer robot is meant to interact with humans and is, therefore, lighter and "safe". A robot can be deployed in different places based on its type and abilities; it can be at home to serve a family or at a factory with mass streamlined production. Robots exist to automate tasks; unlike humans, they can work on a task without stopping, but they are not intelligent. Robots use different integrated internal or external sensors to collect different data from their surrounding, utilised to perform a action [9] [34].

This research used Baxter, an industrial robot manufactured by Rethink Robotics. Baxter is a complete integrated system that consists of hardware, software, controls, user interface, safety, and sensors. Baxter was designed to work for manufacturing applications and fits in a streamlined supply chain to increase task productivity or work on any repetitive tasks. Baxter is a dual-arm robot, as shown in Fig 2.1. It has seven degree-of-freedom arms that provide dexterity and range (#1), an external camera is attached(#2) above its head(#3), and its body sets on a mobility base for manual positioning(#4).

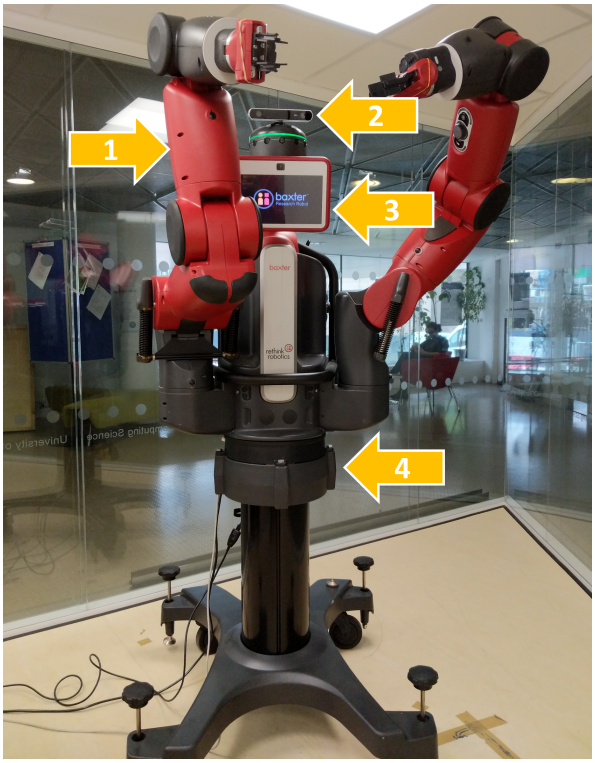

Figure 2.1: Baxter the robot in University of Glasgow, School of Computing Science, CVAS Baxter lab.

2.1.2 Robot Operating System

The control of the Robot is performed using an open source software environment known as the Robotic Operating System (ROS) [37]. It provides tools and libraries to help roboticists build

applications for their robots. ROS is designed based on a graph architecture that represents nodes, the nodes process and exchange messages using topics that follow publish/subscribe protocol paradigm, which allows sharing information by publishing on topics and reading other node information by subscribing to their topics. ROS is language-independent and supports different programming languages such as C++, Python, Lisp, Java, and Lua.

Within the ROS workspace, nodes are imported and defined to manage the proposed research in this thesis. For example, to collect the data for self-awareness models or process data during the demonstration experiment in Fig 2.2, the Data collector node is created that subscribes to read the published messages of Sensor and Camera nodes.

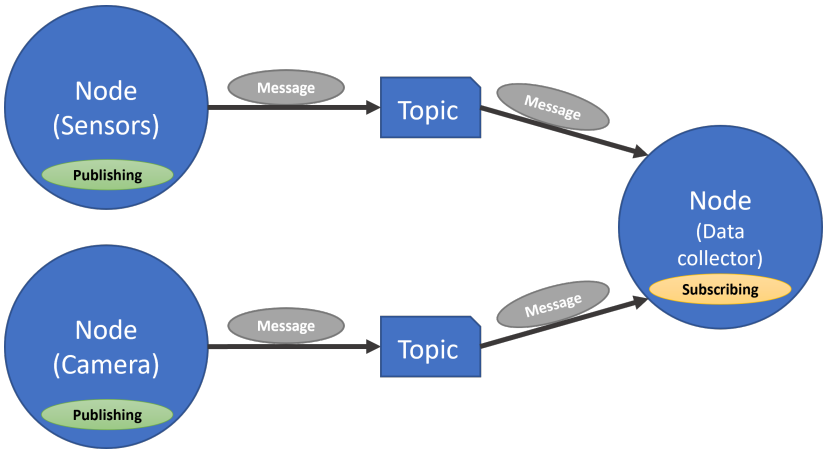

Figure 2.2: ROS paradigm implementation for the data collection of the proposed research.

ROS provides the tools to communicate and control Baxter by integrating Baxter's Software Development Kit (SDK) provided by Rethink Robotics. Baxter's SDK includes all required libraries and tools to interface, operate, and control Baxter.

2.1.3 Robot Simulation

In addition to the libraries and tools that control Baxter reviewed in the previous subsection, this research uses software technologies consisting of different simulator technology. The simulators used are Gazebo [38] by Open Robotics and MoveIt platform [39] by PickNik Robotics integrated with Rviz [38] by Open Robotics.

Gazebo

The Gazebo simulator is a robotic development environment in a 3D world with essential tools to fast design and run proof of concept solutions. Also, Gazebo allows the modelling of different objects and environments to be added within its 3D space. Baxter modelled in the Gazebo 3D space using the Unified Robot Description Format (URDF) provided by the Rethink Robotics

SDK, which consists of Baxter's kinematics characteristics, dynamics, and sensors. The Baxter simulator in Gazebo is fully accessible by ROS controlling and operations.

In this research, the Gazebo simulator is used as the first step to help verify the potential of an artificial self-awareness approach with less complexity and uncertainty in the real environment. This is helpful to demonstrate the concept and to get familiar with issues that might affect the approach when it is implemented in a real environment.

Rviz And MoveIt

Rviz is a primary 3D tool for visualising the ROS framework. Rviz is a powerful tool for debugging robot applications, showing ROS sensor data and status information, and helping users to verify the intended task easily. Using Rviz, Baxter state information from either a simulated Baxter by Gazebo or physical Baxter can be represented, and it can reflect current states of Baxter's and other different connected sensors.

MoveIt is a complete package for robotic motion planning. MoveIt calculates the robot motion plan and can preview it before motion execution. Also, MoveIt provides different planning algorithms for diverse planning solutions for robots. Moreover, it provides collision-checking capabilities during planning motion solutions.

MoveIt in Rviz enables visualising Baxter's kinematics and provides an easy motion planner user interface allowing easy ways to interact with and control the robot. In Fig 2.3 Baxter simulated model is loaded and visualised in Rviz, showing markers on Baxter's right and left end-effectors, allowing a user to interactively plan and move the simulated and physical Baxter robot. Different data points were selected while capturing the dataset for the artificial self-awareness research. Their planning path was verified by positioning the end-effector markers into different random positions.

2.1.4 Robot Sensors

Most robots are equipped with different sensors that help work on different tasks. Therefore, Baxter is built with sensing technologies that can perform useful tasks. Baxter's motors sensor includes torque, position, and force sensing at every joint [40]. Fig 2.4 shows Baxter's joints on one arm that presents seven degree-of-freedom (DoF). The motor sensors represented by joint states are used as proprioception information in this proposed research. The joint states information has three main elements: position, velocity, and efforts for all joints in both Baxter's arms. The topic published by ROS is "/robot/joint_states", which contains nineteen values for each of the three main elements, including the head pan elements values (for more details, see Appendix A).

Furthermore, Baxter has three built-in cameras in its head pan, left arm, and right arm. However, they are not used in this research because the head pan camera field of view angle is

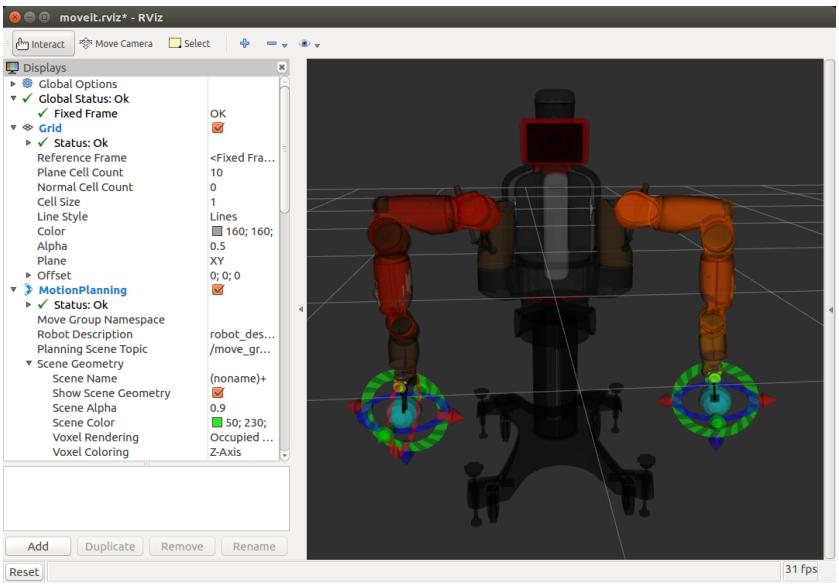

Figure 2.3: Rviz uses MoveIt to control Baxter's right and left end-effectors.

limited, and the cameras in the arms are located in the end-effectors, which cannot view Baxter's arms from an egocentric view. Therefore, an external ZED camera by Stereolabs [41] is used to capture Baxter's arms in the field of view. The ZED camera has dual lenses. It provides a wide-angle view, also, it is flexible to be adjusted as required to fulfil the proposed research requirements. The ZED camera SDK provided by Stereolabs, has a node publishing the camera captured information. Only the right camera is utilised in artificial self-awareness research by processing RGB images. The information captured presents Baxter's arms in and out of the field of view, each in four different environments.

The artificial self-awareness research in this thesis depends on processing two inputs fed by the ZED camera and the robot joints sensors as in Fig 2.5. The scoop of this research considers applications where vision and proprioception senses are needed in a robot. The vision sense can not be replaced by another sensor, such as laser sensing, because the robot arms' visual appearances are important in learning the arms' visual specification. The proprioception information of robot arms is also important, reflecting the joint's position, velocity, and efforts.

2.1.5 Safe Workspace

A robot workspace is a space in which a robot operates. Specifically, it is a set of all possible reachable points surrounding a robot. Moreover, this workspace represents a dangerous space because it is a working area for a robot to perform an intended task; interfering with a robot in its space results in severe consequences to the surrounding humans and environments.

Based on a robot type and supported functionality, different techniques and sensors are equipped for a robot to prevent accidents that might occur. Some techniques are mechanically limiting the range of motions of a robot joint to prevent intersecting with the other joints of a

Figure 2.4: Baxter left arm with labelled joints each produces three elements of proprioception. The full labels of the diagram's annotated arrows are as follows: "S" is Shoulder, "E" is Elbow, and "W" is Wrist.

robot and also utilising robot sensors and control software for collision detection and prevention.

Baxter is supported by different safety and compliance features from its factory, such handle human contact, avoiding accidental contact, and handling physical contact between workers and the robot [17]. That features allow Baxter to work collaboratively with people in its workspace and allow Baxter to be deployed without traditional safeguards workspace. These features did not interfere with the artificial self-awareness conducted research. However, they provided a safe environment to be in person in Baxter's lab while capturing the dataset and a safe environment to capture simultaneous Baxter's dual arms data.

2.1.6 Robot Capability

Robot capabilities are far removed from the science fiction representation. The current robots cannot adopt unexpected changes that might occur to their tasks, and they lack resiliency. Current robotic manipulators are statically deployed to a task or manually controlled by humans [42]. The robots are evolving into many life aspects, and the environment has uncertainties that might affect robots' functionality and their tasks; this raises a need for a dynamically adaptable robot.

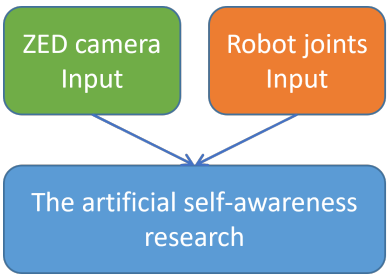

Figure 2.5: The main data sources used in this thesis are the camera and joint inputs.

Humans perform a task with flexibility, and their abilities surpass current robot technology, but they are limited to their physical body potential. In contrast, robots cannot perform as flexibly as humans but perform certain tasks precisely.

Advancing robots to level it up towards the humans level needs to consider different technological aspects. Human self-awareness is an essential ability that allows humans to recognise themselves as an entity in the world; this allows them to control their body's abilities toward their tasks.

2.1.7 Deep Neural Networks

The Deep Neural Network subsection contains general studies that used Deep Neural Networks to support their method and reflect on deep neural networks. Machine learning is a technique where a system learns from its data, not from a set of rules, and the learning is mainly performed by training from a dataset. The deep network is a subset of machine learning methods which are inspired biologically by the human brain's neurons of processing the information and learning from it [43] [44]. It is a deep neural network because the network structure usually contains more than 3 layers of neurons connected and passing information. Because the deep neural network learns from the dataset, the network must consume much data to perform better. Furthermore, the quality of the dataset is also important. PyTorch [45] is used in this study; it is a machine learning framework for deep learning methods.

The deep neural network architectures are widely used in different applications such as classification, vision recognition, and speech recognition [46]. Deep neural network architectures have many types, each with a different approach to solving a problem. This self-awareness research uses a Fully Connected network (FC), a Convolution Neural Network (CNN), Autoencoder (AE), Variational Autoencoder (VAE), and Long Short Memory (LSTM) to facilitate self acquisitions.

Fully Connected Layers

The linear layer or fully connected (FC) is a network consisting of different connected hidden layers of neurons. It is widely used in its simple form for classification decision problems. Its

weights and biases, along with the neurons, help to strengthen the connection between layers and output the final decision prediction. After each layer, a non-linear activation function is used to apply a non-linear transformation, deciding which neuron should activate in the forward direction as it is relevant and important to the data.

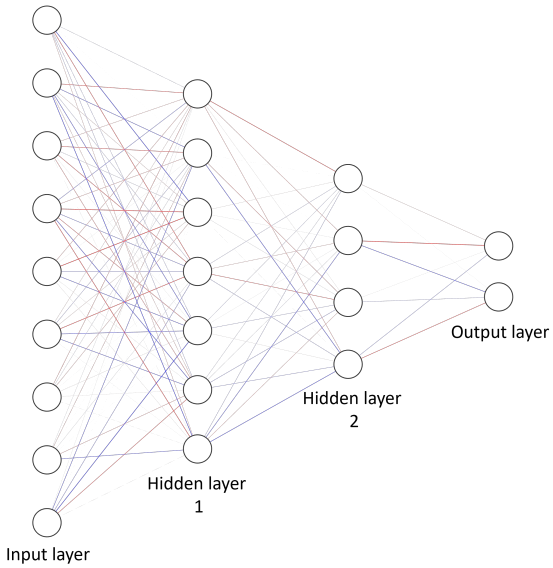

Figure 2.6: An example of a fully connected layers network consists of input and output layers, including two hidden layers.

Different FC networks were designed for classification and feature extraction in different DNN architectures created for level-1 and level-2 self-awareness research architectures that designed in chapter 3, chapter 4, and chapter 5.

Convolution Neural Network

The convolution neural network (CNN) is a network architecture that learns patterns in the image to recognise objects using convolution layers that convolve its image input and feed its result to the next layer. Its first application was recognising different handwritten letters [47].

The basic architecture of CNN consists of two main parts: feature extraction and classification. The feature extraction part is composed of different convolution and pooling layers. The convolution layers are used to extract features from an image. The mathematical operation of convolution is performed on an image with a filter representing a dot product between image pixels and a sliding filter of a particular size. The output from the convolution operation is a feature map that describes distinct image information aspects, followed by a pooling layer that reduces the convolved feature by averaging or maxing it, which helps in reducing the calculation cost.

The classification part resides at the end of the structure. It contains fully connected layers, which take the extracted flattened features as an input and output the prediction class of the processed image.

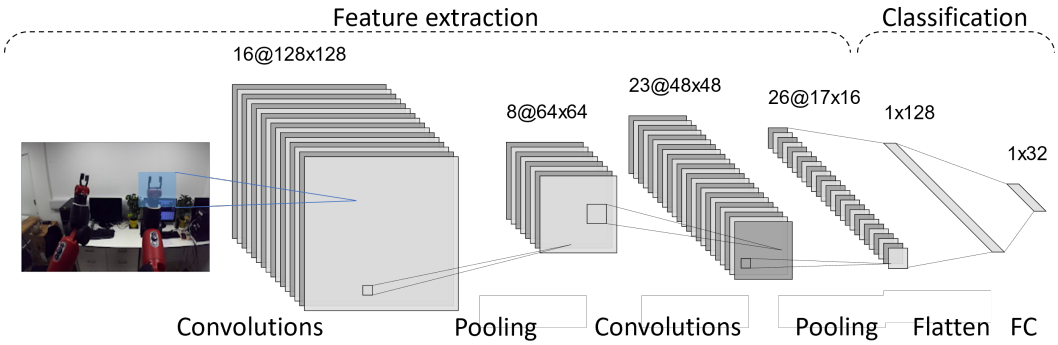

Figure 2.7: An example of a basic CNN shows its parts of convolutions, pooling, and classification functions.

Moreover, a special type of CNN architecture known as Resnet18 [48] is used in this study. The Resnet18 is a CNN architecture used widely for object detection and classification. It contains eighteen convolution blocks used in object detection and feature extraction. Also, it contains skip connections within its deep layers, which aims to avoid the network's vanishing gradient problem where the network's weights are not updated because the loss function shrinks to zero [49] [48]. The reason to select Resnet18 over a plain CNN network is to depend on a state-of-the-art network to extract robot arms features. Also, Resnet has achieved a remarkable image classification result among CNN network architectures such as VGG [48].

The CNN and Resnet18 architectures are used in level-1 of artificial self-awareness in chapter 3, section 3.4 to process the robot vision part of 224x224 pixels as image size input.

Autoencoder

The autoencoder (AE) [50] is a type of unsupervised neural network that learns to reduce the input data to a compressed representation and reconstruct the compressed representation into its close original data form. The Fig 2.8 shows an autoencoder architecture which contains the following:

Encoder: The encoder network contains layers that compress the data input to a reduced form of latent vector. **Latent space**: The latent vector represents the compressed information, also known as a bottleneck. **Decoder**: The decoder network contains layers that uncompress and build the data back to its original state.

The encoder and the decoder usually contain a form of network layers which reverse each other. Its latent vector size is associated with the quality of the decoder output because more information in the bottleneck of a latent vector is better for guiding the decoder to represent its original information. The autoencoder is used in many domains, such as learning to denoise an image, image anomaly detection, or utilising its latent space for a purpose. In this research in chapter 4, section 4.4 and section 4.5 the autoencoder was used to investigate what a robot sees through reconstruction and to process multimodal data into mixed representation.

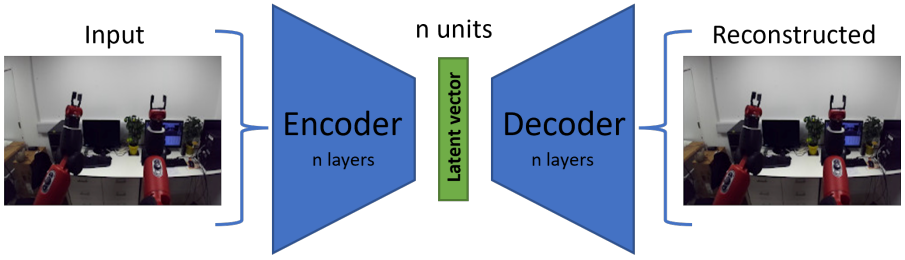

Figure 2.8: An Autoencoder architecture consists of an encoder, latent space, and a decoder. The input image gets encoded into a latent vector and then is reconstructed to its original form.

Variational Autoencoder

The variational autoencoder (VAE) [51] is an unsupervised neural network. Similar to the autoencoder, its main purpose is to compress the data of high dimensionality into a smaller space representation. However, instead of mapping data into a fixed vector, the variational autoencoder is structured to allow maps of its input into multivariate latent distribution.

In Fig 2.9, the VAE network structure contains the mean and the standard deviation vectors of the distribution, allowing a decoder to learn from a variation of the distribution of the same input. Moreover, decoding the data is enabled by sampling from the distribution.

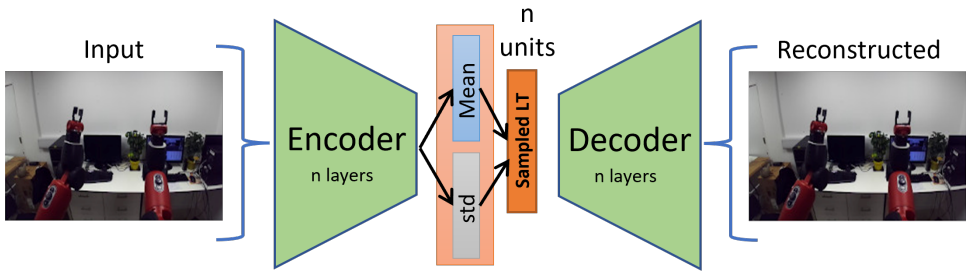

Figure 2.9: A Variational Autoencoder architecture consists of an encoder, mean and standard deviation vectors, and a decoder. The input image gets encoded into distribution and then sampled and decoded to reconstruct its original form.

The Variational autoencoder loss function contains two parts. The first part represents the reconstruction loss calculation similar to the normal autoencoder type. The second part is the Kullback–Leibler divergence (KL) [52], which ensures that the learning distribution is not far removed from the normally distributed Gaussian.

The VAE is used in level-1 of artificial self-awareness research to represent distribution for proprioception and vision input sensors. Also, used to verify both robot input states by reconstructing and visualising what the robot sees and feels in chapter 4, section 4.5.

Recurrent Neural Network

The Recurrent Neural Network (RNN) [53] [54] is a type of artificial neural network used to process time series data or sequential data. This type of network is used mainly for temporal

or ordinal problems, such as natural language processing, video sequences, and historical data problems [55]. The RNN has a concept of "memory", where the information of previous inputs is used to generate the next output of the sequence. The main issue with RNN is the vanishing gradient problem, where the weight becomes too close to zero. Therefore, The Long Short Term Memory (LSTM) architecture is a type of RNN that addresses the vanishing gradient problem by using more gates in its architecture that decide which information of a sequence should retain. The benefit of recurrent neural networks is that to transport information across sequences uniformly. Using the LSTM ensures that it works across long sequences; this presents an attention mechanism to make the network more efficient.

The LSTM in Fig 2.10, gets as an input vector X_t and hidden layer vectors of c_{t-1} and h_{t-1}, its output is represented by h_t, and hidden layer vectors c_t and h_t.

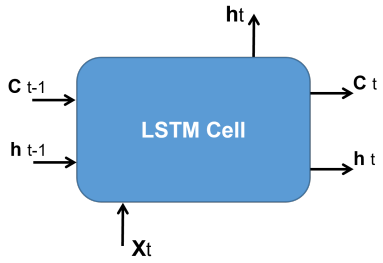

Figure 2.10: The LSTM cell, has hidden layer vector input of c, h, and input vector X.

The artificial self-awareness level-2 implements the LSTM to process the number of input sequences by unfolding the LSTM cells. Each cell processes an input of sequence frame vector and processes previous cell output and hidden states. For example, in Fig 2.11 the unrolled LSTM processing features of 3 sequences, the hidden layer vectors c_{t-1} and h_{t-1} in the first cell only they are initialised by zeros. The unfolded LSTM process each sequence as input x_n and predict the next state of the sequence. Level-2 depends on the final predicted state and is used to be validated the target state.

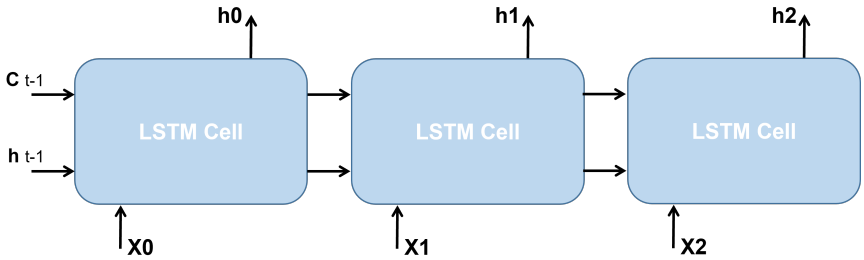

Figure 2.11: The unfolded LSTM cells, the LSTM cells unfolded based on sample input size.

2.1.8 Data Fusion

The information sources used in this study represent two different modalities captured from the vision and proprioception of a robot. This artificial self-awareness depends on both data sources for self-acquisition. Radu et al. emphasise that fusing different modalities is challenging because each data modality might be intrinsically different [56]. Even with two modalities, such as audio represented by waves and video by frames and pixels, it is not easy to associate them as both are non-linear [57]. The sensor used, the sampling rate and the nature of the modalities are variables that affect and make the fusion hard [56].

Two modelling strategies categorise the methods of fusing multimodality in a deep neural network or classical inference, as Radu et al. described in [56]. The first is "Feature Concatenation": where the modalities are concatenated together to form a larger vector of multimodality representation. After that, it gets processed by a fusion model. The second method is "Modality-Specific Architecture": each modality gets processed by a separate network, and the output features are concatenated to be processed by a fusion model network. Their empirical study [56] shows that the Modality-Specific Architecture achieved better accuracies of 5% over the first method of Feature Concatenation on all four tasks. The second method of Modality-Specific Architecture is proposed for video and audio modalities fusion in [57], also used to fuse modalities of human activities using CNN's [58].

This study of artificial self-awareness level-1 uses the Modality-Specific Architecture architecture method to form a shared modality representation with an ability to get an initial sense of self features used for further processing in level-2 of artificial self-awareness to get a higher ability of self acquisition.

2.2 Classify the Artificial Self-Awareness Studies

This section contains a summary and a discussion of the relevant research studies.

2.2.1 The Need for Dynamic Robots

When humans become self-aware, they can recognise themselves in any environment. That is possible because they can distinguish their bodies as separate entities from the world, allowing them to adapt to different situations and scenarios. Robots, however, lack this capability because they are limited to fixed configurations engineered to work in constrained environments. Researchers have theorised [31], [25], [59] that an adaptable robot can increase its productivity and that a self-aware robot can increase task efficiency in different settings and environments.

Despite robots being used in many aspects of our lives, such as serving drinks to people [60], helping rehabilitate and support Alzheimer's Disease [61], and helping in the agricultural field [31], robots work in different environments with fixed setups and configurations. In a case such

as agriculture, where humans and robots are collaborating on a task, Juan et al. [31] expressed that robot adaptability increases its productivity, and the existence of the robot self-awareness will increase robot behaviour efficiency. Also, Kwiatkowski and Lipson [25] consider artificial self-awareness a path to an adaptable and resilient system. Carme [8] believes in a future need for dynamically adaptable robots to fit different tasks and environments. Chatila et al. [59] have stated that robots need to be capable of understanding their actions and adaptable to their environment.

2.2.2 Robots Environment

The real world has different environments associated with uncertainty, and the previous sections emphasised the need for resilient robots. However, Agostini et al. [14] argued that robots could not accommodate all human environments, and hard-coding all possible situations is challenging. To mitigate this, they used AI technique for planner and learning. Moreover, researchers [3], [21] have proposed learning an awareness model inspired by the free-energy principle [62] in robotics, which states that the interactions with the environment are aimed at reducing the internal entropy (i.e. maximising the robot's self-certainty) of an agent. For example, [3], [21] has shown that a robot or its environment might change, and the robot's capability to adapt to different environments is predicated on the assumption that a robot learns continuously using an active inference model. They thus enabled a robot to adjust its control to the task at hand by minimising the distance between the robot's hand and the target object [21] or where the robot's hand is to its internal belief [3]. However, the authors constrained the robot to have reduced visual perception capabilities to simplify the inference task, relying on an observed action within an uncluttered, simple operating environment. Finn and Levine [63] pointed out that integrating a "learned predictive model" with "model predictive control" allows the robot to manipulate unseen objects during training in different environments.

2.2.3 Self-Awareness Dataset

Studies discuss the importance of the dataset for deep learning and the unavailability issue, as Pat et al. [64] linked that the Deep Neural Network prediction performance depends on a dataset and the quality of how it reflects the real data. Some datasets are unavailable or difficult to obtain. Therefore, Konstantinos et al. [65] suggested solving the data unavailability by a simulated dataset of objects and different environments. Brenden et al. [10] have a study focused on the difference in the way humans and machines learned and argued that the present algorithms need much learning to perform as expected while a human can learn from one or few examples. They describe a Bayesian program learning computational model that performs better than current deep learning models in generating a handwritten character. Currently, there are no potential online/benchmark datasets for the proposed artificial self-awareness study. Thus, for this research

study, custom datasets are assembled using real and simulated robots of Baxter and different environments. Due to the importance of the dataset in research and because of availability of a dataset is a problem for the community, the captured dataset for this artificial self-awareness research is published in the Zenodo platform. The artificial self-awareness dataset contributed to the community as open access and can be accessed using the following DOI: 10.5281/zenodo.75391.

2.2.4 Multimodal Sensory

The cognitive studies inspire most approaches of the current self-aware literature. Riva [66] mentioned that integrating different sensory information can guide the location of self in the space. Also, Rochat [4] stated the importance of the modalities to self and the modalities conjunction, which helps form self to be an independent entity in this world. Hoffmann et al. [30] mentioned that for a robot to work on a goal-directed action, it needs knowledge about its physical self and its multimodal sensory.

2.2.5 Self-Aware Model

Hefner et al. [13] have reviewed biological studies and robotics studies related to self. The authors have concluded that self-exploration of behaviours, body representations and sensory-motor simulations, and predictive processes are the three components that could represent self-agency in a robotic system. They have also suggested that self-agency can be measured as the prediction error. The latter is similar to what Amos et al. [24] have demonstrated, where they based self-awareness on predictive control models to allow a robot to create a link between itself and the environment. Similarly, Haber et al. [67] have developed an intrinsically motivated agent using world-model predictions via a supervised learning strategy to model agent awareness to generate different behaviours in complex environments. Similar to this work, Lanillos et al. [20] have used a hierarchical Bayesian computational model to define the self in a robot and argued that understanding sensory mapping changes is core to self-perception. Also, Gold and Scassellati [2] have observed that motor actions and visual motion using probabilistic reasoning allow a robot to self-recognise in front of a mirror and test for self-recognition of the robot self.

Recent research efforts enable robots to develop several forms of body representation that present the self robots. For instance, Gold and Scassellati [2] used a mirror test for self-recognition for their self-model by observing motor actions and motion. The motor motions are processed to decide where each model reflects a situation of either "self", "inanimate", or "inanimate other". Robot Nico of 6 DOF of one arm is used in this study. The camera in Nico's right eye generates 320×240 images from a wide angle. The environment is uncluttered as it gets filtered through a background subtraction filter, leaving only the active objects included within the scene. Their method used three probabilistic models to categorise each part of the

filtered objects based on their moving state, not moving or noise. A mirror test is used for self-recognition, inspired by cognitive studies of human self-awareness [68] [1]. Nico's full body was not fully reflected in the mirror, especially parts that do not move. The result for the three models using four minutes of learning shows that the probabilities changed from 0.5 to 0.020 for inanimate, 0.73 for self, and 0.11 for others. Their algorithm depends on kinesthetic-visual matching, which does not represent a modality fusion, which is fundamental in self-recognition [13]. They use an implicit way to learn self-awareness, while we explicitly learn self-awareness. They get the sense of self but mix levels 1 and 2 as the robot depends on movements.

The authors' next extension study in [27], they experiment using the self-model with an adversary human imitating the robot motion. The idea is to ensure that the result of the robot's reflected self represents Nico's recognition and is not just a match of the simultaneity of motion. The adversary result shows the same probability range of 0.7, representing the self, but this result fluctuates with the time progressing within the range of 0.7-0.2. Comparing the adversary experiment result with Nico self mirror result shows that Nico was uncertain of self as it fluctuated with the adversary case. It also demonstrates that Nico is certain about itself when mirrored and has the self-recognition ability. From this thesis study perspective, the adversary experiment shows that their Nico robot is limited and highly tight with the motion only and not its appearance. This thesis study fuses the robot's motion and visual appearance, which gives a robot high self-certainty and not fooled by other motion that imitates it.

Sturm et al. [69] aim to learn the robot's body schema using self-perception. Their approach used active exploration to map motor commands to body pose observation. The arm learned using visual self-observation only using markers on each motor and an external monocular camera. A Bayesian network is used to identify the arm body structure and its kinematics chains. They demonstrate the approach in real and simulated environments. The results show that the arm can lean to present kinematic structures and adaption awareness to changes or online failure. They argue that learning the kinematics from proprioception would disconnect the motor command action; thus, they depended on the motor command, which can drive the proprioception.

According to Hart and Scassellati [22], self-identification algorithms are the first step toward a complete representation of the robotic self. They focused on mapping Nico's arm kinematics and its vision system. Comparing their approach with other relevant research, they state the use of fewer data sets to train their model. Also, they argue that some approaches attempt to fuse the robot's sensory input. However, in their approach they simply mapped the two inputs of stereo vision and kinematics, making engineering conventional for robotics implementation. When operating at its best, the system can identify the end-effector's position in the visual field to an accuracy of 2.93 pixels (SD=3.83) with respect to the camera view. It must be noted that the authors did not include proprioceptive information to substantiate the ecological self of the robot. However, it can perceive itself because of engineered parametric pairing between kinematics and vision using a motion tracker and vision tag. The Implementation of the ecological self depends

on external entities within the environment. This thesis does not use tags or trackers to define the self in a robot; sensing joint positions visually and proprioceptively should therefore provide a full representation of the ecological self.

Nagai et al. [23] focused on representing self from others by learning to enhance the vision system, inspired by infants' vision development incapable of distinguishing self from others in the first stage of development [70]. Thus, they propose a computational system to learn from the basic interactions with others to develop the vision capability. The authors argue that immature perception helps to detect correspondence between self and others, which is also useful in robots imitating others. They added that the importance of defining self allows a robot to perform contingent interaction and turn-taking. Their method to define self is by maintaining the immature vision with motor command association to distinguish the self from others using optical flow and form clusters to associate them using Hebbian rules. The experiment used an infant-like humanoid robot (M3-Neony) with embedded two eyes cameras and 6 DoFs in its dual arm. The experiment scene is simple and uncluttered and has a caregiver issuing an interaction by Vertical and horizontal patterns gesturing in front of the robot. The result shows that clusters of the robot "self" and others are differentiated in a later observation stage by mapping and maintaining the motor commands association.

The environment of the robot was static and uncluttered. However, the cluttered environment is important to recognise a robot in an unprepared environment, helping recognise a robot in the natural world. Moreover, the system was not fusing the sensors and not using proprioception which is related to what robots feel, not command actions. In this thesis, data fusion of proprioception and vision senses are considered and motivated by cognitive research, suggesting that the robot self is developed by associating multimodal information. Moreover, this thesis on artificial self-awareness includes different environments with clutters, and people passing in the lab's background are included with the study data. Furthermore, the robot is trained to know itself before interacting with the environment or external humans.

Stoytchev [71] was inspired by the efferent and afferent stimuli event proposed by [72], which states that the temporal contingency between the efferent and afferent allows for achieving a self-detection. Thus, in the author's approach, the efferent is represented by motor command and the afferent is represented by visual stimuli. A discrimination threshold estimated by efferent-afferent delay is used to decide whether self or others. A CRS Plus robot arm was used in the experiment with six colour markers placed on the arm joints, and they were tracked and located by using OpenCV library colour segmentation. The camera used is Sony EVI-D30 mounted externally on a tripod putting the arm in its field of view. Self-observation data was collected by motor babbling, a random motor command of joint movements. The results showed that the arm could distinguish and classify either self or others, even with a similar arm in the scene, because the movements are not correctly correlated with the motor commands of the robot. In addition, the self-image model could detect itself by recognising its reflection on a TV

monitor. The perceptual features are attached to one orientation side of the robot, limiting full orientation handling and arm flexibility. Getting data is time-consuming because it depends on a fixed interval. The robot's internal performance degradation might affect the decision of self or others. Also, the system might not be able to handle a scene with many movements by more than two robots as distinct perceptual features will confused.

Hart and Scassellati in [28] extend their study [22] to include a mirror test inspired by Gallup [73]. The authors believe that the mirror as an instrument can be meaningful for reflecting and reasoning the spatial objects in the space. The prior work of self-knowledge allows a robot to build a relationship between its body and the visual sense. In this study, self-knowledge allows a robot to recognise itself in a mirror by locating its reflected end-effector. The new Perspective model allows a robot to calculate its appearance in the mirror. In the experiment, 50 arm positions were utilised in training and 100 for testing. The authors obtained a difference in position between the End-Effector Model and Perspective Model less than 5 cm. The results were calculated and compared between its mirrored and current end-effector position, achieving a mean of 31.55mm.

Lanillos et al. [20] argued that understanding the body's perception and actions is the way to self-perception. They were inspired by O'Regan and Noë [74] theory of human behaviours sensorimotor contingencies, which discussed that human behaviours support sensorimotor mapping. Thus, they attempted to understand the sensorimotor consequences as promoted by their actions. They focused on the modalities changes during an interaction. Also, they mentioned that understanding sensory mapping changes is the core of self-perception. A hierarchical Bayesian computational model is used to integrate the visual and skin of proprioceptive and tactile senses. One arm of the robot was used to perform self-distinction, self-detection, and interaction with an object on a predefined task within one unclutter environment (table). They obtained an accuracy at the pixel level of 74.87%. This performance increased up to 81.9% when classifying the segmented attended proto-objects. Their robot interacted with the environment to define the self by the theory of cause and effect instead of self-exploration. This might add more complexity and tighten the robot's existence with the environmental object. Moreover, the touch sensor can have different effects if the object reaches and touches the robot, which will reverse the cause and the effect entities. This thesis aims to define the self in a robot before interacting with the environment; this occurred without any environment dependency on objects. Also, this method does not allow a robot to learn forward or inverse kinematics.

Amos et al. [24] agreed that artificial intelligence has advanced to the next level, shaping different creatures, assistants, and autonomous systems. The author classified the physical world into two parts: platform and everything else. Indicated that the platform part is what has been created by us, and its capabilities are known. However, the "everything else" part is what the platform is exposed to with uncertainty. The aim was to build predictive control models that are aware and can control their world. Therefore, interacting with shapes builds a dynamic model

that can adapt to its environment, predict its shapes, and define the object from the environment based on a series of tasks. These tasks build and identify the external object properties into the construct proprioception models using sequence-to-sequence modelling. The experiments were conducted using a simulated and real modular prosthetic limb. The training data was gathered by an interaction such as grasping a limited number of different shapes of objects that represent the world. The awareness models explore and predict the state of the objects in the world, which also reflect the object's effect on the robot arm.

They suggested that the communication that happens between self and world, helping in forming predictive models that are aware and represent the external objects that interacted with them in the external world. While this approach defines the external objects to models by interacting with the external world, this thesis approach defines a sense of self (level-1 of artificial self-awareness) by differentiating it from the external world. Focused on learning about a robot model by familiarising itself visually and proprioceptively. Therefore, aiming to understand itself and utilise it to control the world. The author's approach is limited to the objects they interacted with, making their approach hard to generalise to the whole world of objects.

Haber et al. [67] study was motivated by the strategy of infant learning through different activities and spontaneous behaviours by Twomey and Westermann [75], and Fantz [76]. The main focus was to achieve an agent that learns from its environment. They built an architecture with two models: a World-model where the agent learns the consequences of its action and a Self-model where it helps the agent track error of the world model. They aim to have a self-aware agent that can understand its dynamic and generate different behaviours. The authors argue that most of the current advanced robots cannot function autonomously in complex environments without the human help of pre-configured tasks and rewards. Therefore, the self-aware agent has intrinsically motivated agent architecture using the world-model prediction supervised by the self-model to generate different behaviours to control its environment. Deep reinforcement learning was used for the agent. Also, they used a self-supervised convolution deep neural network to train the world model and the self-model using a realistic 3D virtual world environment. The agent results produce behaviours such as ego-motion prediction, object attention, and object gathering. In this architecture, the agent learned to deal with the environment and extracted knowledge by interacting with the environment. Despite the good outcome of generating different behaviours, the agent still has a missing part: a real self-model that can recognise itself before doing an action.

Lang et al. [77] paper defined self in terms of showing the ability to use forward models such as CNN to predict the sense of agency in robots. This paper represented the sense of agency by processing sensory state data and motor commands. The second experiment, related to object permanence, allowed a robot to detect an object even if it occluded it. They concluded by mentioning some limitations related to their approach, such as not considering different environments; they used self-exploration behaviour presented by babbling for generating image data

labelled with joint configuration data. Their environment is simply uncluttered, which CNN predicts. They do not use both arms of the NAO's robot, and only 4 degrees of freedom were used. They did not use the visual modality in their architecture as an input; their model predicts the visual state based on recorded vision states. Finally, they mentioned that unlocking these would potentially improve the stability and detection rate of the results.

Lanillos and Cheng [78] used a bio-inspired technique of predicting coding by Friston [79] [62] to learn and predict robot body perception. They introduced a predictive model that allowed a robot to learn, infer, and update its body configuration. Their work aimed to build an adaptive robot system that can know its configurations for safe interaction applications. The data collected was generated by body exploration. The body perception was transformed into a latent space distribution representing the arm schema to the real arm distribution with the sensory information and forward learning of the arm sensors modality of tactile, proprioceptive, and visual using local Gaussian process regression. The latent approximation calculates the sensory fusion inference, and the experiment showed how sensor availability participates in body adaptations. The prediction error was calculated based on free-energy minimisation for every sensor. The authors simplified the implementation by restricting the robot from perceiving the sensor signal's gradient, representing a passive static perception. Their extension study [21] addresses the static perception using the non-linear function of locally weighted projection regression instead of the Gaussian process. This change provided a high-dimensional learning scale while the arm was moving, but the prediction resulted in the first method were better.

Kwiatkowski et al. [25] showed that a robot can model itself without prior knowledge of its structure and constructs a self-model that can adapt to mechanical changes that occur to the robot arm. Rochat's [1] human development inspired their research. The authors argue that self-image developments are happening in humans and allow them to have the ability to understand the environment around them. However, robots are working within hardcoded abilities. They aimed to make a self-aware agent aware of its changes to continue its task in case of damage. They used a real robotic arm of four degrees of freedom using pairs of data presenting action-sensation captured by moving the arm of 1000 random trajectories, trained using a deep learning method. Their self-model was tested by letting a robot arm perform tasks of writing and pick-and-place. Also, the self-model was capable of detecting future damage that occurred to the arm. Their work demonstrated that self-modelling is the conduit to adaptable and resilient robotic systems. However, the proposed self-model architecture learns about the robot's internal mechanical structure and cannot distinguish itself as an entity in the environment without being explicitly defined. The basic robot's existence as an entity reflects the first level of self-awareness, and Kwiatkowski's self-model is unaware of the distinction between itself and the environment. The multi-modal representation must be considered to provide the self-aware agent more flexibility to understand itself as an entity, which is an important aspect of self-awareness development.

Sancaktar et al. [3] used the Nao humanoid robot in a real experiment, generating training data set of 3200 samples consisting of elbow and shoulder joint values and the observed arm images from an internal robot camera. Also, an experiment was conducted on simulated data of the same robot. A subset is taken from the dataset and used for validation. The environment in both experiments was uncluttered with a solid background. The robot's perceptual inference was tested by getting its arm posse from the visual information. Moreover, the validation of active inference was evaluated by testing the robot reaching an imaginary arm pose. The results of both real and simulated environments show that the perceptual inference got close to MEAN and STD to the target. The active inference results showed predictions that comply with the internal belief of free-energy actions generated.

The authors have a previous similar study [29], aiming to provide a robot of self/other distinction using active inference and a simple deep neural network that was built based on a probabilistic simple neural network model to provide scalability and interpolation. Their experiment attempted to map the optical flow of the visual field of the end-effector and the joint velocity. In addition, they used a mirror test to validate their method.

The taxonomy of this subsection information has developed into table 2.1, which reviews different comparison points among the discussed papers.

2.2.6 Mirror and Self-Awareness

Cognitive studies highly inspire the studies of robotic self-awareness. Thus, many studies such as [2] [27] [28] [29] included the "mirror test" as a validation technique to show that their method of robotic self acquisition allows a robot to recognise itself using a mirror. Historically, Gallup [73] came in 1970 with the mirror test to validate the self-consciousness of chimpanzees and stated that his experiment showed that the chimpanzees showed signs that they recognised their reflection. However, monkeys were not able to have that capacity. Rochat [1] also used the mirror test with children, showing that they can recognise themselves, showing different behaviours indicating they can recognise themselves.

The mirror will limit the robot to recognise itself within a limited medium, that is, the mirror. Moreover, the mirror test is not functional and can be tricked easily by considering the reflection and defining it statically as self. In addition, the developed architectures in this thesis utilise a deep neural network with 30K data to train a robot. This data involves different complexity, which is hard to get the same amount from a limited mirror area. Furthermore, this thesis study focuses on letting a robot know itself before interacting with the environment, which lets a robot acquire an artificial self and utilise that knowledge to support tasks.

Table 2.1: The literature summary table.

Paper Title	Authors	Define self as	Environment focused / Task interaction	Method/approach	Aim	Technology used	Real or Simulated	Uncluttered/Cluttered Environment	Multimodal	Used markers	Used mirror	External camera	Arm/Robot? Dual arm? DOF
A Bayesian Robot That Distinguishes "Self" from "Other"	Gold and Scassellati	Self-recognition	Mirror focused	Simple dynamic bayesian method - probabilistic reasoning	Self-recognition	Nico	Real - sweeping gestures	Images passed through a background subtraction filter, leaving only objects that had moved since the experiment began and scattered noise.	Kinesthetic-visual matching model - no fusion	No - pixel matching	Yes	Used 320x240 images pulled from the wide-angle CCD camera in Nico's right eye at 30 frames per second	Full robot but one arm used / 1 year old / 6 DOF
Using probabilistic reasoning over time to self-recognize	Gold and Scassellati	Self-recognition	Mirror/Human adversary	Simple dynamic bayesian method - probabilistic reasoning	Self-recognition	Nico	Real - sweeping gestures	Images passed through a background subtraction filter, leaving only objects that had moved since the experiment began and scattered noise.	Kinesthetic-visual matching model - no fusion	No - pixel matching	Yes	Used 320x240 images pulled from the wide-angle CCD camera in Nico's right eye at 30 frames per second	Full robot but one arm used / 1 year old / 6 DOF
Body schema learning for robotic manipulators from visual self-perception	Sturm et al.	Body schema	Self-perception from camera	Self-perception using Bayesian networks	Body schema for kinematics - Learning the kinematic from self-observations	Arm minuplator	Real and simulated arm	Simple unified background	Vision	Yes - in the training - visual markers	No	Yes - external - monocular camera	Arm 6-DOF
A robotic model of the Ecological Self	Hart and Scassellati	Ecological self	Self-observation	Parametric kinematics and vision	Learn a robot its kinematics from observing its end-effector in its view filed	Nico	Real	Simple white BG	Mapping / bridging /paring the kinematics and vision	Reflective markers (ARToolkit) and motion tracker (Vicon MX)	No	No	Full robot but one arm used / 6 DOF / but only 4 used due to the marker limitation
Emergence of mirror neuron system: Immature vision leads to self-other correspondence	Nagai at al.	Self	Human interaction required to enhance the vision and develop the self and other.	Develop from immature vision with motor command association to distinct the self from others using optical flow to form clusters and associate them using hebbian learning.	Learn self from others	Infant-like humanoid robot	Real - based on primal interaction - Vertical and horizontal patterns	One simple static experiment environment - in front of a person.	Mapping vision and motor command	No	No	CMOS USB cameras (640 x 480 pixels) embedded in the eyes.	Dual arms, but arms has limited appearing
Self-detection in robots: a method based on detecting temporal contingencies	Stoytchev	Self-detection	Self-observation of colour markers	Discrimination threshold estimated by efferent-afferent delay is used to decide whether self or others	Classify different stimuli of either self or others	CRS Plus robot	Real- motor babbling	Simple white BG	Mapping vision and motor command	Six colour markers	TV is used	Yes - external Sony EVI-D30, image resolution was set to 640x480, 30 frames per second	Arm limited to three joint were allowed to move

Paper Title	Authors	Define self as	Environment focused / Task interaction	Method/approach	Aim	Technology used	Real or Simulated	Uncluttered/Cluttered Environment	Multimodal	Used markers	Used mirror	External camera	Arm/Robot? Dual arm? DOF
Mirror Perspective-Taking with a Humanoid Robot	Hart and Scassellati	Mirror Self-Recognition	Self-observation in a mirror	Mirror self-recognition be six component each utilising ones provides self-knowledge and model the visual appearance of the robot.	Develop an architecture that allow to pass the mirror test	Nico robot	Real	Simple static	Using previous self-knowledge based spatial relationship between its body and visual sense - no fusion.	Yes	Yes - mounted on movable table whiteboard.	Two cameras egocentric / virtual by mirror	4 DOFs / right arm
Yielding Self-Perception in Robots Through Sensorimotor Contingencies	Lanillos et al.	Self-perception	Yes getting self-perception while interacting with the environment - objects	Hierarchical Bayesian computational mode	Learn self-perception by cause and effect	TOMM robot	Real - interacting cause and effect	Predefined task and unclutter environment	Visual and skin(pro, tactile)	No	No	Single camera See3CAMCU50, 640x480 pixels at 30 frames/s	Only one arm used
Learning Awareness Models	Amos et al.	Awareness models	Yes, interacting with shapes to build a dynamic model that able to adapt its environemnt by predict its shapes.	Predictive control models seq-to-seq	Build a model to aware of its world	MuJoCo HAPTIX and shadow hand	Simulated and real	Simple of limited shapes	Proprioceptive and tactile (Haptics: touch)	No	No	Yes	Two: limb only - Modular Prosthetic Limb and Shadow limb
Learning to Play With Intrinsically-Motivated, Self-Aware Agents	Haber et al.	Self-aware agent	Yes, Exploration	Self-supervised learning strategy CNNs	Build an intrinsically motivated model to structure and be aware of its environment.	Simulated agent	Virtual 3d realistic environment	Uncluttered Virtual world	Vision		No	Yes	Simple virtual agent in a 3d world
A Deep Convolutional Neural Network Model for Sense of Agency and Object Permanence in Robots	Lang et al.	Self-agency - yes	Yes - in respect to the permanence side	CNN - forward model	Self-agency and object permanence	NAO robot	Real - Self-exploration behaviour by babbling	Simple uncluttered blue BG	Sensory state/Motor command	No	No	Robot's forehead camera again at 30 frames per second	One arm / 4 DOF
Adaptive robot body learning and estimation through predictive coding	Lanillos and Cheng	Body learning	Body exploration.	Predictive processing (predictive coding)	Learn and predict robot body perception	TOMM robot	Real - body exploration	Simple uncluttered	Visual, pro, and tactile	No	No	RGB camera mounted on the head of the robot with 640X480	One arm

Paper Title	Authors	Define self as	Environment focused / Task interaction	Method/approach	Aim	Technology used	Real or Simulated	Uncluttered/Cluttered Environment	Multimodal	Used markers	Used mirror	External camera	Arm/Robot? Dual arm? DOF
Active inference with function learning for robot body perception	Lanillos and Cheng	Body perception	-	Learning active awareness model inspired by the free-energy principle.	To actively perceive and predict the robot arm.	Simulated arm	Simulated -arm in matlab simulink	Uncluttered and simple operating environment	Visual and pro	No	No	Mounted monocular camera	Simulated 2 DOF arm
Task-agnostic self-modeling machines	Kwiatkowski and Lipson	Self-modeling	Two tasks	Deep learning to train a self-model	To create adaptable self-aware agent	3d printed arm	Action-sensation pairs by moving through 1000 random trajectories	Simple uncluttered	Action-sensation	No	No	No	Four coupled degrees of freedom
Robot self/other distinction: active inference meets neural networks learning in a mirror	Lanillos et al.	Self/other distinction	Mirror	Using active inference and a simple deep neural network	provide a robot of self/other distinction using active inference and a simple deep neural network	TIAGo robot	Real	Simple uncluttered	Joint angel and optical flow(visual)		yes	Head monocular RGB camera	One arm 7 DOF
End-to-End Pixel-Based Deep Active Inference for Body Perception and Action	Sancaktar et al.	Body perception	Real and simulated in simple env.	Learning an awareness model inspired by the free-energy principle and deep neural convolution decoder.	To allow a robot to perceive its body using deep active inference	Aldebaran NAO humanoid robot	Simulated and real of left arm	Uncluttered, clear BG, simple operating environment, and used portion of subset for evaluation.	Visual and sensation	No	No	NAO's bottom camera	Left arm only

2.3 Discussion

The limitations/gaps of the research studies reviewed in the previous section of 2.2 are listed in this section, as well as the suggested thesis methodology to advance state-of-the-art of artificial self-awareness.

2.3.1 The limitation/gap of the Current Self-Aware systems

- The current self-aware systems depend on the environment by using a mirror [2] [27] [28] [29] or interacting with an object [23], [20], [24]. These dependencies in previous work limit the agent by framing itself as an object within the environment.
- The implementations are mostly self-contained architecture, making it difficult to troubleshoot and scale to achieve a full self-aware system. Except [28] which has six models, each concern with a function within its self-recognition system.
- Almost all of the self-model studies mentioned above use a simple and uncluttered environment. Also, they used a solid colour background.
- The data modalities are not fused in many self-model studies. Mainly they are mapped together using different distributions.
- A simple experimental setup utilises a real or simulated robot with minimum degrees of freedom. Moreover, others used markers [69] [22] [71] [28].

2.3.2 Advancing the State of Art

According to the limitation and gaps of the self-aware system, this thesis provides the following solutions to advance the artificial self-Awareness field.

Focus on self from inside to out

The above robotic agents in the literature have learned to deal with the environment while acquiring self-awareness. However, this thesis argues that a robotic system must have the capability to recognise itself (i.e. be self-aware) before performing actions within an environment, as shown in Fig. 2.12. This allows a robot to understand its capability, which helps deal with a task. The proposed approach in this thesis follows the idea that starts by defining a sense of self (level-1, section 1.2) by differentiating it from the external world. Then, integrate another module that defines a situated self (level-2, section 1.2) by situating it within this world with a robot belief intrinsically from its ego-enteric view. The approach is to build self-awareness models from

inside to outside and defines the internal self before a model interacts with the external environment. The thesis inside to out approach is also supported by Riva [66]. The author stated that the volitional studies agree that the self utilises the body for its needs and goals.

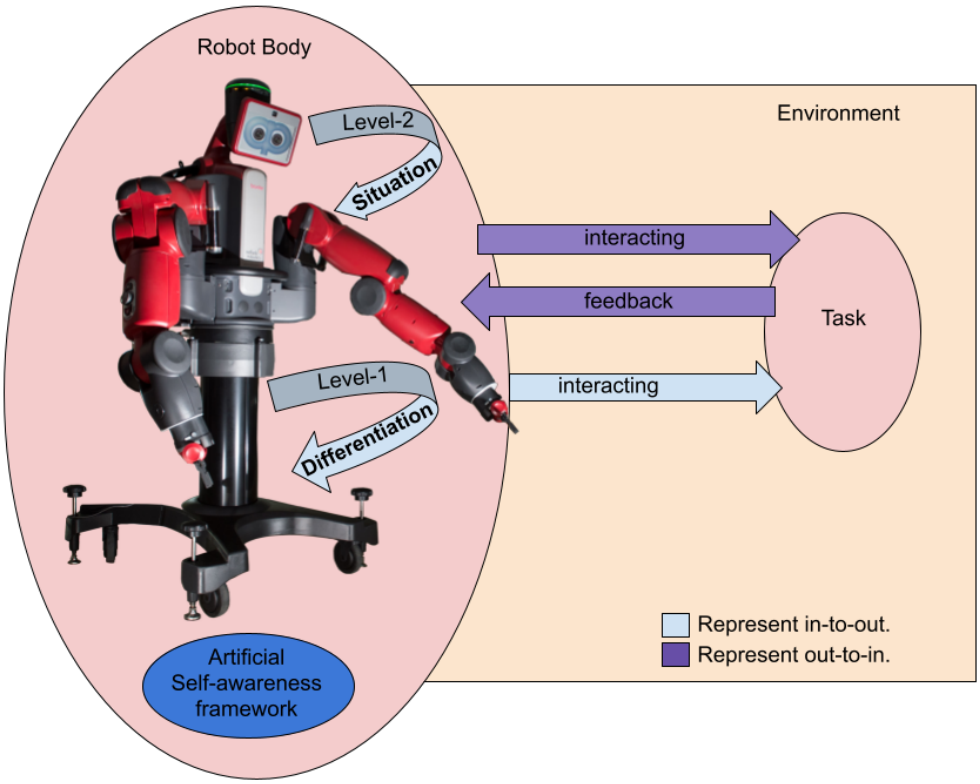

Figure 2.12: A robot starts to learn itself by differentiating and situate itself in this world, then utilise self-knowledge to interact with the environment.

Modular acquisition of the self

The way Rochat's [1] classified the self-awareness development motivates the design of the thesis research implementation, which is focused on a modular approach as explained in section 1.6. Also, self-awareness is presented by levels where each level shows an advanced competence in self-awareness acquisition. The independence between each implementation of the self-awareness levels makes it easy to control the modules and build future levels. In addition, future individual changes do not affect other level modules. Thus, it is easy to scale the artificial self-awareness models as needed.

Acquiring self in different environment

The experimental setup of the artificial self-awareness consists of Baxter, the robot. The 14 degrees of freedom dual-arm of Baxter are used in level-1 and level-2 experiments of artificial self-awareness. Moreover, four environmental scenes were involved in training, validating, and

testing the models. The data was captured into four groups of three subsets with leave-one-out. The reason for leaving one out is to check the models' abilities to generalise over the unseen world. Moreover, the dataset was arranged to represent confounding cases for the robot.

In addition, a fusion network was introduced in level-1 of artificial self-awareness that helps to fuse vision and proprioception data modalities. The fused features are used in level-2 to further increase a robot's self-certainty.

This thesis artificial self-awareness level-1 and level-2 process the dataset with the configured experimental setup complexity, enabling it to generalise and work under unseen test data.

Chapter 3

Enabling the Sense of Self in a Dual-Arm Robot

This chapter investigates the artificial self-awareness level-1 requirement and acquisition process to initiate a robot's sense of self. This requires preparing a dataset of different modalities and preparing a model architecture that is able to associate the dataset modalities. Two experiments are conducted to verify the initiation of a sense of self, each with a different architecture network and different data modalities sources.

The first experiment presents a pilot model to initiate a robot's sense of self by associating simulated multimodal data. This includes the acquiring sense of self details process of capturing the simulated multimodal data, the architecture network design, the experiment, and the results.

The second experiment presents an extended model to initiate and devise a robot with a robot's sense of self by using real multimodal data. This extended model includes the acquiring sense of self details process of capturing the real multimodal data, the advanced level-1 architecture network, the experimental groups formed with leave-one-out and confounding cases, the experimental validation, and the results.

3.1 Introduction

This chapter reviews the first level of artificial self-awareness by underlining the notion that a robot learns to construct a sense of self before interacting and dealing with the environment and objects. The sense of self in this study represents the first self-awareness level "differentiates" based on development levels inspired by Rochat [1]. This chapter proposes to investigate the first level of self-awareness which serves as the building block for enabling a robot to become an adaptable and flexible autonomous machine. For this, the robot must correlate its visual and internal sensing modalities to initiate the sense of self. The level-1 self-awareness is framed as a binary classification task in which a robot can answer whether it can differentiate itself as an entity in an environment with a degree of certainty (i.e. certainty is the accuracy of the

classification prediction).

The first and second experiments are divided into two main sections in this chapter: A pilot model in section 3.3.4 presents the initial model development and experimentation of the sense of self. An extended model in section 3.4 presents the advanced development and experimentation to acquire the sense of self. The two sections of the pilot and extended models are independent, each with a different approach and architecture design to achieving the sense of self.

In this chapter, two hypotheses of this thesis reviewed in section 1.6 are revisited to be assessed throughout the sections:

- The pilot model hypothesis: **The sense of self can be enabled in a robot using a DNN model architecture and the association of simulated multimodal data.**
- The extended model hypothesis: **Level-1 for artificial self-awareness in the robot increases its self-certainty in an unseen environment.**

This chapter presents the technical details of the pilot and the extended models' developed using a deep neural network. The proposed pilot architecture uses a simulated dataset, while the latter uses a real dataset. The pilot model section details an investigation of pilot model architecture comparing two groups and shows the classification result. The extended model section shows the details of the extended architecture by comparing the accuracy between four environmental groups of confounding cases in terms of classification accuracy, mutual information similarity, and saliency map visualisation. The pilot and the extended experiments indicate that a robot can get a sense of self by differentiating itself from the environment using simulated and real datasets.

In Section 3.2, the motivation and objectives are given. In Section 3.3, the pilot model of artificial self-awareness is investigated. In Section 3.4, the extended model of artificial self-awareness level-1 is discussed and reviewed, and the limitation of this chapter's models in Section 3.5, Finally, this chapter is concluded in Section 3.6.

3.2 Motivation and Objectives

Rochat [1] argues that self-awareness is an essential perceptual experience individuals acquire during their early life stages. According to Rochat [7], self-awareness in humans leads to distinguishing themselves from the environment, allowing humans to comprehend themselves and control their capabilities. Also, Kwiatkowski and Lipson [25] consider artificial self-awareness a path to an adaptable and resilient system.

Implementing the first levels of self-awareness artificially in a robot will allow a robot to get a degree of awareness and acquire knowledge about itself. This knowledge lets a robot spot itself within the environment, which can be utilised to support the decisions in tasks.

Rochat's self-awareness development study emphasises the importance of comprising multimodal information to form self-awareness levels [1]. Therefore, level-1 of self-awareness acquisition incorporates proprioception and visual inputs. The main idea is to allow a robot to capture an initial sense of self, using a basic snapshot of its sensory input. The relation of both sensory inputs is important for a robot to form a basic recognition of itself.

This chapter's initiation of a sense of self motivates proposing a pilot model representing the first DNN model to acquire a sense of self. Moreover, this chapter will investigate the pilot model's ability to encode the initial sense of self, using simulated multimodal data of proprioception and visual inputs. Therefore, based on the successful preliminary result of the pilot model, the extended model with advanced architecture was developed to evaluate the sense of self, using a real robot dataset supported by Resnet18 as a vision network.

In this chapter, based on the modular approach designed for this research mentioned in chapter 2 section 1.7, level-1 represents the ground level for the next advanced artificial self-awareness levels, as described in chapter 2, section 1.2. Thus, to advance to level-2 self-awareness of higher recognition capability, obtaining the level-1 represented by the sense of self is mandatory.

In order to achieve level-1 acquisition of artificial self-awareness, the objectives of this chapter are as follows:

- To understand self-awareness level-1 requirements and devise it artificially for a robot.
- To devise multisensory data to enable a robot to associate its movements with its physical body.
- To initiate a pilot model architecture of artificial self-awareness that enables the sense of self, using DNN and simulated data.
- To develop an extended model architecture of artificial self-awareness that enables the sense of self using real robotic data.
- To demonstrate the artificial self-awareness level-1 acquisition with a physical robot.

3.3 Pilot Model: Initial Sense of Self

The proposed pilot module of artificial self-awareness level-1 aims to devise a possible way to initiate the self in a robot. This section reviews a prototype of the first architecture aimed to implement the level-1 self-awareness that focuses on constructing the sense of self by letting a robot differentiate itself within an environment as described in the introduction section and Chapter 2, section 1.2.

The pilot model processed a simulated dataset of vision and proprioception senses. The purpose is to investigate the possibility of the sense of self that can be captured using a deep

neural network and the sensory mixture of vision and perception inputs. There is no defined dataset for artificial self-awareness systems; this pilot model used a simulated dataset of robot and environment scenes to enable the sense of self. The robot's perception of its arms and proprioception associations present the information required for the pilot module to initiate the sense of self within a robot.

3.3.1 Simulation Setup

The simulation dataset for the pilot model is generated using the Robot Operating system (ROS) and Gazebo multi-robot simulator. The simulation setup involves a simulated Baxter robot from Rethink Robotics company, imported into the Gazebo environment. Also, a simulated Kinect camera was positioned at the top of Baxter's head, as shown in figure 3.1. The simulated Kinect camera is used for the pilot model because the ZED camera doesn't have a simulation available. Also, it is supported by ROS and Gazebo environment. Finally, a ROS capturing node is created to synchronously capture and store the data needed from the simulated Baxter and the Kinect camera.

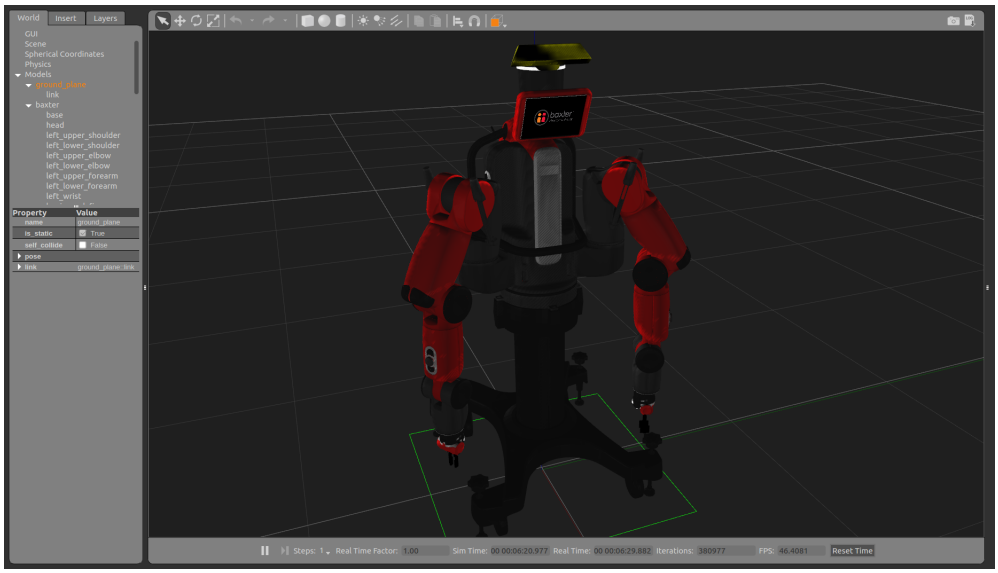

Figure 3.1: Baxter and Kinect in Gazebo simulation setup.

3.3.2 Simulation Dataset

The simulation dataset for the pilot model is generated using Baxter and the Kinect camera in the simulation setup. While moving Baxter's arms among random preplanned points presents Baxter's arms in/out of the field of view defined and controlled using the Rviz simulator, a ROS capturing node is created to capture and store the visual and proprioceptive information. The dataset has been captured and cleaned by removing any image missing a proprioception association and vice versa. Because this is the initial model, a small size of dataset is used. The total

data units in the dataset are 1761 was enough to train and test the pilot model, divided into two sets proportions of approximately 80% and 20% used for training and validation stages, respectively. Moreover, each proportion set contains two subsets labelled "self" and "environment". The training set has 988 data units of "self" and 485 data units of "environment". The validation set has 213 data units of "self" and 75 data units of "environment". The dataset labelling is based on folders that accommodate the captured dataset. For example, the "Environment" folder has all images and proprioceptions associated with random points relevant to Baxter's arms out of the camera field of view, and the "self" folder is relevant to Baxter's arms in the camera field of view.

The "self" datasets present an egocentric view of Baxter's arms with different orientations and backgrounds associated with proprioception states. Similarly, the "environment" datasets present an egocentric view of scenes without a robot arm/arms visible within the simulated environment, and this is also associated with proprioception states. The captured dataset samples are shown in Fig. 3.2. The backgrounds were varied and ranged from uncluttered (simple) to cluttered (complex) environments.

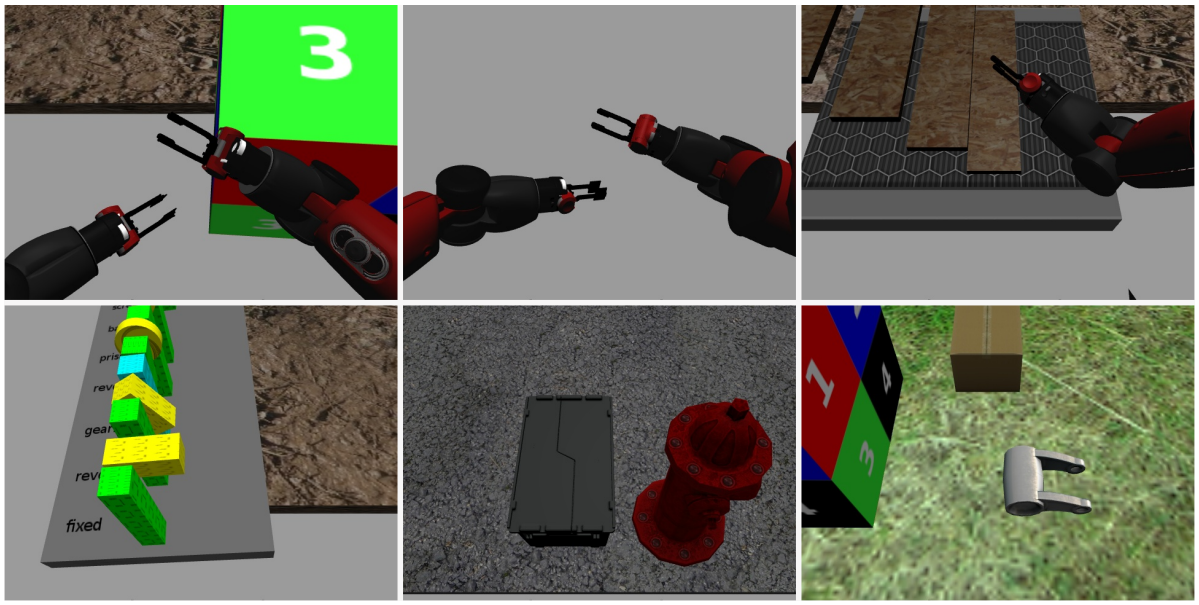

Figure 3.2: The images above represent samples from the dataset captured within Gazebo simulated environment. The upper row images consist of Baxter's arms within the camera field of view representing the "self". The lower row images have no Baxter arm representing the "environment". These images are associated with Baxter's proprioception values representing Baxter's internal states.

The proprioception data captured consists of information elements such as velocity, position, and effort of a robot's arms, where each element has a list of 19 values with grippers left and right fingers state included, representing a total of 57 values stored in a file. The proprioception information expresses Baxter's arms joint states as described in chapter 2, section 2.1.4. In the pilot model architecture experiment, because it represents the first basic experiment, only the

proprioception's velocity element is used to investigate the sense of self acquisition. That is to avoid any complexity affecting the acquisition process by considering all the proprioception elements within the process. The captured images' width and height are 640 and 480 pixels, respectively. Also, transformations are applied to the data input, such as cropping images to enrich the training data and make the trained network more resilient, normalising images to get digested by the DNN network activation functions, and converting the images and the proprioceptive information into tensors, preparing them for DNN training and evaluation.

3.3.3 Pilot Model Architecture

The pilot module employs robot vision and proprioception senses as the primary inputs. A robot's perception of its body and world combined with proprioception information is processed to distinguish itself. Moreover, the visual sense is used to recognise and understand that the robot exists there, associated with proprioceptive information to confirm the robot awareness status. Specifically, The velocity element is only considered from the proprioception elements for the following reasons: to reduce the complexity between the simulated multimodal data association, the robot's velocity within the scene can feed the pilot model with the required internal state parameter, indicating the arm's motion in the scene, and ensure that DNN can enable a sense of self in its basic form. The pilot model comprises a collection of Deep Neural Networks to initiate the artificial self, using the above inputs. This pilot model architecture is created using the PyTorch DNN framework to initiate a self. The model processed the vision and velocity element from the proprioception in the training and validation stages.

The pilot module architecture design was built with a combination of Convolution Neural Networks (CNN) and fully connected layers. The CNN network is used widely for images features extraction and is proven for many tasks. Therefore, the pilot module architecture shown in Fig.3.3 consists of four consecutive CNN layers processing the visual input state of the robot. The output from the last convolution layer block is a tensor size of 3360 units that is used as an input to the fully connected network layer (FC1). Simultaneously, the proprioception represented by the internal velocity state of the robot is processed with a single fully connected network layer (FC2). After that, the FC1 output a tensor size is 1000 units, and the FC2 output tensor size is 19 are get concatenated to present tensor of size 1019 units and is passed to a fully connected layer (FC3), followed by a max-pooling and ReLU (rectified linear unit) activation function to get a more sparse distribution of the output value. The output of FC3 is a tensor of 2 units that predicts self or environment.

3.3.4 Experiment

The experiment of the pilot module is conducted to check the potential of initiating a sense of self by utilising the simulated dataset and DNN architecture designed in the previous section.

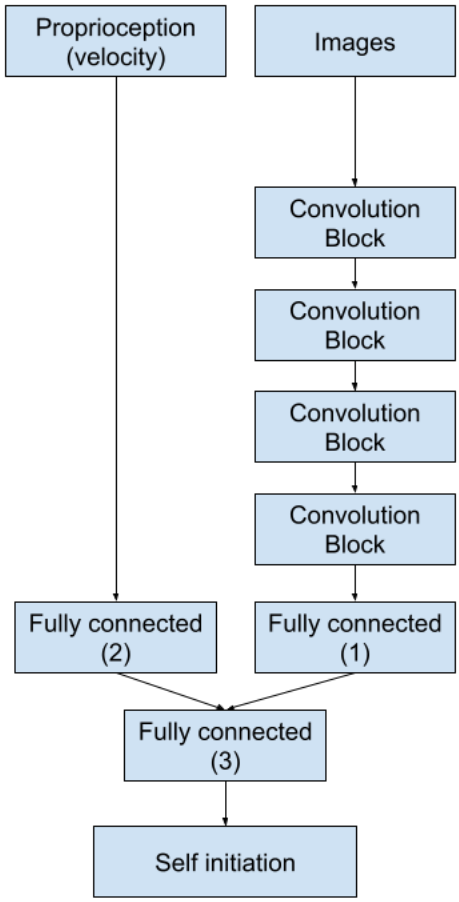

Figure 3.3: The pilot artificial self-awareness module architecture combines vision and proprioception (Velocity) of robot sensors input to predict self or environment. As shown above, the model process the data input through four convolution layers and a linear layer, respectively. Concatenate the output features from both Fully connected (1) and (2) networks and then pass them to fully connected layers (3) to carry out a prediction.

For level-1 - differentiation, the initial basic architecture was constructed to investigate a possible way to the sense of self. Therefore, a simple multimodal neural network and a binary classification task are designed to allow us to understand how a sense of self can be elicited within a neural network and ultimately using the simulated multimodal data.

This experiment answered the following research question:

- **RQ1**: Is it possible to initiate a sense of self in a robot by processing simulated visual and internal velocity feedback?
- **RQ2**: Is it possible to associate simulated visual and internal velocity using DNN to enable a robot to acquire a sense of self?

The prediction output depends mainly on two signals: the presence of the robot arm within the scene and the sense of the movement. The predicted output signal falls into either self or the environment. The robot arm and the proprioception (velocity) establish an association that allows a robot to sense itself.

Training

This pilot module architecture is trained using the training group of the simulation data. The pilot module architecture shows that the training trend Fig. 3.4 of the average loss over 32 epochs is decreasing, indicating that the pilot architecture can capture an association and learns from simulated data using DNN.

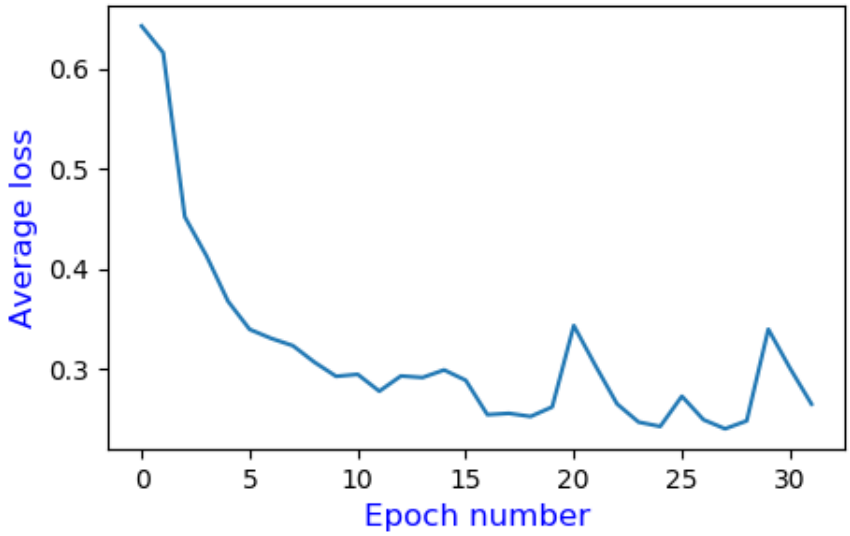

Figure 3.4: This chart shows the training loss of pilot module architecture trained over 32 epochs, where the trend of the average loss starts to settle between approximately 0.25 and 0.35 averages. The loss average trend is decreased gradually from the first epoch to the last reaching 0.25.

Validation

The validation dataset is processed after the pilot model is trained, and the model achieved an accuracy of 93.05% derived by eq. 3.1, where TP is the true positive value, TN is the true negative value, and T is the total units of the dataset.

$$\frac{(TP+TN)}{T} \times 100 \tag{3.1}$$

The validation dataset is visualised in Fig. 3.5 using a Confusion matrix across the two prediction categories of self and environment to assess the model accuracy result. As mentioned in the above "Simulation Dataset section", the validation has two subsets: the "self" group has 213 data units, representing 73.96%, and the "environment" group has 75 data units, representing 26.04%. The percentage of the correct prediction for the self group is 92.49%, and for the environment group is 94.6%.

The false-negative (predicted as the environment but true self) represents 16 units, that is 7.51% of the self group, and the false-positive (predicted as self but it is true environment)

represents only 4 units, that is only a 5.4% of the environment group. The false-negative and false-positive percentage of the total validation dataset represents the miss-classified classes of 6.95%, which shows the pilot model architecture achieved a high classification accuracy in distinguishing self and environment classes using the validation data; this implies the pilot model is able to acquire the initial sense of self as the model classified the major of the evaluation dataset correctly.

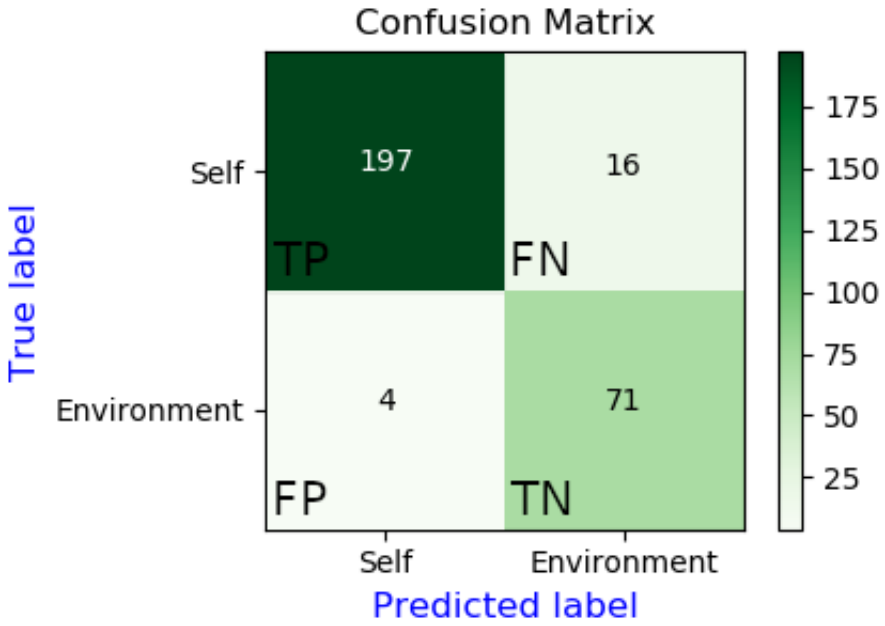

Figure 3.5: The confusion matrix of the validation dataset shows that the trained pilot module architecture learned to classify self and environment data groups.

3.3.5 Pilot Module Outcome

The proposed pilot module architecture aims to figure out a possible way to initiate the self in a robot by answering the research question proposed in the experiment section 3.3.4. In this regard, the initial acquisition of self was successfully initiated by designing a pilot architecture module using DNN and processing simulated robot multimodal data. This pilot model allows classifying the self by seeing the arms associated with the proprioception data. The model achieved an overall classification accuracy of 93.05%. Based on the aim of this section of assessing the potential of enabling self-awareness ability for a robot using DNN and simulated data, the outcomes of this section are as follows:

- The initial sense of self can be enabled using simulated multimodal data sensors of vision and proprioception (velocity).
- The pilot model architecture built using DNN is able to capture the sense of self by processing a multimodal dataset.

In the next section, an extended model is proposed to process a realistic dataset, as the environment is more complex with uncertainty because it has many uncontrolled variables that might affect the robot and its surroundings. Furthermore, a new architecture is proposed and evaluated to embrace the environment uncertainty using a real robot and real environment data.

3.4 Extended model: Enable the Sense of Self

The initial sense of self is enabled in the previous section of the pilot module of artificial self-awareness. Its experiment classification result shows a successful pilot exploratory self-awareness acquisition using a simulated dataset and deep neural network basic architecture. The motivation was to investigate the potential initiation of the sense of self for a robot using the simulation dataset of vision and velocity element of the proprioception. Based on the success of the pilot model experiment, this section proposes an extended model, which aims to enable a sense of self in a robot by using real robot data and an advanced DNN structure.

The approach in the extended model focuses on building an initial sense of self in the robot by enabling it to differentiate itself from the environment using proprioception and vision-sensing modalities. For this, a Deep Neural Network architecture is designed for the extended model architecture (Fig. 3.6) to support and understand the sense of self in the robot. The sense of self represents the level-1 of the implicit self-awareness discussed in Chapter 2, section 1.2. It also represents the self-awareness level one of Rochat's five self-awareness development levels [1] discussed in Chapter 2, section 1.1. Therefore, the first level of self-awareness inspires the extended module architecture design, and therefore, the proposed extended model can be mapped for robots as follows:

- Level-1: "Differentiation" is the level at which a robot is capable of getting an initial sense of self.

The proposed level-1 enables a robot to differentiate itself by seeing its arms and gripers associated with its proprioception in a scene. The assumption at this level is that the robot has a high-level description of its arms via perception and kinematics and can move its arms via motion planning. Therefore, a binary classification enables a robot to have a high-level prediction of its sense of self. The objective is to confirm if the arms and grippers belong to the robot. The outcome from this level is the initiation of self and consists of sensing the distinction between robot and environment.

In the self-awareness level-1 for humans reviewed in chapter 2, section 1.1, Rochat [1] emphasised that individuals begin to observe a unique experience on what they see with respect to their movements. Similarly, [3] and [62] stated that humans' inference is based on mapping multi-sensory input interpreted by their brains. Moreover, the previous section of pilot model architecture demonstrates a successful preliminary result of self by incorporating simulated multimodal data sensors such as vision and proprioception. Thus, the anticipation is that vision and

body movement are crucial components to initiate the self in a robot. Therefore, in this section of the level-1 artificial approach, the prediction of self depends mainly on two elements: the arms' presence in the robot's field of view and the sense of its movement.

The rationale behind the extended model is to embrace the environment uncertainty using a real robot and real environment data and define a neural network architecture that provides a way to learn the first level of self-awareness and encodes the internal mechanisms of level-1: Differentiation. Therefore, the predicted output of the neural network is a supervised binary classification task that predicts the sense of self of the robot by learning the distinction between the environment and the self as separate entities.

3.4.1 Extended Model Architecture

In this level-1 architecture, a real robot uses its visual and proprioception senses to discriminate its arms. For this, the robot's vision and proprioception senses complement each other to get an initial sense of self. The vision input comprises RGB images captured using a Stereo ZED camera from Stereolabs configured to output images at 720p resolution at 60Hz rate. The captured images represent the robot's arms or environment. The proprioception captured consists of the robot's joint states of three elements: angular position, velocity, and motor torque.

The architecture for level-1 of artificial self-awareness is shown in Fig. 3.6, consisting of a Resnet18 network [48] path to process the visual input state of the robot. Resnet18 is a residual neural network that utilises shortcuts known as a skip connection to avoid vanishing its network gradients, which is important to improve the network learning weights [48] [49]. It is a state-of-the-art architecture used widely for object detection and classification. The Resnet18 process vision data of 224X224 pixels of 3 RGB channels and output tensor of size 19 units. Similarly, for proprioception, a single fully connected network layer (FC0) processes the robot's internal state of 51 units and output tensor of 76 units. That is concatenated with the output of the Resnet18 network of the tensor size of 19 units. The concatenated tensor represents stacked features of 95 proprioception and vision modalities units. That is passed next to fully connected layers (FC1) and fully connected layers (FC2) to combine and associate the proprioception with visual representations features. A ReLU activation function follows the FC1; the output of the ReLU is a tensor of 32 units that inputs into the FC2 and outputs two units, which are used to predict self or environment.

3.4.2 Experiments

In level-1, the output of the experiment represents the initiation of self and consists of sensing the distinction between robot and environment using a supervised binary classification task that predicts the sense of self of the robot. This experiment result answered the following research questions:

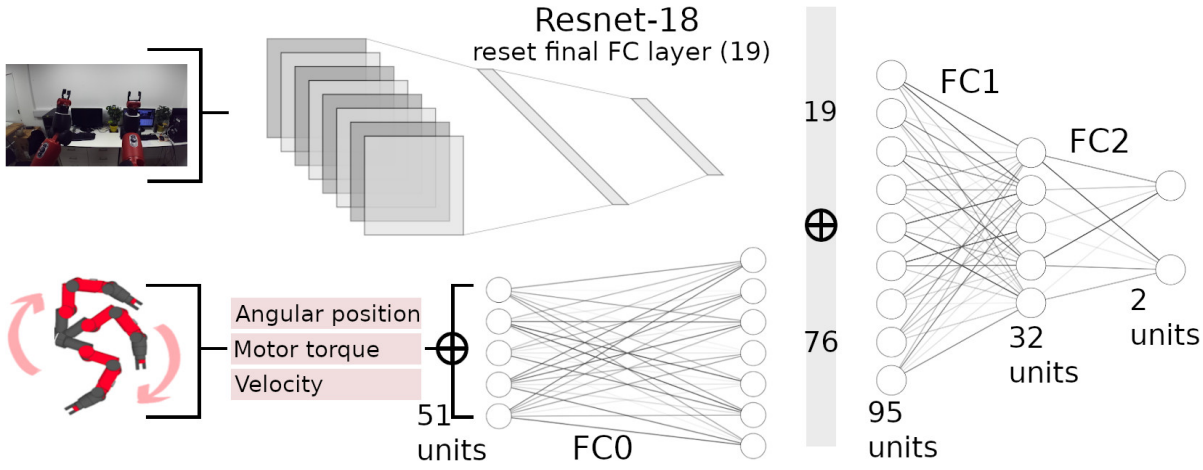

Figure 3.6: The extended model: level-1 architecture incorporate vision and proprioception inputs of the robot sensors to predict self or environment. As shown in the above architecture, the model process vision and proprioception through two subnetworks (Resnet18 and Linear layer, respectively) concatenates the outputs features into a linear layer and pass it to three fully connected layers to carry out a classification prediction.

- **RQ1**: Does developing the sense of self allow a robot to know itself?
- **RQ2**: Is the robot able to associate the observed and the felt movements in different environments with high accuracy?

The next subsections contain the level-1 real dataset description and the dataset groups, the conducted experiment details, and the results that conclude the level-1 implementation.

3.4.3 Real Dataset

The robot arms are calibrated before capturing the real dataset by performing a robot default factory calibration on both arms. The experimental design is framed based on the robot's capabilities. Specifically, the robot cannot move to a different place in the lab; thus, it is manually allocated into a different environment scene view. Similarly, the robot's vision camera sensor is fixed on top of the robot's head and cannot actively move its head. The camera's field of view complies with what a robot can see intrinsically. Moreover, the robot can move its arms freely within its predefined working volume, and there are no obstacles introduced in each of the captured scenes. According to Rochat [1], infants move their hands randomly, which helps build an experience that differentiates them from the world as a separate entity. We, therefore, commanded the robot to wave its upper limbs without a predefined task in the environment to enable the robot to learn to perceive and differentiate itself from the environment.

In order to train, evaluate, and test the level-1 of artificial self-awareness, a ROS node script was implemented to capture and store synchronised visual and proprioceptive sensor information of the physical robot (Baxter). Another ROS-Rviz simulator script is configured to guide Baxter through gripper poses of predefined points selected randomly to control Baxter's arms. During

real data capturing, Baxter's hands are recorded using the predefined points for both arms and within its working volume. A total of 30k images and proprioception states were captured over four different environmental settings, as shown in Fig. 3.7. Each scene represents a unique group that ranges from simple uncluttered (front towel and front glass) to cluttered (front computers and in the lab) environments. The scenes include two classification categories labels, namely self and environment. Furthermore, the description of the considered scenes are as follow:

Figure 3.7: Sample images captured from four scenes, these images represent "self" in a different environment (The robot arms are available in the field of view in the scenes)

Front Towel: A robot sees its arms within the field of view. A basic white background (a white towel) gives a robot a higher opportunity to focus on its features and not get distracted by the environment.

In Lab: A robot sees its arms while a lab scene is used as background, and this lab has objects scattered around. Moreover, the objects are not fixed in the scene; the object was interchanged by other different objects manually during the capturing.

Front Glass: A robot sees its arms within an environment with reflective glass in the background. Also, the environment is active as people pass during capturing.

Front Computers: A robot sees its arms in an environment representing a computer lab that has many objects. Moreover, different objects are added and removed to make an inconsistent changing environment.

The scenes mentioned above have been used with the robot to capture the dataset in different settings, such as the availability of both robot's arms in the field of view, only the left arm, only the right arm, and no arms existing in the field of view.

3.4.4 Experiment Groups

The level-1 self-awareness extended model was trained, evaluated, and tested with four different experimental groups formed from the captured real dataset scenes. An experimental group combines three scenes while leaving one out for testing purposes, as shown in Fig. 3.8. For example, experimental Group-1 consists of In Lab, Front Glass, and Front Towel captured scenes used to train the level-1 model and leaving out Front Computers scene for testing the level-1 trained model. The objective is to have broader and more diverse real data groups for training our proposed DNN architecture (Fig. 3.6). Also, to inspect the proposed DNN architecture's ability to generalise within an unseen group environment.

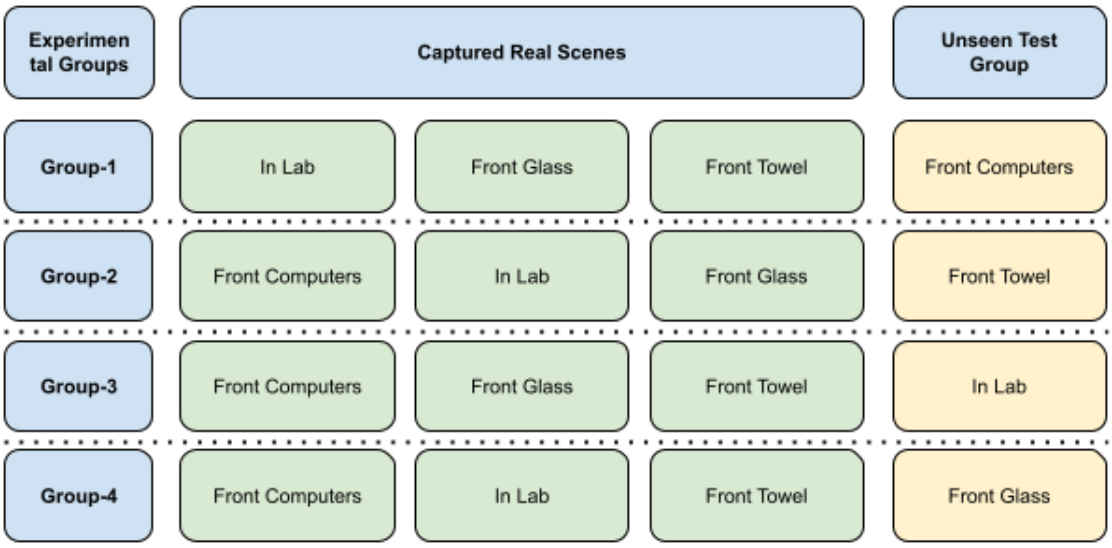

Figure 3.8: The experimental group (Blue) combines three different scene environments (Green), and the uncombined scene represents an unseen test group (Yellow) for the corresponding experimental group.

In order to create training and validation sets for each experimental group, a split of 80:20 proportions is created, respectively [80]. Accordingly, the training and validation sets represent approximately 20K+ and 5K+ images and proprioception states, respectively.

3.4.5 Confounding Cases

The proposed groups in the above Experiment Groups subsection have been prepared to manifest different signals. In order to further test the level-1 hypothesis that artificial self-awareness in the robot increases its self-certainty in an unseen environment, an ablation study is carried out represented by four confounding cases to understand the effectiveness of the combination of proprioception and vision within the proposed level-1 model. The objective is to confirm that the robot can differentiate itself with a degree of certainty while presented with confounding sensor signals this includes image data plus proprioception data.

Therefore, this experiment framework consists of four experimental groups, and each consist of four confound cases to compare with different robot perception signals. Similarly, the confounding signals are applied to the unseen test groups. As shown in Fig 3.9, Case-1 comprises images and proprioception correspond to the self class, while Case-2, images and proprioception, correspond to the environment class. Case-3 comprises confounding samples where the robot's arms are in the visual field of the robot, but the robot's proprioception corresponds to the environment class. While Case-4 is composed of environment images, the robot's proprioception comes from the self class. The two sensory inputs are important for the model to produce the decision of self or environment. Based on the defined confounding cases, the model will most presumably output an environment as the classification decision if any sensory input did not report as self.

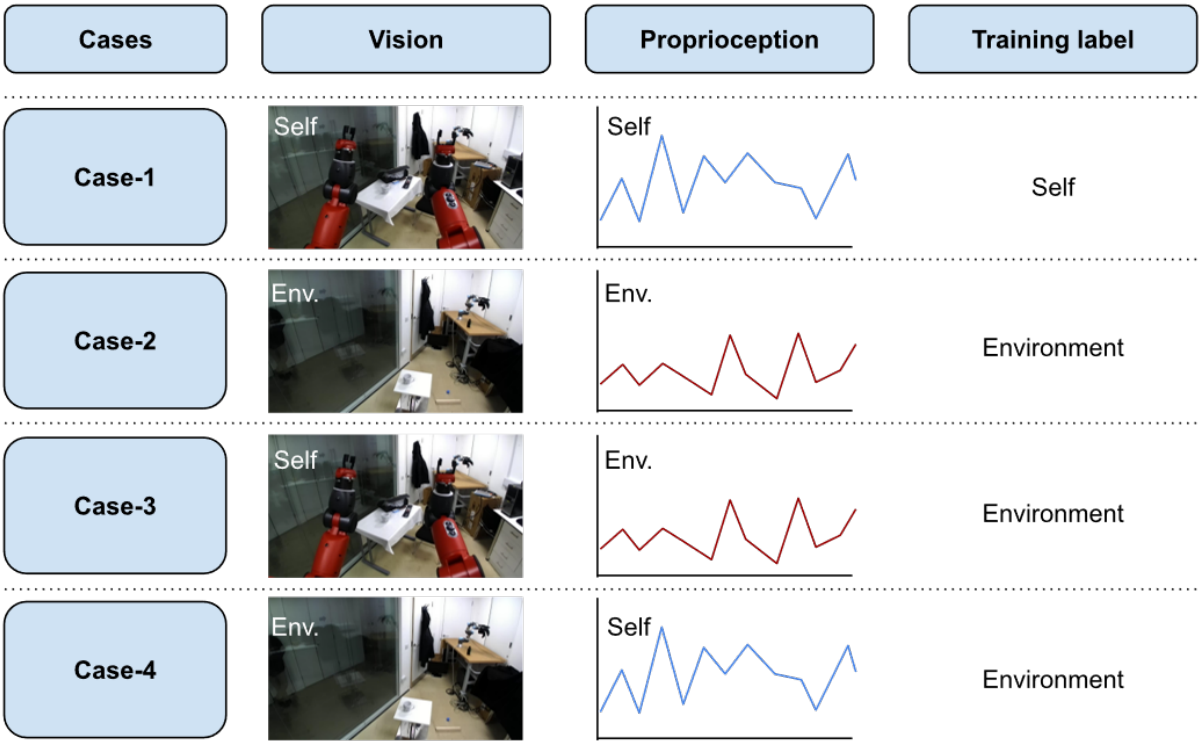

Figure 3.9: The confounding case presented by each row shows a different combination of vision and proprioception signals. The first row represents the self class, while the second, the third, and the fourth rows present cases related to the environment class.

The confounding case was formed based on the same group's information. For example, in Fig. 3.10, the group consists of two sets of Self and Env, each set divided into four equal subsets that combined to form the needed cases. The data distribution over all the cases is equal, and the unmapped data shown in Fig 3.10 was kept aside for troubleshooting.

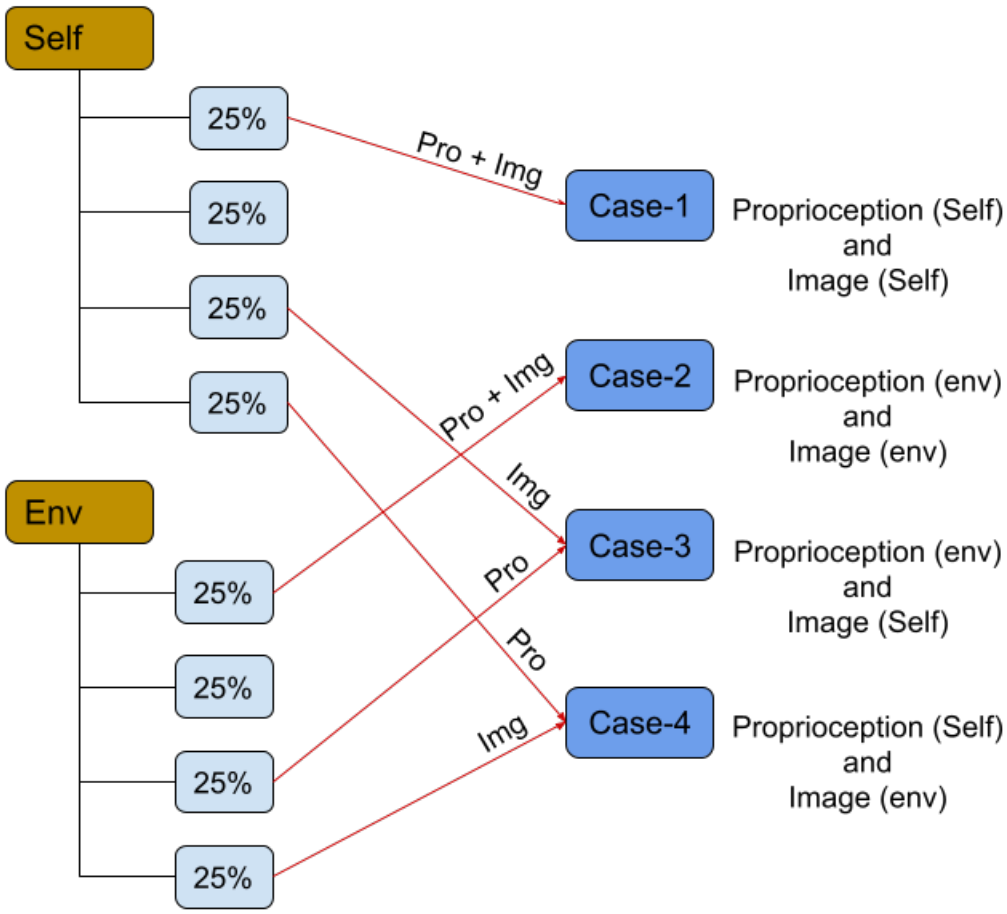

Figure 3.10: Case divisions structure creation.

3.4.6 Training and Validation

The experiment is conducted using PyTorch version 0.4 [45] to implement level-1 of artificial self-awareness architecture with PyTorch's default Binary Cross-Entropy loss [45]. A pre-trained version of Resnet18 is used and fine-tuned with the real dataset [48]. The training consisted of 24 epochs with a batch set to 64 for loading the proprioception and images dataset partially to fit the 8GB of GPU memory and a learning rate of 1e-3 that controls adjusting the learning weights for the loss gradient descent. A learning scheduler is used with a step size of 7 and a Stochastic Gradient Descent (SGD) as an optimizer. In comparing this Extended model with the earlier Pilot model from the hyperparameters perspective, this Extended model trained with 24 epochs due to the bigger dataset size and the fast learning convergence between the training loss and validation loss during the training stage. To understand and process level-1 in different environments, a leave-one-out cross-validation strategy is adopted to test each trained experimental group set reviewed in Fig. 3.8.

The level-1 model was trained and validated using the aforementioned experimental groups shown in section 3.4.4, Fig 3.8. The training and validation average losses in Fig 3.11 show that the level-1 architecture model is learning in all the groups. The training and validation loss

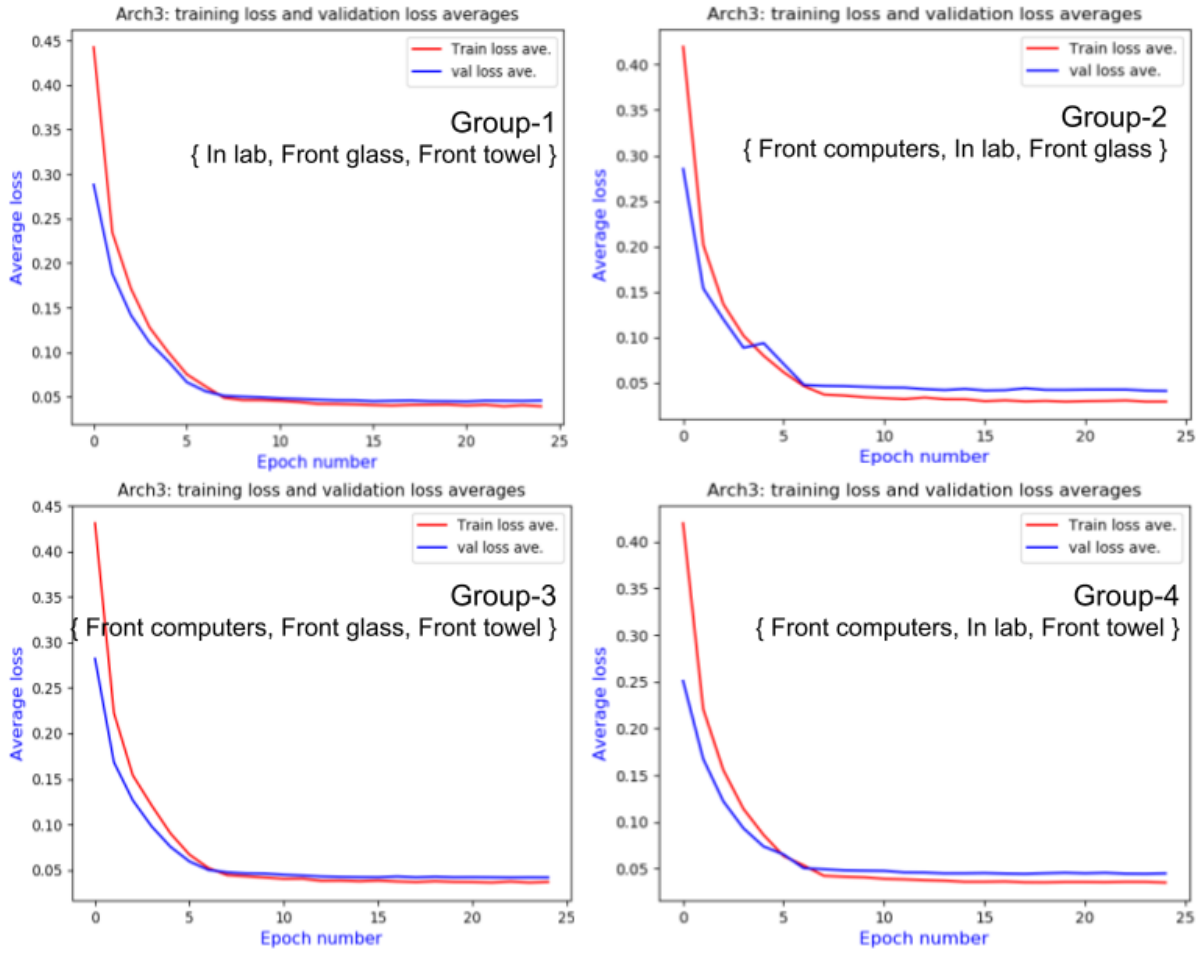

Figure 3.11: Training and validation average losses for the four experimental groups shown in Fig. 3.8

averages for all the groups decreased smoothly from 0.45 to 0.05. Accordingly, the validation accuracy for the groups Group-1 is 99.35%, Group-2 is 99.07%, Group-3 is 99.2%, and Group-4 is 99.1%.

The results of each unseen test group were validated in terms of classification accuracy, Saliency Map, and Mutual information in the following sections.

3.4.7 Testing

Classification Accuracy

The level-1 hypothesis validity can be verified by having an unseen experimental group from the dataset as the sense of self of the level-1 hypothesis assumes that devising self-awareness in a robot increases its self-certainty in an unseen environment. Accordingly, confusion matrices for each unseen test group in Fig. 3.8 are shown in Table 3.1. The classification accuracy for each unseen test group is: Group-1 is 88.1%; Group-2, 90%, Group-3, 82.1%, and Group-4, 94.7%. Therefore, the achieved results stated that the self-awareness level-1 architecture enables the

robot to differentiate itself from the unseen environment with an average accuracy of 88.7%. This result is based on the robotic self-awareness methodology carried out in this thesis study. Rochat [1] has no parallel quantitative data for the human self-awareness population at level-1 that can be directly mapped with the current thesis result of robot differentiation.

Table 3.1: Test groups classification accuracies for each unseen group (dark grey row) and their derived confounding cases(light grey rows).

	Group-1 Front Computers		**Group-2** Front towel		**Group-3** In lab		**Group-4** Front Glass	
Unseen Test Group	Predicted Self	Predicted Env.	Predicted Self	Predicted Env.	Predicted Self	Predicted Env.	Predicted Self	Predicted Env.
True Self	23.4%	1.6%	21.5%	3.5%	23.5%	1.5%	23.8%	1.2%
True Env.	10.3%	64.7%	6.5%	68.5%	16.4%	58.6%	4.1%	70.9%
Confounding **Case 1** Class: Self								
True Self	93.7%	6.3%	86.1%	13.9%	94.0%	6.0%	95.2%	4.8%
True Env.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Confounding **Case 2** Class: Env.								
True Self	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
True Env.	0.0%	100%	0.0%	100%	0.0%	100%	0.0%	100%
Confounding **Case 3** Class: Env.								
True Self	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
True Env.	0.0%	100%	2.4%	97.6%	0.1%	99.8%	2.2%	97.8%
Confounding **Case 4** Class: Env.								
True Self	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
True Env.	41.1%	58.9%	23.6%	76.4%	65.0%	34.4%	14.3%	85.7%

Saliency Map

A visualising technique is used for the neural network to further gain insights into the model's behaviour to show what it learns about the scene and the objects. A FlashTorch framework [81] is used to analyse how the self-awareness level-1 model perceives the input images through saliency maps. Saliency map shows which pixels in the image the neural network focuses on to predict a class output. In this artificial self-awareness model, vision and proprioception are used in predicting a class of self or environment, but only the visual part is evaluated with FlashTorch. With saliency maps, the level-1 model attention of the input images is analysed to get an insight

into regions that contributed most to the prediction output of self or environment, as shown in 3.12.

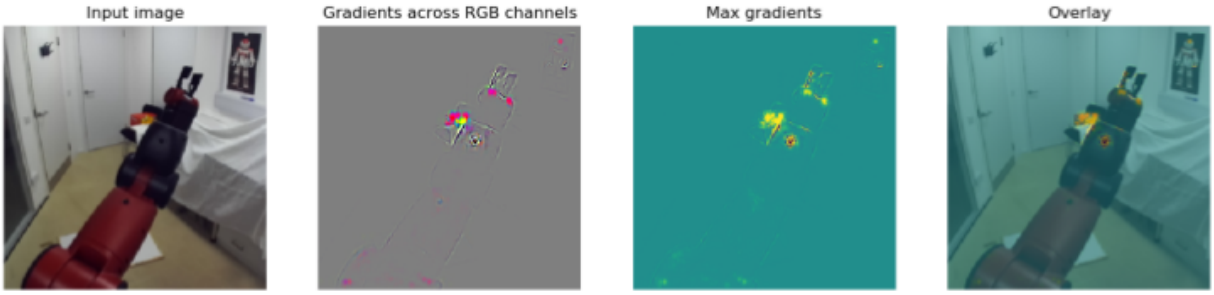

Figure 3.12: The first image (far left) represents a sample from the "In Lab" group input image. The three right images represent the saliency maps of the input image based on Gradient cross RGB, Max gradients, and the overlay of the gradients with the original image. They show that the level-1 module mainly focuses on the robot's arm related pixels.

In associating the source of information of the first row of unseen test groups in the classification accuracies, Table 3.1 (dark grey) with their information produced by saliency maps in Fig 3.13, the following are the common observations for each unseen test groups in Fig 3.8:-

In the Front Computers group, about 10.3% of the Environment class is predicted as Self; as the network fixates on the yellow pot in the environment (Fig 3.13-A), the network interprets the yellow pot as part of the robot (i.e. self class).

In the Front Towel, the classification accuracy is high at 90% in predicting the correct class, and the edges of the bottom towel are distinguished for the environment class (Fig 3.13-B). Moreover, Baxter's hand saliency pixels are highlighted more clearly (Fig 3.13-C). When Baxter's hands are within the scene, the network focuses less on the bottom of the towel, showing that when the robot's arms are within the field of view, the main focus is on the arm's features and discards the Environmental features such as the edges of the towel.

The In-Lab test group achieved a 17.9% of incorrect classification as the red cube, and other red objects got a network focus (Fig 3.13-D, -E, and -F); this shows the network is sensitive to objects coloured the same colour as Baxter's hand (red). Also, the network shows sensitivity toward bright colours such as red and yellow, and other bright environmental objects where the network gave greater weight to that objects.

The final Front Glass test group's classification accuracy is high at 94.7%. Despite Baxter's hand's features being well recognised (Fig 3.13-H), the network shows sensitivity toward bright colours such as yellow and gold (Fig 3.13-G). The level-1 architecture is biased towards bright regions in the images (i.e. picture frame in the background and table corner on the left-bottom).

The above overall groups' observations show that the level-1 network is sensitive to bright colours and brightness. Furthermore, some cluttered objects consistently coexisted in the robot field. Also, during the data capturing session, part of the clutter with distinctive features was randomly placed in the scene that attracted the level-1 network's focus.

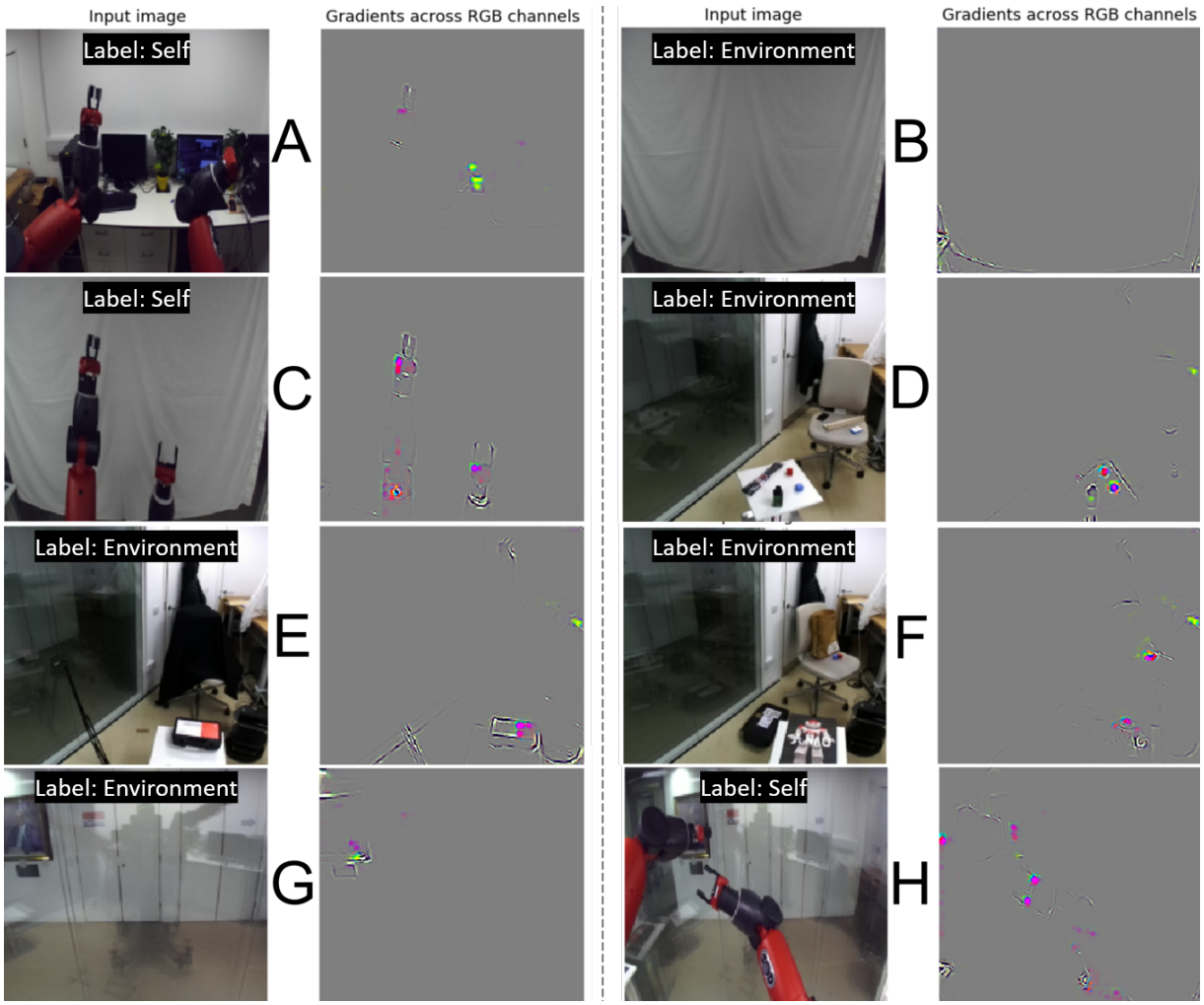

Figure 3.13: These images are representing the saliency maps of different environment groups as described in Table I, where A corresponds to Group-1; B and C, to Group-2; D to F, to Group-3; and, G and H, to Group-4. For each group, the right image shows the predicted label, and the left image shows the regions the model focused on.

The unseen test dataset is separated into four confounding cases described in Fig. 3.9 and each case was evaluated separately to investigate its classification accuracy errors association. The results for the cases are shown in Table 3.1 (the light grey rows). The unseen test of Front Computers (Group-1) and In Lab (Group-3) have noticeable classification errors of 11.9% and 17.9%, respectively, and reveal that Case-4 is the most misclassified case in both Front Computers and In Lab. The reason for these misclassifications is that Resnet18 is biased towards bright colours [82] [83], as discussed above. For Front Towel (Group-2) and Front Glass (Group-4), where the robot faces uncluttered environments, their confusion matrix results show better accuracy than the previous two groups because the evaluated test data for both Front Towel and Front Glass contain fewer environmental objects, which helps to yield a higher classification prediction. Also, from the observation in Fig. 3.13-C and -B that when the robot's hands become more predominant on uncluttered backgrounds, the level-1 architecture predicts the correct clas-

sification regardless of confounding signals coming from proprioception. Also, Table 3.1 shows proprioception signals have a high contribution to predicting the correct class. For instance, in case-3, where images contain the robot's arms, but proprioception corresponds to the environment class as mentioned in Fig 3.4.5, the level-1 architecture can accurately predict the correct class for all groups.

Mutual Information

The mutual information [84] is used to understand further whether the level-1 architecture learns to differentiate the robot from the environment. Accordingly, the Mutual Information is computed for the four models trained using the experimental groups' training datasets (Fig. 3.8). The mutual information is computed at the last layer of the self-awareness level-1 architecture; thus, proprioception is considered during the mutual information computation.

The objective is to measure and compare if the four level-1 trained models have a degree of similar knowledge invariant to the training set. The multimodal sources are compared with mutual information, and how well two sources are matched is measured by mutual dependence between two variables. Different sources of information mean more distributed points in the joint histogram and, consequently, low mutual information metric. The spread in the joint histogram is associated with uncertainty. In Fig. 3.14, joint histograms show minor variability in the correlation between the group's model weights. This confirms no significant differences between the trained models despite the differences in the training datasets (Fig. 3.8). Also, it confirms that the misclassification in the confusion matrix results (Table 3.1) is based on the environment noise as other objects within the environment distract the network's attention. Therefore, this demonstrates that the level-1 network architecture captures a degree of self-awareness and, consequently, a certainty; about the robot's body knowledge within different environments.

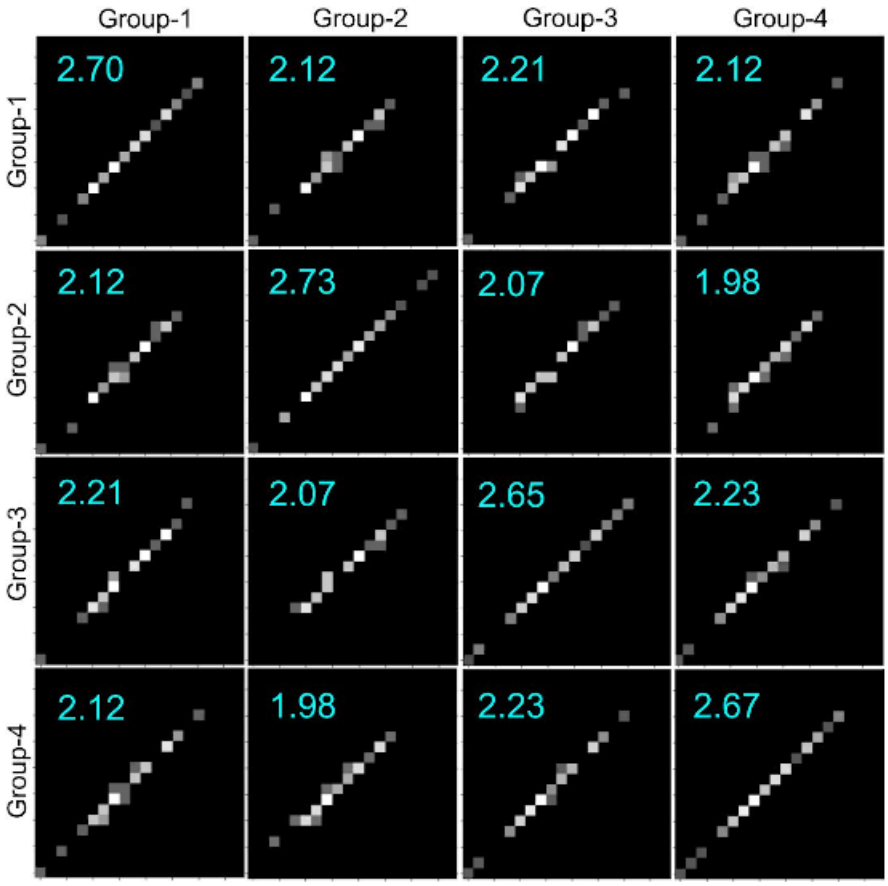

Figure 3.14: Mutual information 2D joint histograms of the trained weights of four level-1 architectures, and each is plotted across different groups' weights. The mutual information is noted at the top left corner of each joint histogram plot.

3.5 Limitation

Group-1 and Group-3 in the extended model experiment show that the resnet18 used in its architecture influences the classification accuracy because it is biased towards bright colours; thus, the Resnet18 is detached in the next chapter's model architecture. This study of self-awareness is planned to be developed modular, which means any issue that affects the current model will transfer to the next model of self-awareness, which will present unexpected results for the next levels.

In addition, the DNN architectures in both pilot and extended models are difficult to troubleshoot. They cannot present their data to its original state, so qualitatively decision can be made; the idea is to employ a modular architecture with scalability support and interpretation ability. That will allow easy troubleshooting and analysis of any issue during the model's training, testing, and expansion with the next levels of self-awareness. The pilot and the extended models cannot show what their models are presenting. Thus, the next chapter comprises developing level-1 of artificial self-awareness revision to mitigate the current limitations of the pilot and the extended model's architecture.

3.6 Conclusion

This chapter proposed two evaluation approaches for a possible start of artificial self-awareness acquisition, inspired by the first level of self-awareness defined by Rochat [1]. The first approach of pilot model architecture investigated the potential of initiating a sense of self by using simulated data input consisting of multimodal data. The pilot study experiment result shows:

- A high possibility of initiating the sense of self from simulation multimodal input by average classification accuracy of 93.05%.

The second approach of the extended model initiates level-1 of artificial self-awareness sense of self in a real dual-arm robot. This includes using the Resnet18 network as vision, three proprioception elements, and processing real scenes multimodal datasets of a real robot with different environments. The experiment demonstrated that a robot can differentiate itself from the environment by achieving the following outcomes:

- Average classification accuracy of 88.7% using unseen test groups and across four different scenes' groups are presented in Fig. 3.8.
- Presented a saliency map showing which pixel regions the neural network focused on to predict the underlying output. The prediction output helps fetch different observations in conjunction with the classification results from different groups.
- The presented mutual information between the group's trained modules does not vary significantly, thus representing a basic, minimal self [13] within the neural network and robot.

This chapter verifies two of the hypothesises of this thesis:

- The pilot model hypothesis: ***The sense of self can be enabled in a robot using DNN model architecture and association of simulated multimodal data.***
- The extended model hypothesis: ***Level-1 for artificial self-awareness in the robot increases its self-certainty in an unseen environment.***

The proposed artificial self-awareness pilot and extended architectures of the level-1 experiment outcome show that a robot can develop a sense of self. Nevertheless, the robot cannot locate its limbs within the environment and put them into context for a task, as that requires the next level of self-awareness. Moreover, before moving into the level-2 artificial self-awareness model investigation, a new artificial self-awareness level-1 revision is proposed in the next chapter to mitigate the issues reviewed in the limitation section 3.5 of this chapter.

Chapter 4

Artificial Self-Awareness Level-1

In this chapter, the preliminary level-1 of artificial self-awareness acquisition reported in Chapter 3 is modified and advanced to a more scalable and interpretable architecture, that can reconstruct the data to its original state, it can also produce a stable fuse multisensory state. The proposed artificial self-awareness level-1 revision architecture is designed using neural networks generative modules. Differing from the preliminary investigation in Chapter 3, in this chapter, a Variational Autoencoder (VAE) is used to process the data input and an Autoencoder (AE) to fuse the sensory signals. The designed architecture is used to devise a robot with level-1 of artificial self-awareness.

4.1 Introduction

This chapter presents the revised version of self-awareness level-1 architecture built based on generative modules networks. The proposed self-awareness level-1 architecture used a variational autoencoder to learn a low-level representation of high-dimensional data and used an autoencoder to fuse the data to represent a relation state that complies with the sense of self. This work presents a robust version of artificial self-awareness level-1 architecture (Differentiation), incorporating a fusion of vision and position data modalities.

The level-1 hypothesis of this thesis reviewed in Chapter 1 section 1.6 are revisited to be assessed throughout this chapter:

- **A robot's self-awareness sense of self can be enabled using an unsupervised learning method to associate and reconstruct the fused real multimodal data.**

The chapter is structured as follows: Section 4.2 gives the motivation and objectives of the chapter. Section 4.3 reviews the new artificial self-awareness level-1 architecture. Section 4.4 presents the Level-1 baseline experiment. The level-1 VAE experiment is reviewed in Section 4.5. The discussion of level-1 baseline and level-1 VAE experiments is in section 4.6. Finally, the conclusion of this work is given in Section 4.7.

4.2 Motivation and Objectives

The proposed approach to designing the self-awareness levels follows a modular format that starts by defining a sense of self by differentiating it from the external world and then integrating the other models representing the remaining self-awareness levels. The main idea of this approach is to build self-awareness models from inside to outside, as discussed in chapter 1, section 1.8 and define the internal self before a robot model interacts with the external environment. In this chapter, the proposed architecture for artificial self-awareness level-1 dropped the Resnet18 favouring the Variational Autoencoder (VAE). Resnet18 did the required feature extraction, but within level-1 boundary only, and it did not fit the overall self-awareness incremental module design approach. The main issue was that Resnet18 had shown sensitivity toward bright objects [82] [83], which affected the classification accuracy results, especially within cluttered environments. To avoid future feature extraction problems and ensure that the latent space reflects the feature representation, AE and VAE-based architectures are implemented instead. It is crucial to have an architecture that can learn a low-level representation of high-dimensional data and reconstruct its original data state for interpretability. Thus, two architectures are implemented: an architecture based on AE networks for level-1 artificial self-awareness to extract features into a latent space and fuse them. An exemplary architecture based on VAE for level-1 artificial self-awareness will ensure feature extraction based on probability distribution and model expandability to accommodate the next level-2 and future levels of self-awareness.

In addition, in the previous chapter 3, the DNN module architectures verified its ability to process the real dataset of robot modalities. Furthermore, it demonstrated the ability to integrate the robot signals of different modalities. These outcomes participated in the design of this chapter proposed approach by continuing using the robot's real dataset.

The objective of this chapter is to build a final level-1 module based on the following features:

- Reconstruct a scalable and interpretable level-1 of artificial self-awareness architecture.
- Build an architecture that processes multimodal sensory data.
- Build an architecture that fuses the vision and positions data into a low-level data representation.
- Construct an output state of integrated vision and position vector to be processed by the next level of artificial self-awareness (level-2).

4.3 Artificial Self-Awareness Level-1

The level-1 of artificial self-awareness is constructed by letting the robot learn to differentiate itself by seeing its arms, including hands and gripers, in association with the position data. The

robot will have an understanding of its limbs' appearance and will be able to confirm if the observed arms belong to it. In this level (differentiation), the self-model is initiated with an initial sense of self.

The rationale for redesigning self-awareness level-1 is to overcome the limitation of the previous level-1 model by implementing a new level-1 architecture that can anticipate extracting robust features from the input data. The previous level-1 model architecture in chapter 3 shows a sense of self and complies with the hypothesis for level-1 of self-awareness. Nevertheless, it was not providing stable multimodal signals. The reconstruction ability is very important for the current approach for the artificial self-awareness study because it can be utilised to see what a robot sees and as a tool to troubleshoot the model. The reconstruction of a low-level signal into its original signal helps advance self-acquisition by tracking and understanding different stages of the architecture. In this modular approach, the data produced by the first module of level-1 gets processed by the second module of level-2, i.e. From the data perspective, the module are dependent on one another; each module has an expected output that allows being manageable for implementation and maintenance. Furthermore, the output signals of the level-1 architecture do not reflect the required cross-modal relationship of the input multi-modality due to the issue found with Resnet18 network outputs, as discussed in Chapter 3. It is biased towards bright colours, which will exacerbate the issue of unwanted data processed by the next level module of artificial self-awareness.

4.3.1 Level-1 Model Revision

The level-1 self-aware model employs real data of vision and proprioception as the primary sensory inputs. The visual sense is used to recognise and understand that the robot exists within the scene and is associated with proprioceptive information to confirm the robot's awareness status internally. The association between the two inputs allows a robot to know itself as an entity. Two implementations of the level-1 in this chapter are divided into two main parts: The level-1 baseline and the level-1 VAE experiments. The level-1 baseline experiment architecture is built based on the classical AE networks. However, the level-1 VAE experiment architecture is built based on VAE and classical AE.

The experiments result of the level-1 baseline was limited to answering only the first research question, but the level-1 VAE answered both the following research questions:

- **RQ1**: Can a robot initiate a sense of self by associating its multimodal data using an unsupervised learning method?
- **RQ2**: Can a robot's sense of self enable it to interpret what it sees and sense by reconstructing its fused multimodal features?

4.4 Level-1 Baseline

This section employs a generative model of the Deep Neural Network for the level-1 baseline design, that learns and captures the sense of self by processing the modalities of proprioception and visual information. The architecture incorporates an unsupervised learning method represented by an Autoencoder and a supervised learning method presented by a classifier network. The designed level-1 baseline of artificial self-awareness architecture must be extendable and able to integrate levels-2 of self-awareness studied in the next chapter.

The baseline experiment used an Autoencoder to implement the feature extraction network for level-1 artificial self-awareness. The baseline artificial self-awareness level-1 model structure consists of three Autoencoder networks to encode and decode robot sensors. As in Fig. 4.1, the model processes input of perception and internal robot state using Autoencoders. The encoder's networks will transform the images and proprioception sensory inputs into a latent space. The next parts of the Autoencoder networks are the decoders that reconstruct the inputs from the latent space vectors. The latent vector output from the vision encoder is a tensor of size 12,544 units, concatenated with the latent vector output of the proprioception encoder tensor of size 10 units. The concatenated tensor size of 12,554 units is passed next to Mix Autoencoder to build the relationship between the perception and the internal robot state latent spaces. The Mix Autoencoder network encodes the tensor for mixed-signal feature representation. The mixed-signal features latent space is a tensor size of 32 units processed by a classification network and output a tensor of 2 units to be used to predict either Self or Environment. Also, the same output of the mixed-signal features latent space of the 32 units size is planned to be used for the next module of level-2 self-awareness.

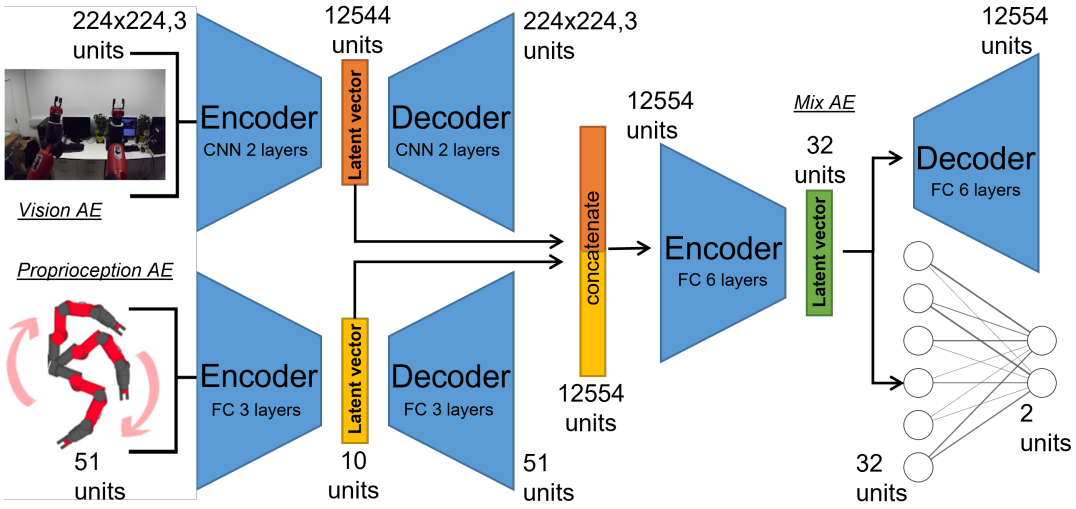

Figure 4.1: The artificial self-awareness level-1 baseline architecture is built based on Autoencoders and classifier networks.

4.4.1 Baseline Dataset

The proposed baseline level-1 artificial self-awareness architecture employed input sources used previously in the level-1 model as primary sensory input, such as vision and proprioception. The dataset used for this baseline level-1 is the same real dataset of the robot captured and used in Chapter 3. The framework PyTorch [45] is used to implement the baseline level-1 of the self-awareness model. In the data loading stage, the dataset loaded from training, validation, and test sets got a preprocessing of data transformation such as resizing images to 224X224 and converting the images and proprioception into tensors.

4.4.2 Experiment Groups

The vision and body movement are essential as input to have different scenes with the same presence of the body object. The relation of body movement sensation with vision confirmation generates a sense of self. In the baseline, the level-1 experiment followed the same experimental framework as in Chapter 3, section 3.4.4, where four experimental groups of datasets are presented, each formed by three different scenes of environment for the training/validation and one for the test as a leave-out group.

4.4.3 Confounding Cases

The same four confounding cases framework used in chapter 3, section 3.4.5, is also implemented for the baseline level-1 self-awareness experiment. These confounding cases test the supervised classification part trained based on the Autoencoder latent vector. Furthermore, only the first two confounding cases of the framework are also used for the same architecture to validate the reconstruction ability of the Autoencoders.

4.4.4 Experiment

The level-1 baseline architecture was trained in two phases, the first phase focused on the reconstruction ability, and the second phase focused on fusion accuracy. Moreover, both iterations used different confounding cases; in the first phase, only used the first two confounding cases of case-1 and case-2 represent self and Env, respectively. This discrimination between the two phases avoids sending conflicting signals for the model architecture. The first phase loss function is based on the cross entropy criterion and mean squared error, sequentially back-propagated across the architecture submodels. The aim of this is to check the reconstruction ability of the module. The second phase used the four confounding cases framework. The second phase's loss function is based on the cross entropy criterion between the predicted and targeted classes processed end-to-end. The aim is to check the architecture's fusion accuracy across different confounding cases.

4.4.5 Level-1 Training and Validation

The architecture processed data with a batch set reduced to 32 with shuffling to avoid GPU memory overflow due to the additional proprioception elements added, such as position and effort. Also, the network architecture used an adaptive moment estimation (ADAM), an optimizer for deep neural networks that can handle sparse gradients and easily adapt to most problems [85]. The optimizer is configured with a learning rate of 1e-3. The baseline level-1 network architecture is trained for 10 epochs, enough to produce a latent vector representation that can be reconstructed using its corresponding network decoders. In addition, the supervised classifier is trained for 25 epochs to predict Self or Env label classes of the encoded mix latent vector. The vision and the proprioception networks' latent vectors are concatenated and processed by the mixed encoder network. On this artificial level-1, the prediction will depend mainly on the presence of the hands and the sense of movement. The fused output represents the self's initiation by sensing the distinction, i.e., the relation in perception related to internal movements; this relation is trained in this level-1 baseline model, which predicts Self or Env.

The baseline level-1 artificial self-awareness network architecture was trained and validated the network architecture's ability of encoding and decoding the four experimental groups of two cases. The validation data is encoded into a latent vector and decoded to its original state. The result of Mean Square Error (MSE) between the reconstructed data and the original data of the first batch for each experimental group are shown in table 4.1.

Table 4.1: The MSEs of batch zero of the validation set reconstruction between the original and the reconstructed data by the baseline decoders.

Validation set Reconstruction	Images MSE	Pro. MSE	Sliced Images MSE	Sliced Pro. MSE
Group-1	0.0350	0.3911	1670.7460	7628.2314
Group-2	0.0277	0.3432	7075.6558	13385.6670
Group-3	0.0248	0.2920	37935.1914	231459.203
Group-4	0.0245	0.3659	34060.8320	69068.3906

The baseline level-1 artificial self-awareness network classifier was trained and validated four times, each using the corresponding trained network with the matched experimental group. 4.2, the classifier training and validation average losses show that the baseline level-1 network is learning across all the groups. Accordingly, the validation accuracy for the groups is; Group-1 is 99%, Group-2 is 99%, Group-3 is 98%, and Group-4 is 99%.

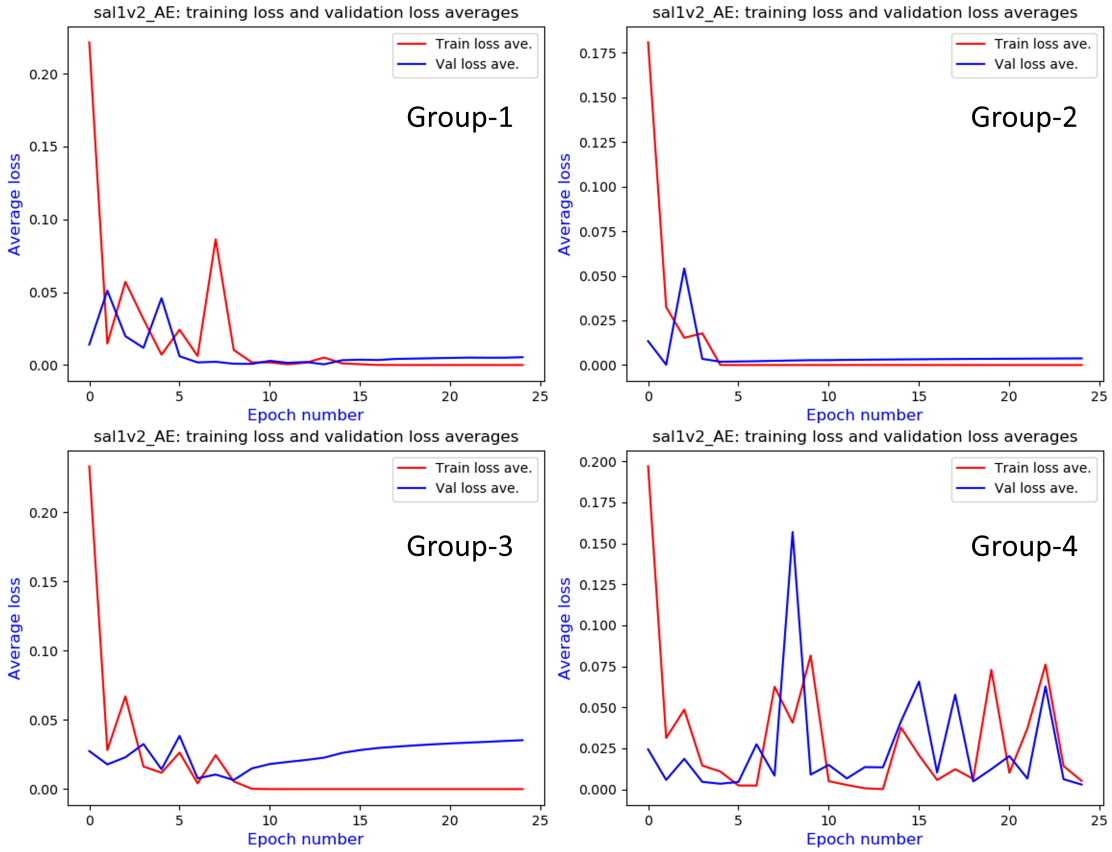

Figure 4.2: Training and validation average losses of the baseline level-1 classifier using the four experimental groups. The Group-1 and the Group-2 charts show that the level-1 network is learning and show fluctuations that are settled after the 10th epoch. Groups-3 and-4 after the 10th epoch produced a different pattern; in Group-3, the learning and validation trends started to diverge, which means there is no more information in the data to keep the network learning. However, In Group-4, the trend kept fluctuating but within a range of low average loss, indicating Group-4 ability for more space to learn from its data group.

4.4.6 Level-1 Classification

The classification accuracy result includes the confounding cases reviewed in chapter 3, section 3.4.5. The four unseen test datasets were tested each with its corresponding experimental group framework as in Chapter 3, section 3.4.4, Fig. 3.8. Accordingly, the baseline level-1 classification accuracy for each test group is as follows: Group-1 is 74.22%, Group-2 is 74.20%, Group-3 is 75%, and Group-4 is 87%. The overall average accuracy overall the groups is 77.55%.

4.4.7 Level-1 Reconstruction

The baseline level-1 reconstructions of the two modalities of vision and proprioception observe a high similarity with respect to their original signals of the four groups with MSE averages of 0.028 and 0.34805, respectively. The vision decoder network reconstruction of latent space reflects what the encoder processed. Also, the proprioception decoder network reconstruction

of latent space reflects the original proprioception signals. The mixing network is not producing the expected latent space; the sliced parts of vision and proprioception produce high MSE values all over the groups, as shown in table 4.1. Thus, the vision and proprioception decoders could not reconstruct their original data, as the observer similarities concerning the original signals are low. Therefore the decoders of vision and proprioception cannot handle the latent vector changes made by the fusion network.

The baseline level-1 reconstructions show that the new network architecture design for artificial self-awareness can encode and decode the information fed to the network. However, this is only valid for the vision and proprioception networks. The second row of "Reconstructed -1-" in Fig 4.3 shows vision reconstructions, which are almost identical to the "original" data samples, and the data lost in comparing both of them have an MSE value of 0.0334 for the four sample images of self and env. The last row of "Reconstruction -2-" represents a noise as the network could not reconstruct the image part split after the vision modality got fused and decoded by the mix network. The proprioception network reconstructed the robot's proprioception signals with minimum data loss of MSE of 0.4037 of a random sample of Group-2, as in Fig 4.4. However, the proprioception signals split part could not be reconstructed by the proprioception decoder after it fused and decoded by the mixed network model.

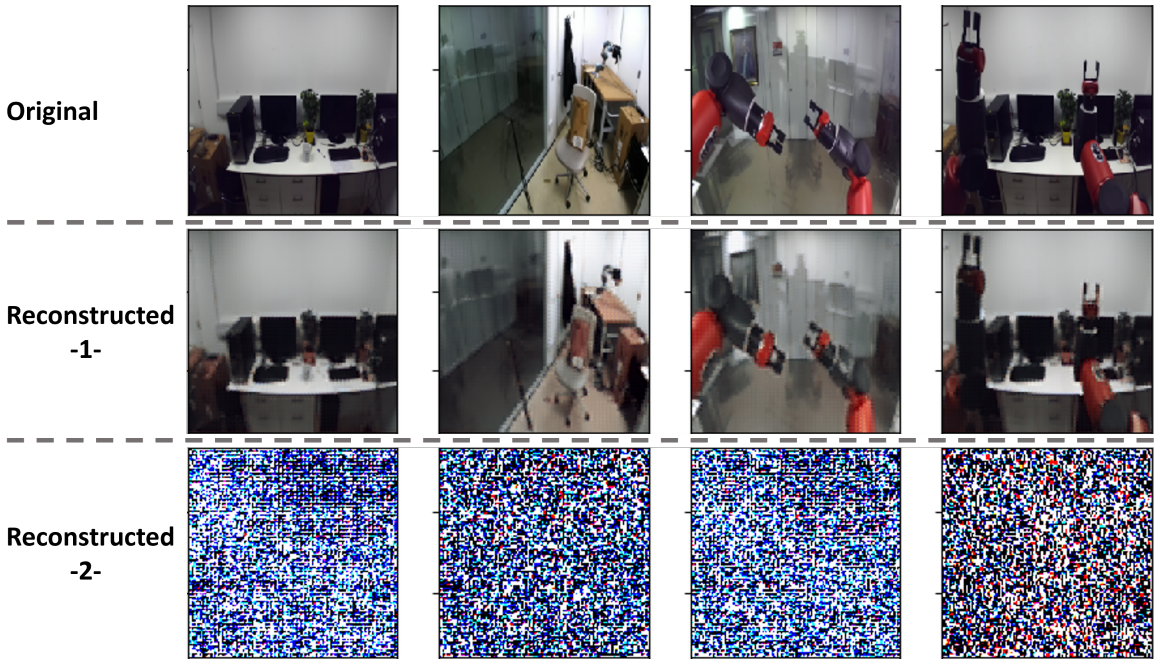

Figure 4.3: The reconstructions of random samples from Group-2 data, the first row has the original images. The second row, "Reconstructed -1-", contains the reconstructed images from latent space representation. The third row, "Reconstructed -2-", has reconstructed images after separation from the predicted latent vector by the mix AE network.

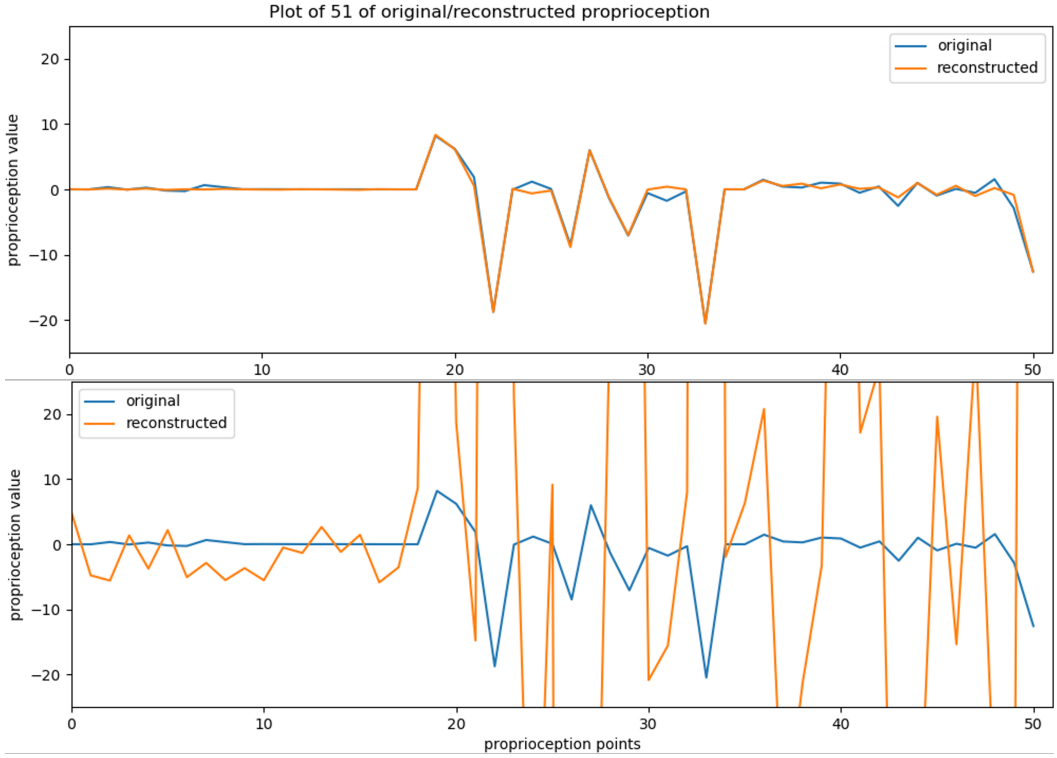

Figure 4.4: The two charts represent reconstructions of a random sample from the Group-2 dataset. The upper chart contains 51 units of the original signals and their reconstructed signals by the proprioception network decoder. The lower chart contains the same original signals and their reconstructed signals using the proprioception decoder after they were fused and split by the mix AE network.

4.4.8 Outcomes

The baseline level-1 network architecture classifier with the Autoencoders performs well by 77.55% as average over the four experimental groups using unseen test datasets. Still, the classical Autoencoder does not guarantee the features extraction part because the Autoencoder might learn to copy the input and not be able to handle a variation to its latent space vector [86]. Thus, another type of Autoencoder is considered to force the features extraction part, such as a Variational Autoencoder (VAE) [50]. The VAE potentially produces better accuracy over the unseen test groups because of its latent continuity ability [86]. It is vital to consolidate level-1 architecture as per the considered modular approach; other planned self-awareness levels such as level-2 and future work implementation utilise the features produced by the level-1 module. The classical Autoencoder is employed in the baseline level-1 of self-awareness architecture. This module is considered a baseline that can be compared with the next section's results.

The level-1 baseline experiment answered the first research question reviewed under section 4.3.1 of this chapter, as the classification result shows an accuracy of 77.55% in distinguishing the self class from the env class using the four groups and confounding cases framework, which means that the model able to devise sense of self for a robot. However, the level-1 baseline

model could not reconstruct the multimodal vision and proprioception after the mix network fused them. Thus, the second research question was not answered by this model experiment. The level-1 baseline experiment did not validate the hypothesis of this chapter reviewed in section 4.1.

In the next section, the level-1 VAE model is presented to overcome the issue reviewed with the level-1 baseline model associated with the inability to reconstruct the mixed modality, which also affects the multimodality fusion features.

4.5 Level-1 VAE

The proposed level-1 VAE architecture utilises the proprioception and the visual information by incorporating a collection of DNN, such as unsupervised learning methods of Variational Autoencoder (VAE) and Autoencoder (AE). Also, a supervised learning method for the classification. The main aim is to restructure the final design of self-awareness level-1 using VAE to overcome the issue associated with Resnet18 discussed in chapter 3, section 3.5 and the Autoencoder copy limitation discussed in the outcome section of the Baseline level-1 4.4.8. VAE has a generative ability based on a latent distribution to investigate and analyse the output, a required feature for the proposed modular approach that considers incremental levels of artificial self-awareness in the future. The VAE sampling latent vectors drawn from different input features create a better multimodality representation from both inputs' sensory sampled latent spaces. The restructured level-1 VAE of self-awareness based on VAE is shown in Fig. 4.5.

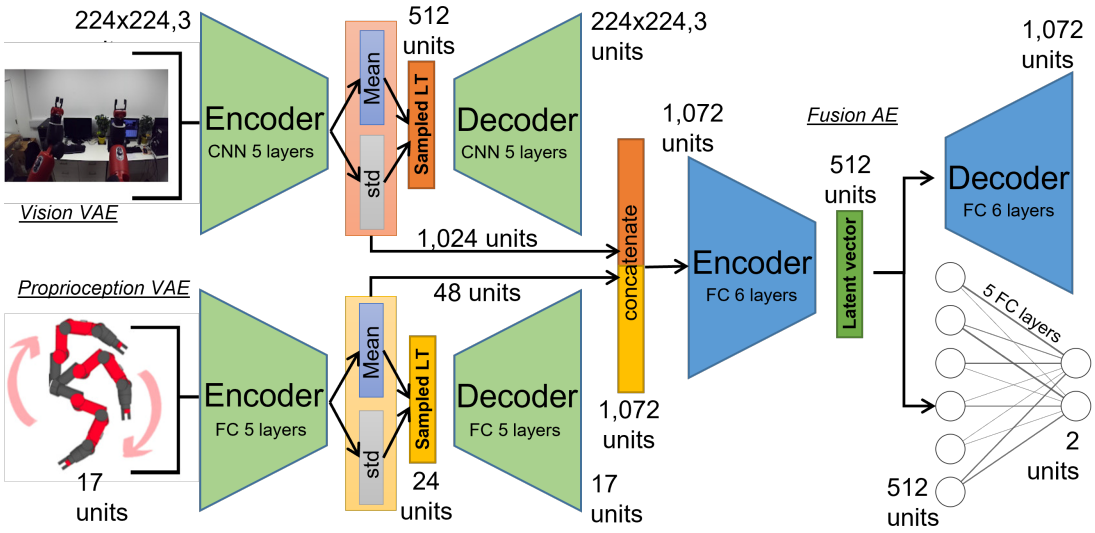

Figure 4.5: The artificial self-awareness level-1 VAE architecture, based on Variational Autoencoder networks as features extraction network, and Autoencoder network to mix the sensory features. The classifier network checks the network's ability to classify the integrated input features.

In the context of the level-1 VAE architecture Fig. 4.5, the VAE has an encoder network

to encode the vision and position inputs into a latent space distribution. It also has a decoder network to sample and decode the latent space distribution into its original input form. Implementation of parallel CNN-VAE and an FC-VAE, each with 5 layers to process the vision input by encoding an image of 224X224px and the position input vector of 17 units into an output of two vectors of 512 units for the image output and two vectors 24 units for the position output, that output vectors describing the mean and variance of the latent state distributions. After that, the vectors of the mean and the variance of both latent distributions VAEs were concatenated in a fully connected layer of size 1,072 units. Next, the fusion network is implemented with an Autoencoder consisting of 6 fully connected layers for each encoder and decoder network. The fusion network gets the 1,072 units output from the VAES parallel networks as input and encodes it as a latent space of 512 units, representing a robot feature state of the input data. After that, a classifier network of 5 fully connected layers classified the features snapshot of the sensory inputs as self or environment. The CNN-VAE, FC-VAE, and classifier used the activation function of ReLU, Tanh, and Tanh, respectively. The ReLU activation function rectifies the data from 0.0 to 1.0, which is suitable for processing the CNN data. The Tanh activation function has the domain of a negative number, which is suitable to process the proprioception as its data ranged from -1 to 1.

4.5.1 Level-1 VAE Dataset

The proposed level-1 VAE artificial self-awareness architecture employed input sensors used previously in the above baseline level-1 model as primary sensory input, such as vision and proprioception. Specifically, in this level-1 VAE, the input of the self-awareness framework is fed by two sensory sources, robot proprioception (angular position) and robot vision. The level-1 VAE experiment follows the same dataset framework of experimental groups and confounding cases. The only position element is considered for two reasons: the elements in the proprioception have different scale units; they can not process all units with the same network. Each element needs to be considered in a different sub-network to comply with the recommended approach of processing different modalities as discussed in chapter 2 2.1.8. Also, the position is needed to fulfil the current stage of self-awareness, i.e. a robot needs to make an association between the arm's internal position state and its vision state to fulfil the sense of self.

The position data tensors P are preprocessed by normalising its tensor elements between the range of -1 and 1 using equation defined in 4.1. And, the vision image tensors are preprocessed by the range of 0 - 1.

$$\frac{sin(P)+cos(P)}{\sqrt{2}} \tag{4.1}$$

4.5.2 Experiment Groups

This level-1 VAE model used the same proposed experiment groups explained in chapter 3, section 3.4.4, but without leave-one-out because more data is good to enrich the VAE distribution. The four groups are included to train and validate the level-1 VAE model. The validation set is not trained and is evaluated instead of the unseen. The model focused on representing a space distribution for the vision and position networks by training their models of four groups to enhance the generative capability. Also, the dataset is bigger by including the four groups; the generative ability is required for model interpretation where the data's original state can be restored from a distribution, as mentioned in the motivation and objectives of this chapter.

4.5.3 Confounding Cases

This level-1 VAE model has used the same proposed confounding cases created and explained in chapter 3, section 3.4.5, and mainly are used for the fusion AE classifier. The Vision, position, and fusion models used only the first two cases (Case-1: Self and Case-2: Environment.)

4.5.4 Level-1 Training: Vision and Position Networks

PyTorch's framework [45] is used to implement the level-1 VAE artificial self-awareness models. The training of the level-1 VAE architecture is phased into separate sub-cycles of training and validation.

The loss functions used for each vision and position VAE network are composed of two parts. The first part is the reconstruction loss which is calculated based on the MSE of the decoder output with its original input. Moreover, the second part is the latent loss presented by the KL divergence loss, which penalises the VAE if the generated information is out of the desired distribution [52].

The VAEs networks use the Adam optimisation algorithm with a learning rate of 1e-4, which is lower than in earlier experiments. The main reason is to let the network learn in smaller steps to reach the best minimum learning point swiftly. Moreover, the data input increased to 64 batches with shuffling because only the position element of the proprioception is involved in this experiment, which can be accommodated within the current GPU memory capacity of 8GB used in this study. The vision VAE and position VAE networks were only trained and validated using the first two cases (Self/ENV); their network's weights state was saved after training of 350 epochs for Vision VAE and 80 epochs for position VAE. The purpose of using higher numbers of epochs for the VAEs used here is to ensure that the networks learn and form sparse distributions representing their data features. The training and validation losses result in Fig. 4.6 show that over 350 epochs, the vision model is progressing in learning the Self and Env features, and the training and validation trends of the average losses are minimising from 0.0006 to 0.0001. Similarly, the result of position model training and validation in Fig. 4.7 shows that the training

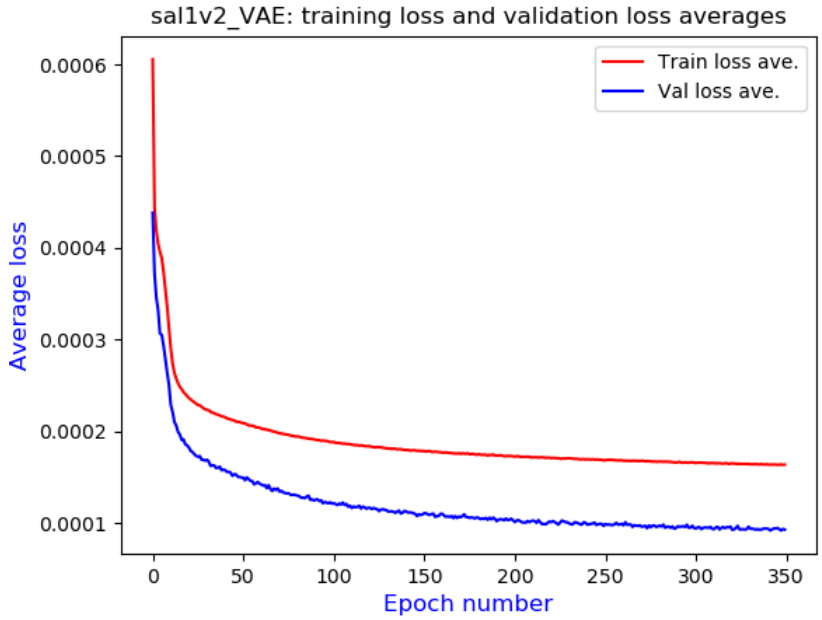

Figure 4.6: Training and validation losses of Vision VAE network.

and the validation average losses are decreasing smoothly over 80 epochs from an average of 0.0024 to 0.0015, which means that the position model is learning to capture the position Self and Env signals.

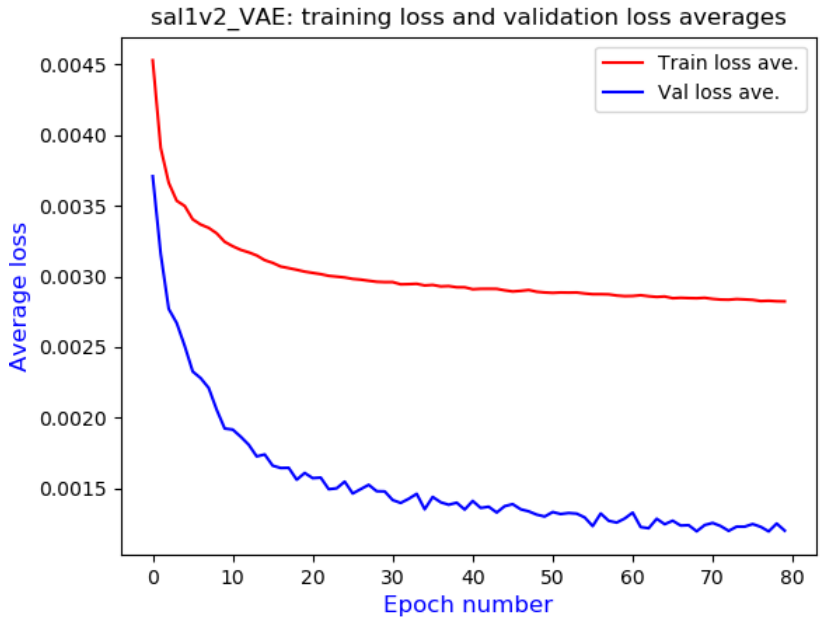

Figure 4.7: Training and validation losses of Position VAE network.

The vision and position networks used classifiers to verify that their networks encoded the required information for the level-1 differentiation. The classifiers use the cross-Entropy loss

function between the predicted latent vectors and the target labels. The classifier networks are trained for 25 epochs and use stochastic gradient descent (SGD) optimisation with a learning rate of 1e-3 and a learning scheduler set to 7 step size. Moreover, the classifier networks are set to use 64 batches with shuffling. The vision and position VAE classifier has fully connected layers of 2 layers and uses ReLU as activation functions.

4.5.5 Level-1 Validation: Vision and Position Networks

The validation results of the vision and position networks are threefold verified: t-Distributed Stochastic Neighbor Embedding (t-SNE) of the distribution [87], Reconstruction ability, and Classification ability.

TSNE

The network encoded latent distribution is visualised by plotting t-Distributed Stochastic Neighbor Embedding (t-SNE). The network encoded latent distribution shows a sparse representation of the validation data. The vision VAE network result shows, in Fig. 4.8, a t-SNE plot with a

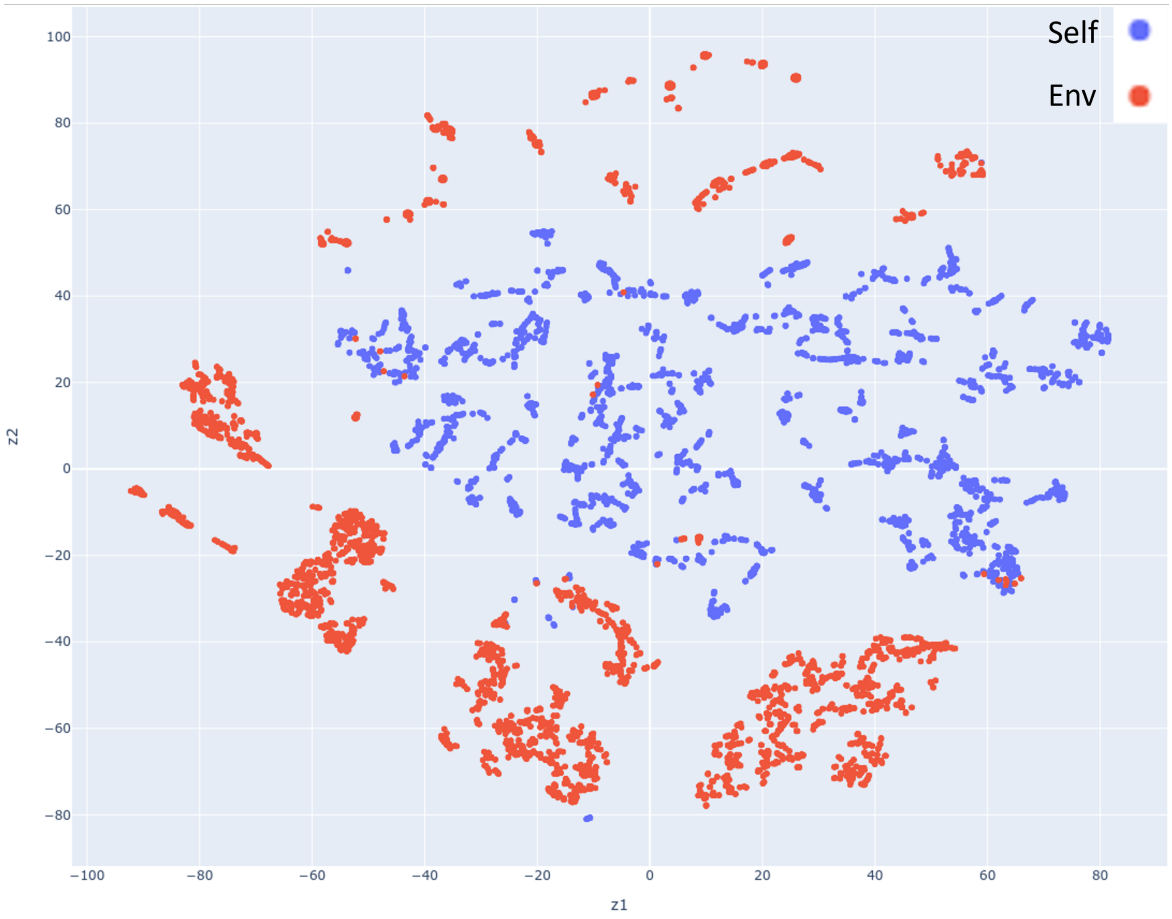

Figure 4.8: t-Distributed Stochastic Neighbor Embedding (t-SNE) of vision VAE network.

continuous separation between the environment and self groups. Similarly, The position VAE

network result shows its ability to learn and sample the self and environment data distribution of the validation dataset. In Fig. 4.9 visualising t-SNE position latent space distribution shows a well continue smooth distribution of self and environment data classes. The continued distribution formed for both vision and position sets ensures that the model sampling can be meaningful for the decoder. Also, continuing to aggregate similar distributions into the same space helps a robot to transition from the distribution point to the neighbour-relevant point of a context.

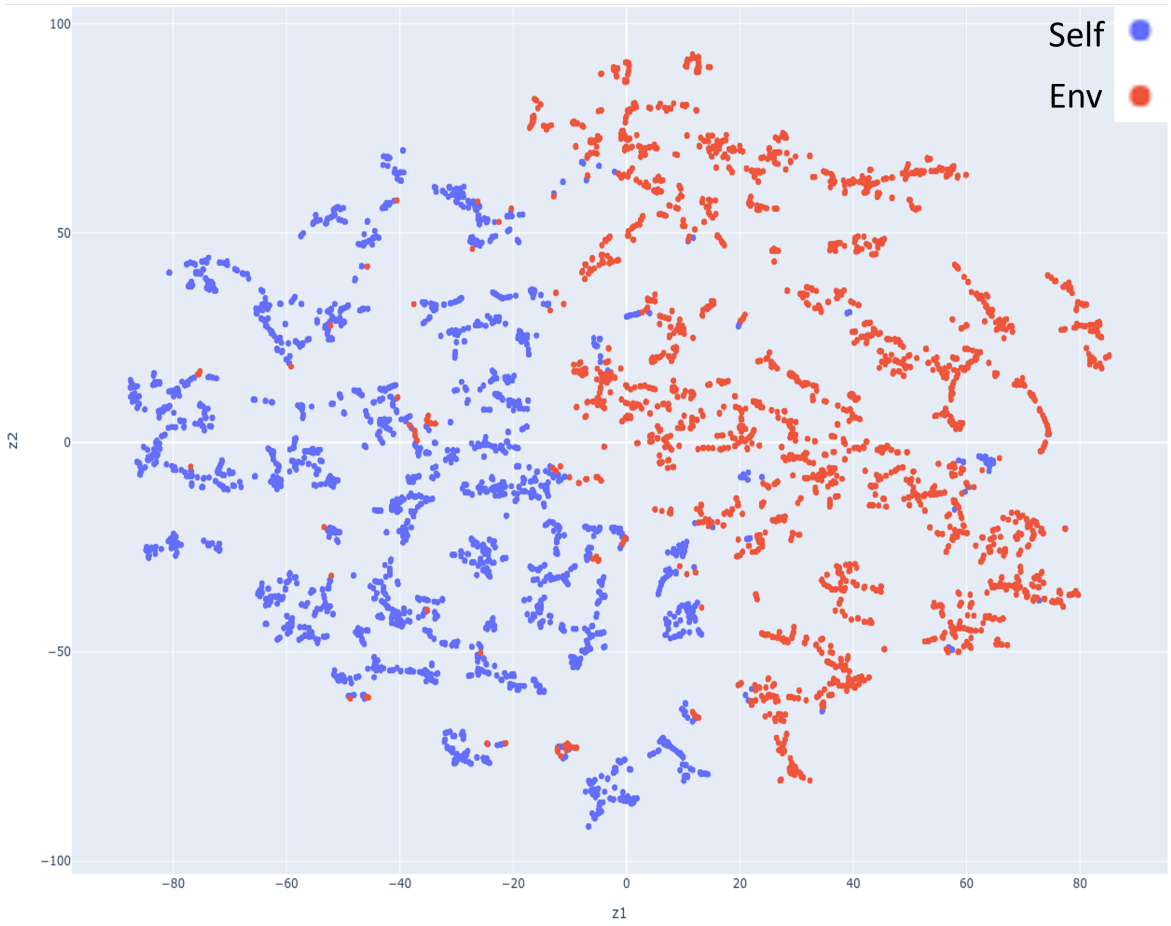

Figure 4.9: t-Distributed Stochastic Neighbor Embedding (t-SNE) of Position VAE network.

Reconstruction

The vision and position of VAE networks were validated by reconstructing random elements from the validation dataset to ensure that the VAE networks acquired the proposed information and produced the expected results. The learned distribution was sampled and reconstructed using the decoder networks. Moreover, Fig. 4.10 shows vision VAE network reconstruction for some environment and self cases. The vision VAE model validation stage result shows its ability to learn to encode and decode the dataset with a minimum 0.00009 validation loss average for the last epoch. Also, Fig. 4.11 shows an example of high relevance between the reconstructed signals of the position network with its original signals. The position VAE model result shows

its ability to learn, sample, and reconstruct the dataset's self and environment distribution with a validation loss of 0.001 for the last epoch.

Figure 4.10: Reconstruction using vision VAE network of random samples from validation data during epoch 322. The top row, represents the original images, and the lower row represents their corresponding reconstructions.

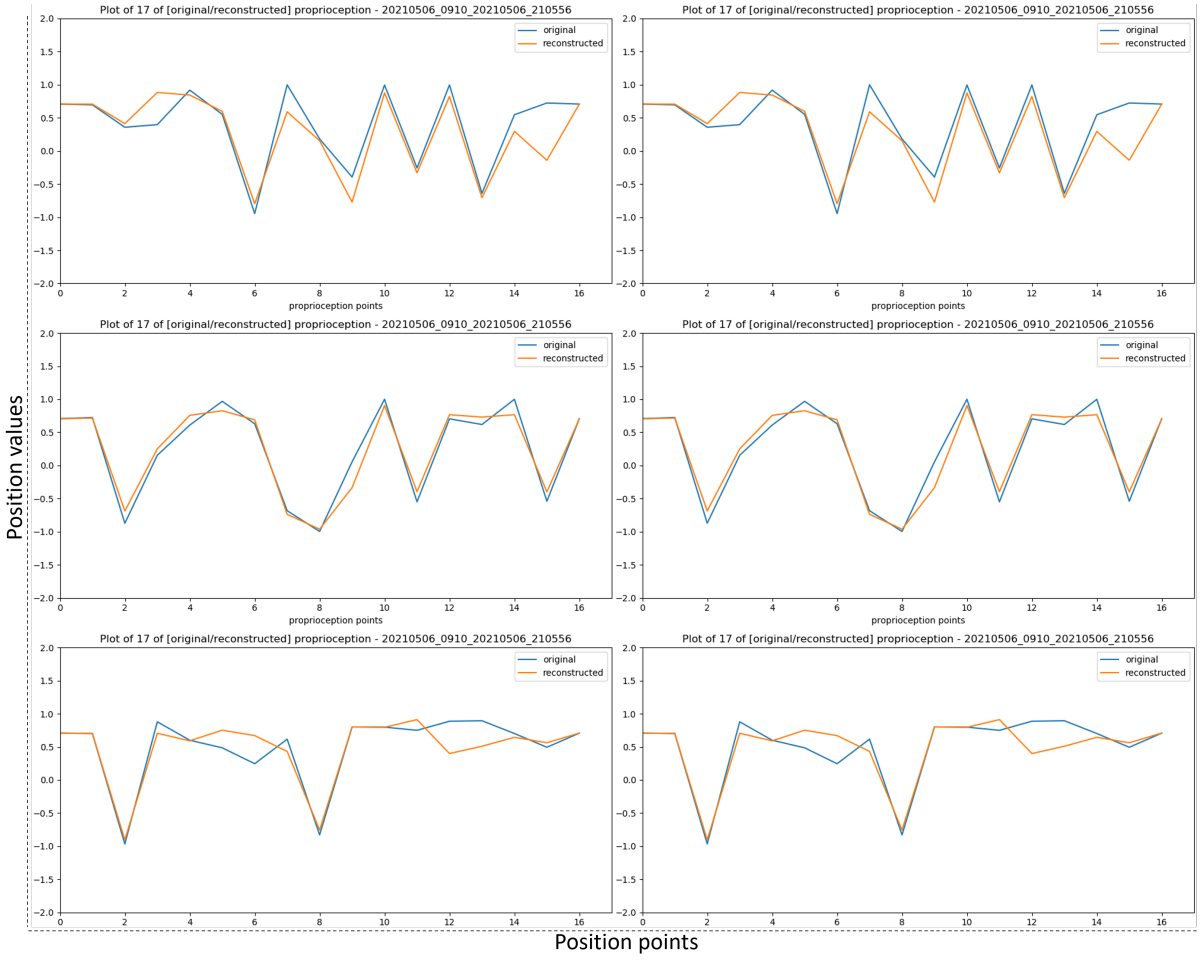

Figure 4.11: Position signals reconstructions using position VAE network of random samples from validation data. The blue lines signals represent the original positions signals, overlaid with the orange lines represents their corresponding reconstructions.

Classification

A supervised classifier is trained and validated to verify the vision and position networks' ability to represent the classes needed, such as self/env. The accuracy results were reported using the eq. defined in Chapter 3, eq. 3.1. The classification result of the vision network achieved an average accuracy of 99.92% of the validation dataset that represented by the true-positive and the true-negative, and only total of 5 elements are misclassified that represented by false-positive and false-negative, as shown in the confusion matrix in Fig. 4.12. Moreover, the classification

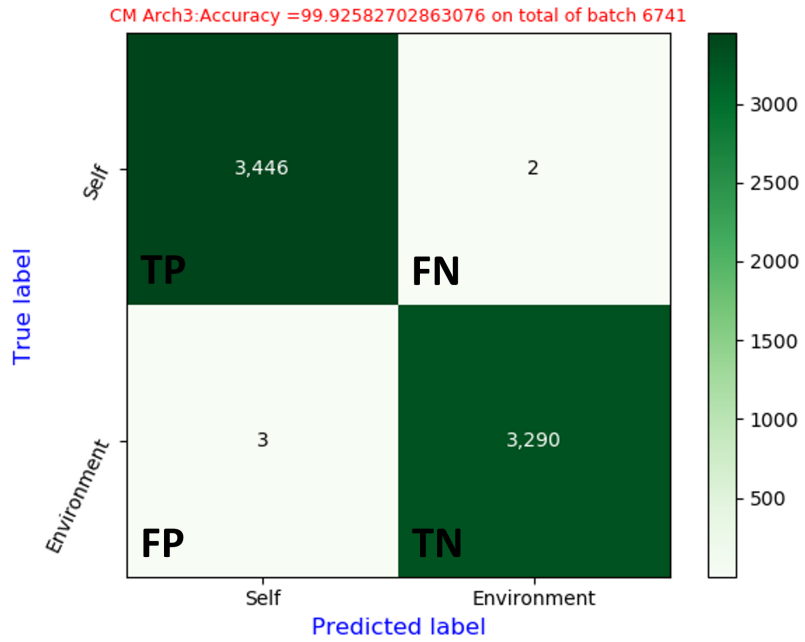

Figure 4.12: Vision VAE network classification achieved a high accuracy using the validation dataset.

result of the position network achieved an average accuracy of 90.5% using eq. 3.1, and only 640 elements of the total of 6,741 elements are misclassified that represented by the false-positive and false-negative, see Fig 4.13. The false-positive elements that predicted as self but they are truly environment elements are 337 that represent 5% of the total elements of 6741, the false-negative that predicted as an environment but they are truly self, which represent 303 elements that is 4.5% of total items population.

The classification result of vision and position networks shows that the networks can accurately classify the data of self and environment. This also shows that both networks captured the required features and distinguished each class which helps in the next stage of fusion of the features of vision and position networks and helps initiate the sense of self by using multimodal data, which helps validates Q1 of the research questions of this chapter.

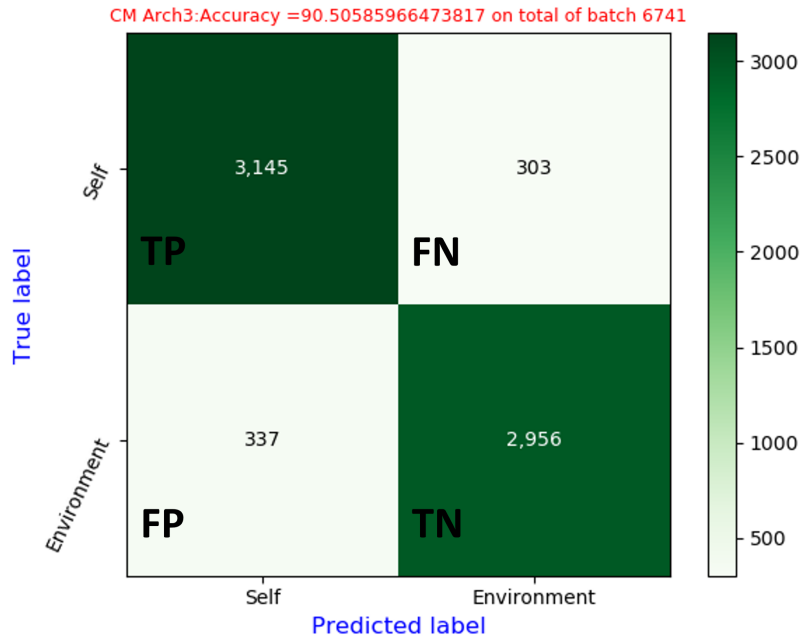

Figure 4.13: Position VAE network classification achieved a high accuracy using the validation dataset.

4.5.6 Level-1 Training: Fusion Network

The fusion AE network is trained using the MSE loss function that measures the squared L2 norm between the fusion decoder output elements and their original elements and backpropagates the level-1 fusion network. The network is trained for 25 epochs and uses the Adam optimisation algorithm with a learning rate of 1e-4 and a learning scheduler set to 7 step size. The model trained with batch set to 64 with shuffling. The vision and position VAEs network's weights loaded and froze in preparing them for the fusion stage. The vision and position VAE's latent space distribution, specifically their mean and variance vectors, are taken to the AE fusion model. The fusion AE network was trained and validated using only the two cases dataset (self/env).

Next, the fusion is fine-tuned by training, validating, and testing the level-1 VAE classifier network of the four cases dataset. The level-1 fusion Classifier is trained based on the loss function of Cross-Entropy Loss between the predicted latent vectors and the target labels. The classifier network is trained for 25 epochs and uses stochastic gradient descent (SGD) optimisation with a learning rate of 1e-3 and a learning scheduler set to 7 step size. The network processes data in 64 batches with shuffling. After good distribution results were achieved from both vision and position networks, the fusion network was trained and evaluated. The fusion model aimed to mix the two input modalities and output a corresponding vector. The result shows that the fusion model is able to mix the vision and position distributions into a fused latent space vector.

The training and validating of the fusion AE network in Fig. 4.14, also shows that the fusion network learns to deal with the concatenated input distribution, the training and the validation

average losses are minimised and converging during the last epochs to 7.3e-7.

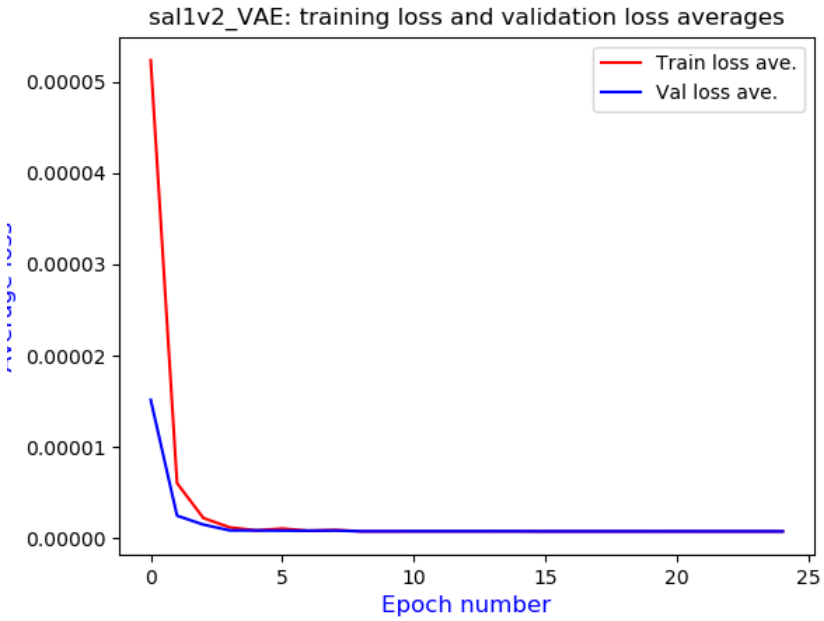

Figure 4.14: Training and validation losses of fusion AE network.

4.5.7 Level-1 Validation: Fusion network

The fusion network was validated by two approaches: its classification ability and its reconstruction of the input modalities to their original states from the fusion latent space.

Classification

The classification result of the fusion network achieved an average accuracy of 98.38% is derived using eq. 3.1, and 1.62%, that 109 elements of the total of 6741 elements are misclassified, that presented by the false-positive and the false-negative elements of the confusion matrix in Fig. 4.15.

The fusion network classifier was trained and tested using an unseen test group by leave-one-out from the framework of the four groups and confounding cases of FC, IL, FT, and FG; the result of each unseen group is plotted in a confusion matrix in Fig. 4.16. The accuracy rate is calculated using the eq. 3.1 for each unseen group's result of the confusion matrix. The accuracy rate result of FC, IL, FT, and FG is 85.09%, 67.7%, 89.6%, and 80.8% respectively, with an average of 80.7% derived by eq. 4.2 overall the total groups.

$$AverageAccuracy = \frac{FC + IL + FT + FG}{n} \tag{4.2}$$

The classification result shows a robot's capability to show a sense of self with 80.7% rate of

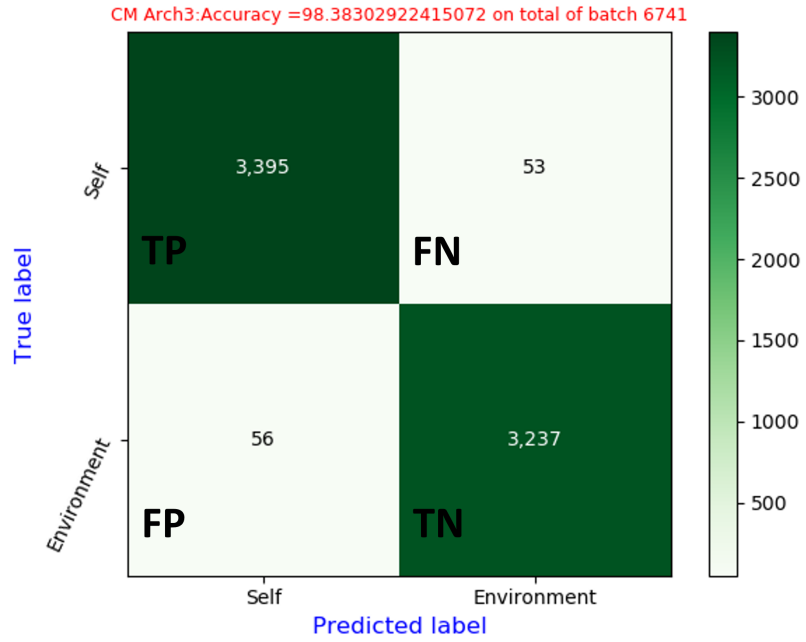

Figure 4.15: Fusion VAE network classification achieved a high accuracy using the validation dataset.

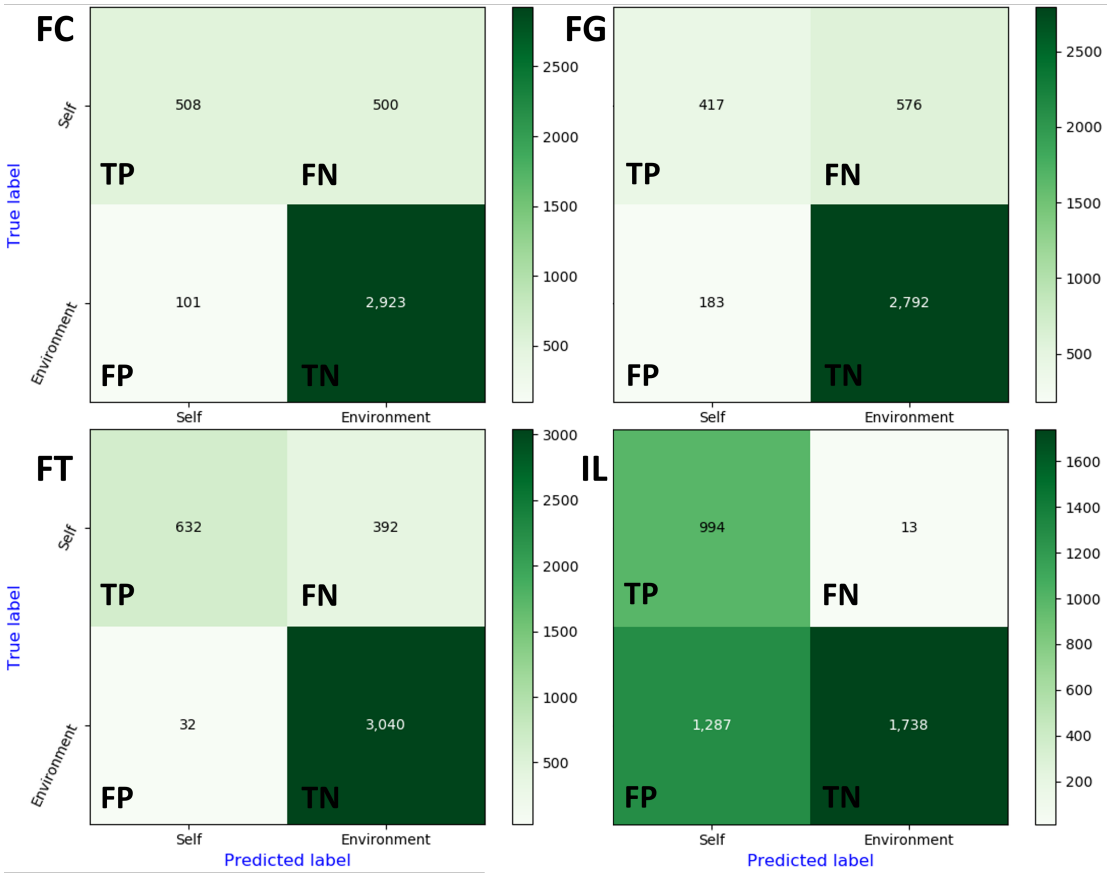

Figure 4.16: The fusion AE network classification was tested on unseen datasets groups of FC: Front Computers, IL: In Lab, FT: Front Towel, and FG: Front Glass.

self-confidence cross-over the experimented groups in artificial self-awareness level-1.

The Front Towel - FT test group's classification accuracy achieves a higher accuracy rate among the four groups. The main difference between the FT and the other groups is the clutters associated with the other three group's environments. The In Lab - IL test group shows the lowest accuracy of 67.7%, and the environment is cluttered (complex). These results indicate that a simple, uncluttered environment helps a robot learn more about itself than a cluttered environment in level-1.

Reconstruction

The Fusion AE networks were validated by reconstructing random elements from the validation dataset to ensure that the fusion network has acquired the input features and produced the expected results. The model's ability to learn mixing modality is evaluated by decoding the fused latent vector with the fusion decoder network, then splitting the reconstructed vector into two parts corresponding to the vision of 1,024 units and the position of 48 units of latent space distributions. Each separated vector is sampled and reconstructed by its corresponding VAE decoder. As a result, the fused data vectors represent their original modalities signals. The reconstruction results in Fig. 4.17 and Fig. 4.18 show the ability of the fusion network to encode and decode the multi-modal features that concatenated prior by the vision and position VAEs.

Figure 4.17: Fusion reconstructions: the result of the fusion decoder gets separated into two tensors. The tensors get reconstructed using their corresponding decoders of vision network decode. The upper row in both images represents the original images, and the lower row in each image represents the reconstructed images.

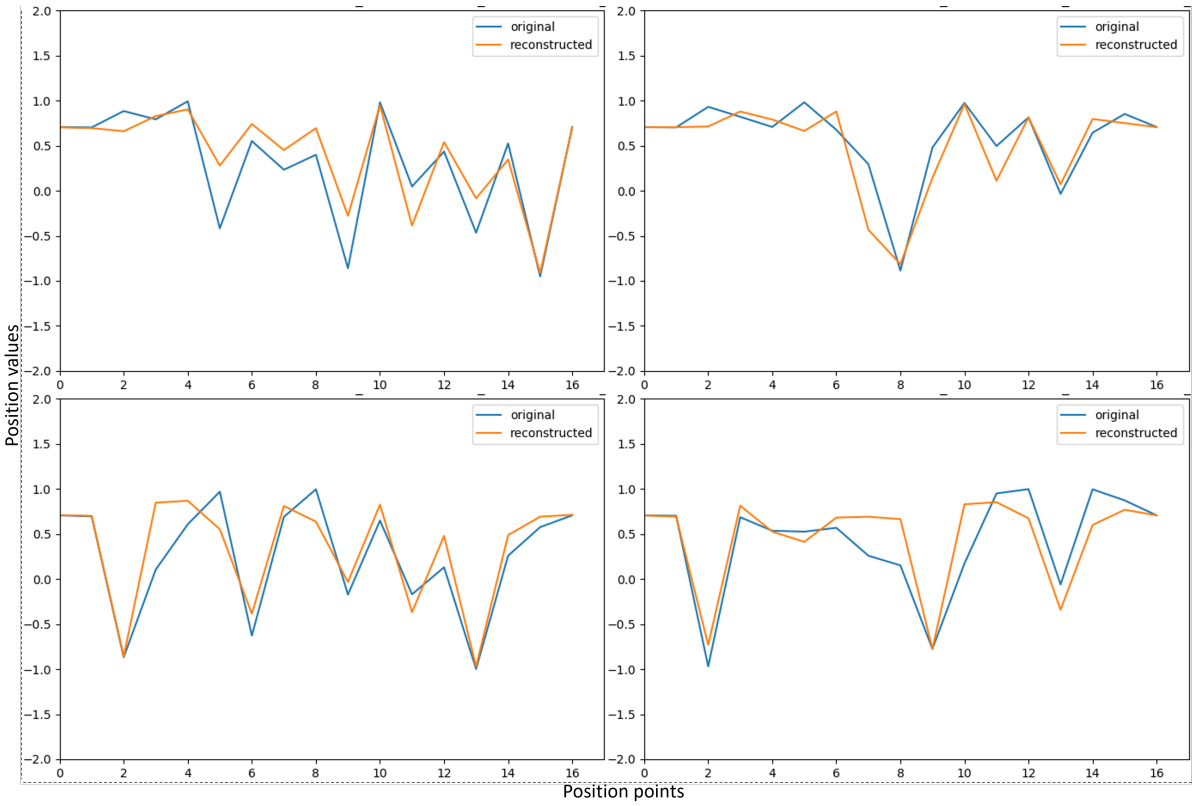

Figure 4.18: Position signals reconstructed using position VAE decoder network after got fused and separated by the fusion network of random samples from validation data. The blue lines signals represent the original positions signals, overlaid with the orange lines represents their corresponding reconstructions.

4.5.8 Outcomes

The Variational Autoencoder is employed for vision and position networks, and the classical Autoencoder is used for the fusion network of the level-1 VAE of self-awareness architecture. This final module is used as the first step toward self-awareness in the next chapter.

This chapter's level-1 architecture design can form a cross-modality latent space from proprioception and vision sensors by the Autoencoder network, as in Fig. 4.5. The level-1 VAE network architecture with the Autoencoders performs well by 80.7% as average over the four experimental groups using unseen test datasets. Moreover, it provides a good reconstruction and is able to fuse the modality into a latent space and reconstruct them back after fusion.

4.6 Discussion

This chapter employs generative models to process and integrates the captured information. The data of proprioception and robot' arms scenes are fused to form a relation between both sensor modalities. The information perceived from senses integration assigns a meaning to the sense of

self. This association is presented by integration that mixes sensors encoded latent space. The integration between proprioception and visual data determines the classifier output, in this case, either a 'self' or 'environment'.

The experiment of level-1 baseline section 4.4 used standard Autoencoder and got an average accuracy of 77.55% over the unseen data groups. This result is considered a baseline and compared with the outcome from this level-1 VAE self-awareness designed based on VAE, which is achieved 80.7% over the same datasets framework. Comparing the results of the baseline level-1 model with the results of the level-1 VAE model, the latter model achieved better in the overall average of the four groups by 3.15%, higher than the level-1 baseline module. This result achieved by the level-1 VAE model answers the first research question in section 4.3.1 of this chapter because of its ability to initiate the sense of self by associating multimodal data sources using the unsupervised learning method.

However, the advantage of the redesigned architecture of the level-1 VAE is providing a stable latent space that represents the extracted features and its ability to reconstruct the sense's multimodality from them, which answers this study's second research question in section 4.3.1 of this chapter.

The VAE is selected over the standard AE because of the following reason: the standard Autoencoder is only able to generate images from the same latent space area. In contrast, the VAE can generate from the discontinuities. This essential feature helps to generate and classify the predicted frame in the next chapter, level-2 of the self-awareness stage. Technically, the existence of the mean and the standard deviation used in VAE allows a decoder to learn from a variation of the distribution of the same input. Thus, the predicted frame in the level-2 model might fall within that distribution in which the decoder will be able to reconstruct and classify it.

4.7 Conclusion

The proposed artificial self-awareness level-1 outcomes show that a robot gets a sense of self; an initial self is represented by differentiation defined in the robot. The differentiating level-1 of artificial self-awareness uses real data sets of different environments represented by different groups. Also, the model was verified by the confounding signal of the unseen data group. Furthermore, the artificial self-awareness level-1 allows the robot to differentiate itself by 80.7% of accuracy. Also, it allows the multimodel data of a robot to get interpretable and representable, which helps consolidate the development of the next level of artificial self-awareness.

This chapter verifies one of the hypotheses of this thesis:

- A robot's self-awareness sense of self can be enabled using an unsupervised learning method to associate and reconstruct the fused real multimodal data.

The differentiation of artificial self-awareness level-1 is defined in the robot, but it cannot locate its limbs within the environment and put them into context for a task. The next chapter's work comprises developing level-2 (Situation; chapter 5) of artificial self-awareness. The idea is to employ temporal sequences of the robot's arms and model visual and proprioception experiences in a recurrent network architecture, which we believe is the next step to let a robot identify itself with higher self-certainty in an environment.

Chapter 5

Artificial Self-Awareness Level-2

The level-1 of artificial self-awareness can recognise the robot from a single snapshot of vision and position data inputs. However, the robot cannot situate itself by stating itself dynamically within a scene. This chapter focuses on defining level-2 of artificial self-awareness, an additional module architecture proposed by which the robot can process temporal data and situate itself. The difference between level-1 and level-2 of artificial self-awareness is that level-2 gives a robot more self-certainty about itself. This chapter follows the modular approach reported in Chapter 1, and level-2 proposed in this chapter is an added module to level-1 of artificial self-awareness.

In this chapter, a robotic self-awareness level-2 model that incorporates sensing of visual and proprioception (position) features is developed to provide a robot's artificial self-awareness of the situated self. This chapter provides the technical details of the level-2 architecture and compares the accuracy between four environmental groups of confounding cases in terms of classification accuracy and reconstruction. This work provided a robot with a situated self as inspired by Rochat's [1], utilising a robot's multimodal sensory input to build relationships that define the situated self in a robot. The experiments indicate that a robot can get situated self by differentiating and predicting itself from the environment in a temporal manner.

5.1 Introduction

The proposed approach for self-awareness research study starts incrementally by defining a sense of self (level-1) and extending the situated self (level-2).

This chapter focuses on self-awareness level-2 that incorporates the output from level-1 of self-awareness to provide a situated self. The output from level-1 contains the fused latent vector of vision and position data. The approach integrates previously reported level-1 of getting a sense of self and extends it to allow a robot to increase its self-confidence by perceiving itself over time.

The level-2 hypothesis of this thesis reviewed in Chapter 1, section 1.6 is revisited to be assessed throughout this chapter:

- **A temporal perception of the robot's dynamic movement increases the self-certainty of the robot in a different environment if the predicted reconstructed signals for vision and proprioception are statistically significant and classification accuracy is above the level-1 for the same dataset.**

In Section 5.2, the motivation and objectives of this work are given. Section 5.3 presents this work's artificial self-awareness level-2 approach, and Section 5.4 details the level-2 experiment and the results. The experimental results discussion is given in Section 5.5. Section 5.6 presents the level-2 statistical analysis. Section 5.8 concludes this work.

5.2 Motivation and Objectives

The motivation of this chapter is to increase the self-certainty observed during level-1 of the self-awareness model. This chapter proposed to increase self-recognition through the means of developing a predicted self over a temporal state. Level-1 (Differentiation) and level-2 (situation) of Rochat [1] inspire the artificial self-awareness framework architecture design. From a robot perspective, the level-1 of artificial self-awareness proposed in chapter 4 is framed as a robot being able to get a high-level description of its limbs via a fused association of its kinematics and its vision by getting a snapshot of its fused association state. In contrast, for this chapter, level-2 of artificial self-awareness is framed as a robot learning a temporal connection between the current and the predictive fused association state by processing a series of observations. The fusion latent vector is output from level-1, representing the experience of felt and seen relation. Level-2 offers to advance the fusion relation by preprocessing the current states and predicting the next future state. The predicted state is validated with the available original state of the dataset. The fusion relations preprocessing are occurred by a recurrent neural network, where the temporal dynamic behaviour between the fusion states is captured.

The motivation for several window sizes explored in this chapter was to find the optimal window size a robot can be processed to temporarily develop a better robot self prediction using LSTM. In level-2 of artificial self-awareness, the model depends on predicting the next state (future state) based on processing several prior states. The validity and accuracy of the next predicted state come from the model's ability to learn from processing previous sequence states. Moreover, because the level-2 model uses LSTM and no common window size is recommended, processing lower to larger window sizes is good practice to reach an optimal point that can be selected and to be used for the artificial self-awareness framework.

Considering an advanced level of self-awareness such as level-2 gives a robot the ability to go beyond getting a sense of self by allowing it to utilise its higher self-confidence and integrate it to control its task. The key contributions of this chapter are two-fold: Firstly, it is the second level of artificial self-awareness piece of work to adapt self over time to enhance a robotic task. Secondly, the proposed level-2 approach is applied to the framework of different

environments with the confounding case and demonstrated to show substantially improved robot self-recognition (higher self-certainty).

5.3 Artificial Self-Awareness Level-2

The design approach of the artificial self-awareness network architecture framework is modular and combines level-1 and level-2, allowing an easy expansion of future levels and allowing model interpretation and validation. Therefore, the level-1 inputs of the artificial self-awareness framework are fed by two sensory sources: the robot proprioception represented by the angular position, where the robot arms location in the scene is important to comply with what the robot visually perceives, and the robot vision is where a robot sees its arm's in/out of the scene. Both data inputs get fused to form an association that is used as an input of level-2, a snapshot of the current association represents the sense of self for the robot. After that, level-2 uses the fused association as input and learns to generate the next fused association state. The next fused association represents the situation of what the robot intrinsically believes of itself as situated.

The approach in the level-2 model focuses on building the situated self in a robot by letting a robot obtain a temporal perception of its dynamic movement using its proprioception and vision data sensing modalities. For this, a Deep Neural Network architecture is designed for the level-2 model architecture (Fig. 5.1) to support the situated self in the robot. The situated self represents the second self-awareness level of Rochat's five self-awareness development levels discussed in Chapter 2, section 1.1. Therefore, the second level of self-awareness inspires this chapter's model architecture design, and therefore, the proposed level-2 model can be mapped for robots as follows:

- Level-2:" Situation" is the level at which a robot situates itself in a scene by obtaining a temporal perception of its dynamic movement using a position and vision data sensing modalities.

Because the recurrent methods can process temporal data problems, the LSTM type of recurrent DNN is used for artificial level-2 implementation. The LSTM in level-2 processes the number of input timesteps by unfolding the LSTM cells. Each cell processes an input timestep vector, including the previous output and the previous hidden state of information.

5.3.1 Level-2 Model

In this artificial self-awareness level, a robot situates itself by observing and confirming the temporal motion perception, represented by its sensory changes within a scene. In level-2 architecture, a recurrent learning model is implemented, such as Long Short-Term Memory (LSTM) network. The LSTM network gets a current state of features and learns the changes in time over the next state of features. Therefore, in Fig. 5.1, the LSTM inputs a series of the level-1 features

of 512 units and outputs the next feature state of 512 units using a single LSTM layer. Different memory cells of 2-10 window sizes are investigated to examine the appropriate number of states level-2 architecture can encode and the network's capability to encode its awareness of the robot's arms through position and vision signals relations. The level-2 is used the same level-1 classifier network of 5 fully connected layers that input 512 units and output 2 units to classify the LSTM prediction state of self or environment.

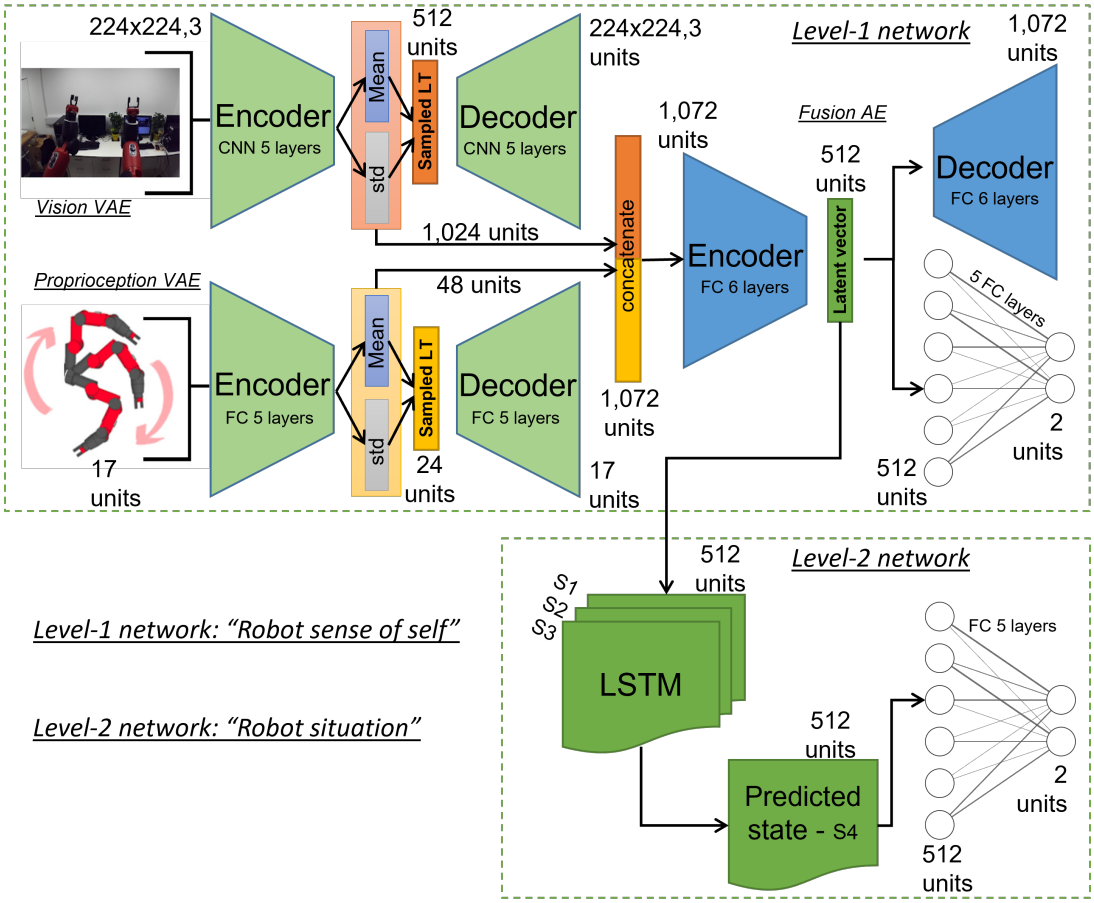

Figure 5.1: The artificial self-awareness level-2 architecture was built based on LSTM as a recurrent network. The classifier network checks the LSTM network's ability to generate the next state features by classifying the generated state of self or env. The above fully connected layers networks in level-1 and level-2 are the same.

5.3.2 Level-2 LSTM

The Long Short Term Memory (LSTM) [53] [54] is a type of Recurrent Neural Networks (RNN) is used to process time series data or sequential data as described in Chapter 2, section 2.1.7. This type of network architecture is used for temporal problems. It memorises the previous input information, which participates in associating the next input, in its architecture, decides which information a timestep should retain, and presents an attention mechanism to make the network more efficient.

The artificial self-awareness level-2 implements the LSTM to process several input timesteps by unfolding the LSTM cells. Each cell processes an input of timestep frame vector and processes previous cell output and hidden state of information. For example, in Fig 5.2 the unrolled LSTM processing features of 3 timesteps, the hidden layer vectors c_{t-1} and h_{t-1} in the first cell only they are initialised by zeros. The unfolded LSTM processed each timestep as input x_n and predicted the next state of the timestep. Level-2 depends on the final predicted state, which is used to validate the target state.

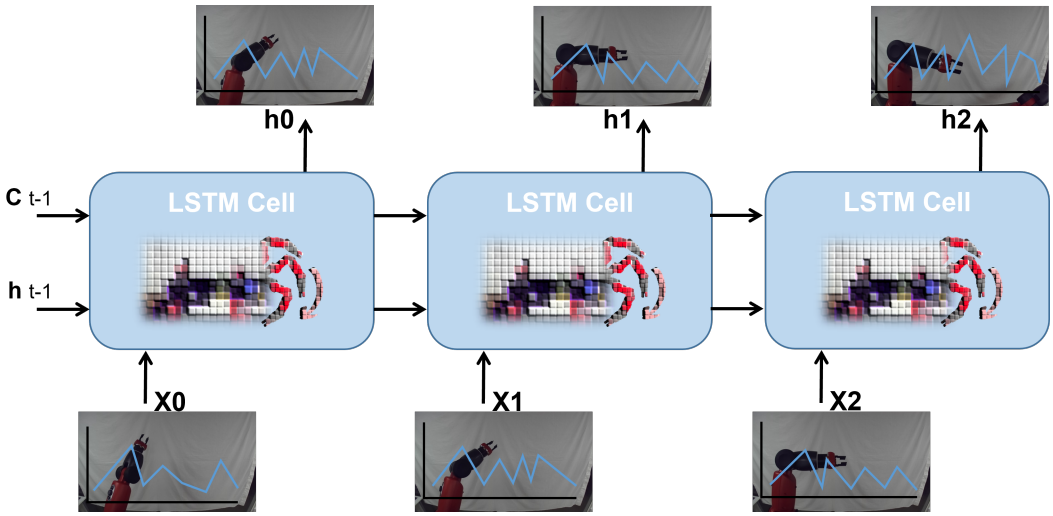

Figure 5.2: The LSTM cells are unfolded based on the sample input size of multimodal information passed by level-1 of the artificial self-awareness model.

5.3.3 Level-2 Dataset

In Chapter 4, section 3.4.3, the level-1 dataset is prepared to validate the artificial self-awareness framework research consisting of real robot arms positions and scenes from the egocentric robot view. Moreover, the position and vision of the dataset contain two robot scenarios, robot arms in and out of the field of view. The dataset captured for both modalities is 30k, divided into 80:20 for the training and validation stages, respectively. As described in Chapter 4, section 3.4.3, the dataset was captured in four groups of different environment settings, where each group represents a different environmental complexity that ranges from simple (front towel and front glass) to cluttered (*front computers* and *in lab*) environments. In addition, each group was distributed into four confounding cases, where each case consisted of a different input combination of signals of position and vision. While a combination of three environment settings was used for the training, leaving one out representing an unseen environment set for testing purposes.

In addition to the above dataset preparation, for level-2, the same dataset was prepared into samples of timesteps to be processed by the level-2 LSTM, where each group is transformed into samples of 12 timesteps capable of accommodating the window sizes of 2-10 considered in this study. These samples were processed by different experiments using different window

frame sizes for training and validation sets of the level-2 module. Fig 5.3 shows an example of 3 samples of 12 timesteps of the self class that correspond to the Front Towel group scene. The arm/arms have a slight move in each timestep, representing video frames associated with their position values. The samples are formed by shifting one timestep in comparing the samples with each other.

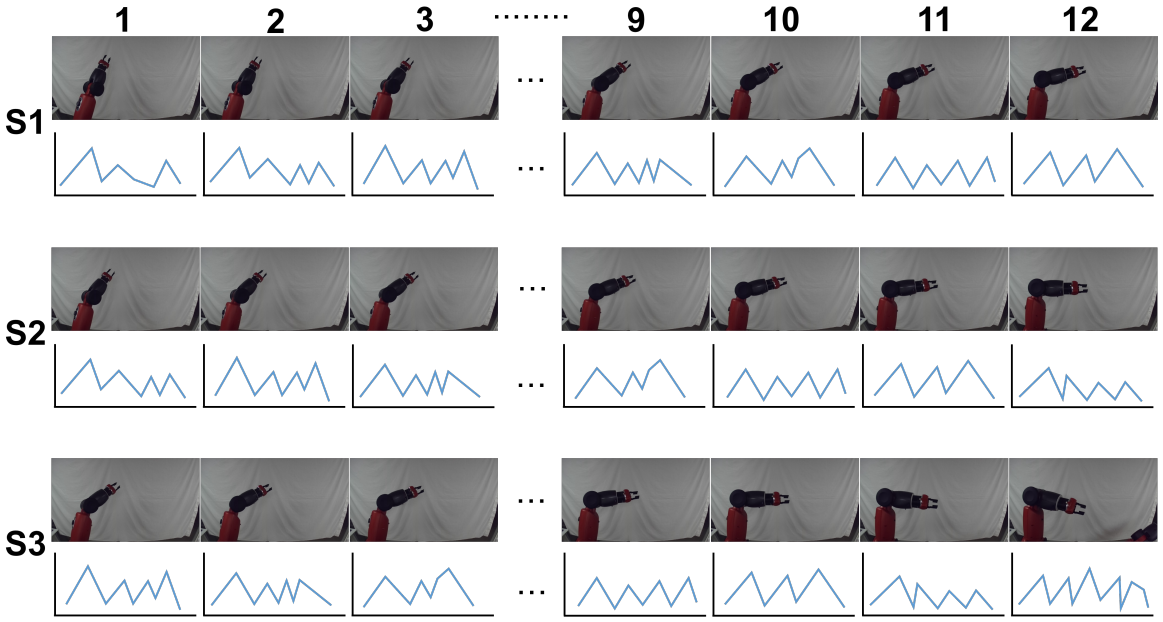

Figure 5.3: The diagram shows three samples, each sample (Sx) represents a timesteps of twelves frames.

A different window frame size is assigned to allow different dataset combinations based on the number of timesteps per sample. The number of different window frame sizes are utilised for the level-2 experiments. Fig 5.4 shows an example of a window size of 6 frames and its target. The window sizes are applied to experiment with artificial self-awareness level-2, such as 2, 3, 4, 5, 6, 8, and 10. For each window frame size, $windowframe + 1$ is equal to the target frame used to validate the LSTM level-2 predicted frame.

To train and validate the level-2 network, two datasets are used: the first set consists of Group-1, Group-2, Group-3, and Group-4 of samples, that contain only Two Cases (Case-1 and Case-2), and the second set represents Group-1, Group-2, Group-3, and Group-4 of samples of All Cases with leave-one out as unseen group.

5.4 Level-2 Experiments

The experiments of level-2 are divided into two main stages, the first phase is to train the level-2 LSTM network, and the second is to train the Level-2 Classifier network. Moreover, the experiments are conducted to process different window frame sizes of 2-10 encoded fusion vectors. The Level-2 LSTM experiments involved the samples of dataset framework with only

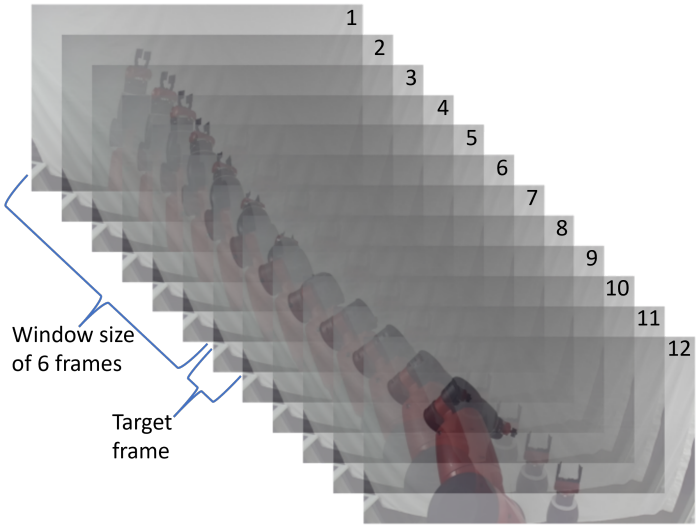

Figure 5.4: Sample of twelve timesteps of image frames from the level-2 dataset. An example of window size of 6 is applied, which results in 6 frames timestep, and the 7th represents a target frame used for validation.

Case-1 and Case-2. The Level-2 Classifier experiments involved the samples of the dataset framework of the groups and their unseen groups with the four confounding cases.

This experiment of artificial self-awareness level-2 answered the following research questions:

- **RQ1**: Is a robot able to relate its arm movements with its body?
- **RQ2**: Is a robot able to build an inter-modal link between what is seen and felt?

5.4.1 Level-2 LSTM Network

Train And Validate LSTM Network

Train level-2 LSTM is based on the MSE loss function that measures the squared L2 norm between the predicted state elements and the target LSTM state elements and backpropagates the level-2 LSTM network. The network is trained for 25 epochs, enough for the network to converge the trends during the training and evaluation, and uses the Adam optimisation algorithm with a learning rate of 1e-5 that lets the network learn in small steps. Moreover, process data of 64 batches with shuffling that can be accommodated with the GPU of 8GB of memory used in this study. The training of level-2 LSTM starts by loading level-1 models and freezing their weights, including vision, position, and fusion networks. In level-2, the LSTM is used to encode a timestep of features extracted by level-1. The level-2 dataset consists of samples of 12 timesteps sliced in different sizes based on the required window size and uses the LSTM to encode them. The LSTM network was trained and validated using the framework of four groups dataset of two confounding cases (Case-1 and Case-2) of samples that were processed

and passed by level-1. The LSTM is trained several times, each with different window size ranges of 2-10 samples, specifically 2, 3, 4, 5, 6, 8, and 10. Fig. 5.5 shows the training and validation trends of average losses of the LSTM experiments over the 25 epochs. The diagrams show that there are no significant differences between the experiment's average trends, and that is due to the dataset used, which is the same overall in the experiments; the only difference in each experiment is that a different window size gets processed. After each experiment, the LSTM network's weights state is saved after 25 training epochs.

The training MSE loss averages are shown in table 5.1 for the LSTM experiments for the window size of 2, 3, 4, 5, 6, 8, and 10. The results show that the level-2 model is learning from the provided timesteps. The converges between the training loss and the validation loss for all experiments are mainly between the third and the fourth epochs. The experiments that processed a higher window size show a slightly faster converging as in the experiment of the 8-window and 10-window timesteps, where both validation trends at the third epoch achieved 1.18e-3 and 1.20e-3, respectively. The 10-window achieved the lowest validation average of 1.60e-4 at the final epoch compared with the 2-window of 1.87e-4.

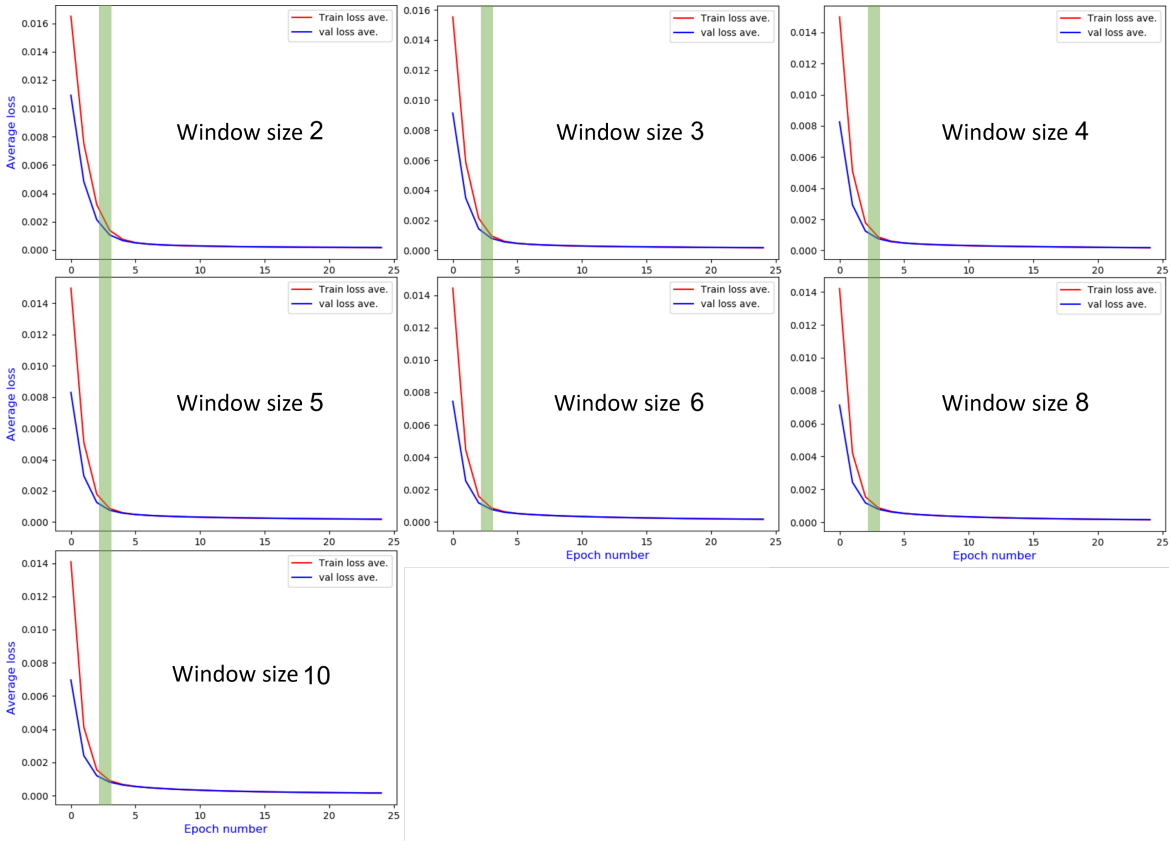

Figure 5.5: The plot lines represent the average training loss (red) and validation loss(blue) over 25 epochs. Each was trained and validated using different timestep sizes. The numbers on each diagram show the number of windows timestep size used for the trained LSTM network. The

The validation MSE loss averages for the LSTM experiments for the window size of 2, 3, 4, 5, 6, 8, and 10 are shown in Table 5.1. The validation MSEs show a gradual enhancement in

MSE loss accuracy while processing a fusion sample with a bigger window size of timesteps. The experiments' training and validation losses are added in Table 5.1 where both MSE averages of training and validation demonstrate that level-2 LSTM processed with larger window sizes yields better MSE loss accuracy result. Moreover, The reconstruction MSE average of each experiment is listed in table 5.1 to show the reconstruction loss associated with each experiment and validate the level-2 of the self-awareness model. Furthermore, the reconstruction MSE averages are visualised in a bar chart in Fig. 5.6 for easy comparison between its reconstruction values.

Table 5.1: The Mean Square Error (MSEs) averages were calculated for training, validation, and reconstruction losses of Level-2 LSTM experiments on seven different sample sizes, such as 2, 3, 4, 5, 6, 8, and 10.

Window size	Training MSE Average	Validation MSE Average	Recons. MSE Average
2	1.75e-4	1.86e-4	3.81e-3
3	1.82e-4	1.89e-4	4.29e-3
4	1.77e-4	1.83e-4	3.23e-3
5	1.69e-4	1.75e-4	3.07e-3
6	1.62e-4	1.68e-4	2.51e-3
8	1.55e-4	1.61e-4	2.23e-3
10	1.53e-4	1.60e-4	1.03e-2

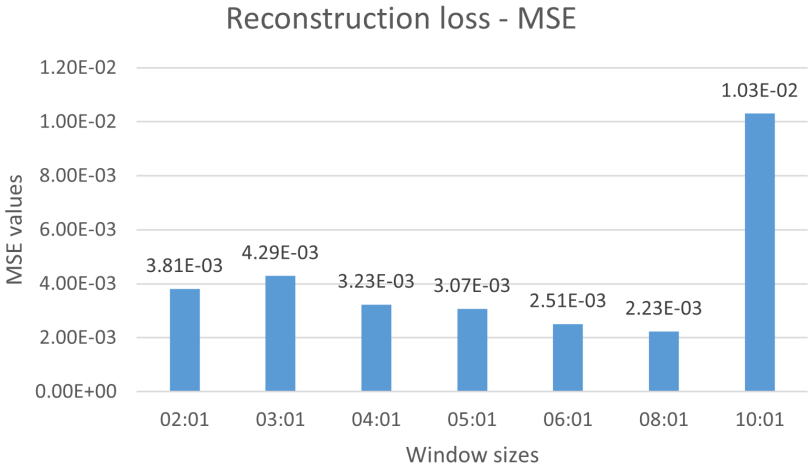

Figure 5.6: In the visualised bar chart for the reconstruction MSE averages, each bar represents a specific window size associated with its reconstruction MSE average result.

Fusion Latent Space Reconstruction

The predicted state of the LSTM represents a future state of fusion sensory of a robot, consisting of 512 units. The predicted state is investigated to validate further the level-2 LSTM network's learning and prediction ability of a future state and ensure that the LSTM network's predicted state is valid and meaningful within the level-2 context. For that, the predicted fusion state for the best experiment is reconstructed. In Figure 5.7, signals from the experiment that used the 8-window timestep is reconstructed, the 8-window timestep selected because it achieves the lowest MSE average as shows in 5.1 and visualised in Fig. 5.6, which means less variation between the original and the predicted signals.. The predicted fusion state signal (orange trend line) and its original target signal (blue trend line) are plotted. Comparing the plots of both fusion signals show that the level-2 LSTM network's predicted state is valid, i.e. the LSTM model has captured and can predict the next fusion state, as both the predicted and the target signals are almost identical. The orange trend line of the predicted fusion state covers and is aligned with the signal changes of the blue trend line of the original signal.

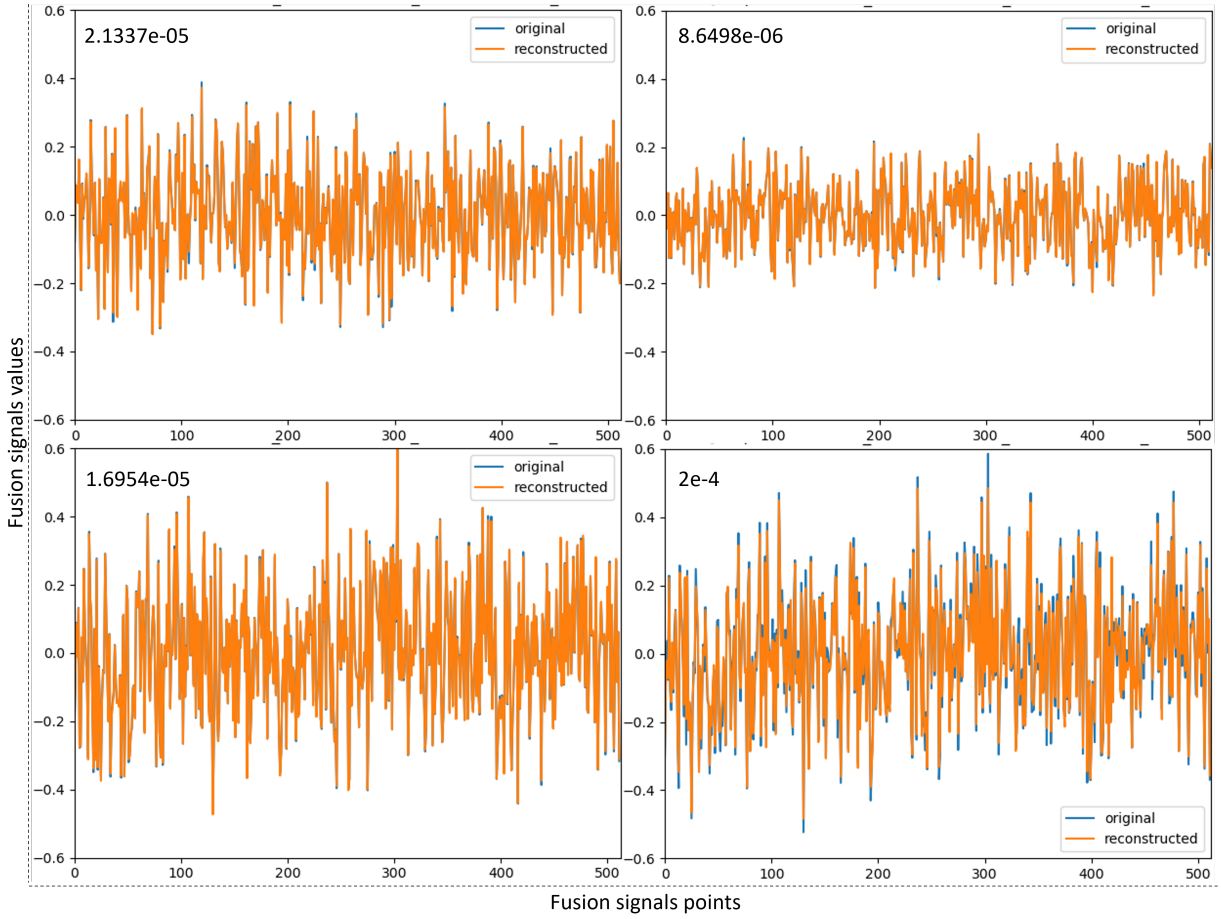

Figure 5.7: The target and predicted state signals of the fusion network latent vector of 512 points are plotted, showing that the predicted states are aligned with the target states.

The output of the level-2 LSTM represents the next time step of the fusion latent space

vector. The seven experiments were conducted using the same dataset, but different window timestep sizes are used to predict the next time step of the fusion latent space vector. The reason for conducting different window timestep sizes in different experiments is to find an optimal timestep size that level-2 LSTM can encode based on the used datasets. The reconstruction loss is calculated by MSE between the predicted state and the target state of the fusion latent space vector. The level-2 LSTM network is validated by the first batch of 64 samples of the validation dataset proportion. The results of the calculated reconstruction MSE averages are presented in Table 5.1. It can be deduced from this table that the MSE average of the fusion latent state vector reconstruction loss is better when considering a large-size window timestep. The table starts from the 2-window timestep average of 3.81e-3, ending with the 8-window timestep average of 2.23e-3, except for the 10-window timestep, where the reconstruction MSE average starts to become worse by 1.03e-2.

The predicted fusion vector is decoded by the level-1 fusion decoder, producing the anticipated concatenated vision and position features. These features are split based on the sizes of vision and position vector. Each vector is reconstructed using its corresponding decoder of level-1 to validate the level-2 LSTM visually and monitor its capability, representing one of the artificial self-awareness architecture features that comply with interpretability. The reconstruction results for both position and vision networks show a reconstruction match between the predicted and the target states.

Vision Frames Reconstruction

The level-1 Fusion AE network decodes the predicted level-2 LSTM fusion Latent space vector; this results in a vision and position concatenated features vector. The resulting features vector are split into two features, each processed by its model. Thus, the vision and position decoders of the level-1 network are used to visualise their predicted vectors by reconstructing both position and vision sensor inputs.

In detail, for the vision part, the level-1 vision decoder is consulted to produce the reconstruction of the corresponding vision vector. The vision part feature contains 1024 concatenated units of 512 Mean and 512 Standard Deviation units sampled and decoded by the level-1 vision model decoder. The first batch of 64 samples of the validation dataset is used for the reconstruction check. The reconstructed results of random samples of the 8-window size by the vision decoder network in Fig.5.8 show a successful reconstruction results match between the target and the predicted states, the above row of images represents the reconstructed images by the level-1 decoder, and the second row represents the reconstructed images by the level-1 decoder, but after level-2 LSTM process and separation. The target and predicted reconstructed images in Fig. 5.8 are almost matched in comparing them; the predicted reconstructed images have no major features lost. In addition, some reconstructed image in the second row shows better feature representation than the targeted image. For example, the rightmost reconstructed im-

Figure 5.8: The first row represents the target image frames, and the predicted image frames are represented by the second row, and show that the predicted image frames are almost close to the targeted image frames. Larger sample images are shown in Appendix C.

age shows a better arm representation than the one represented by level-1; this demonstrates that level-2 participated in representing the features from the probability distribution. This also helps the classifier to classify the reconstructed features.

Additional samples from the first of the 64 samples of the validation dataset are reconstructed, specifically with the number of window timesteps listed in Table 5.1. The reconstruction was generated on different timestep window frame sizes for the same dataset, Fig 5.9 and 5.10 show reconstructed examples of the targets and the predicted states for each window timestep size.

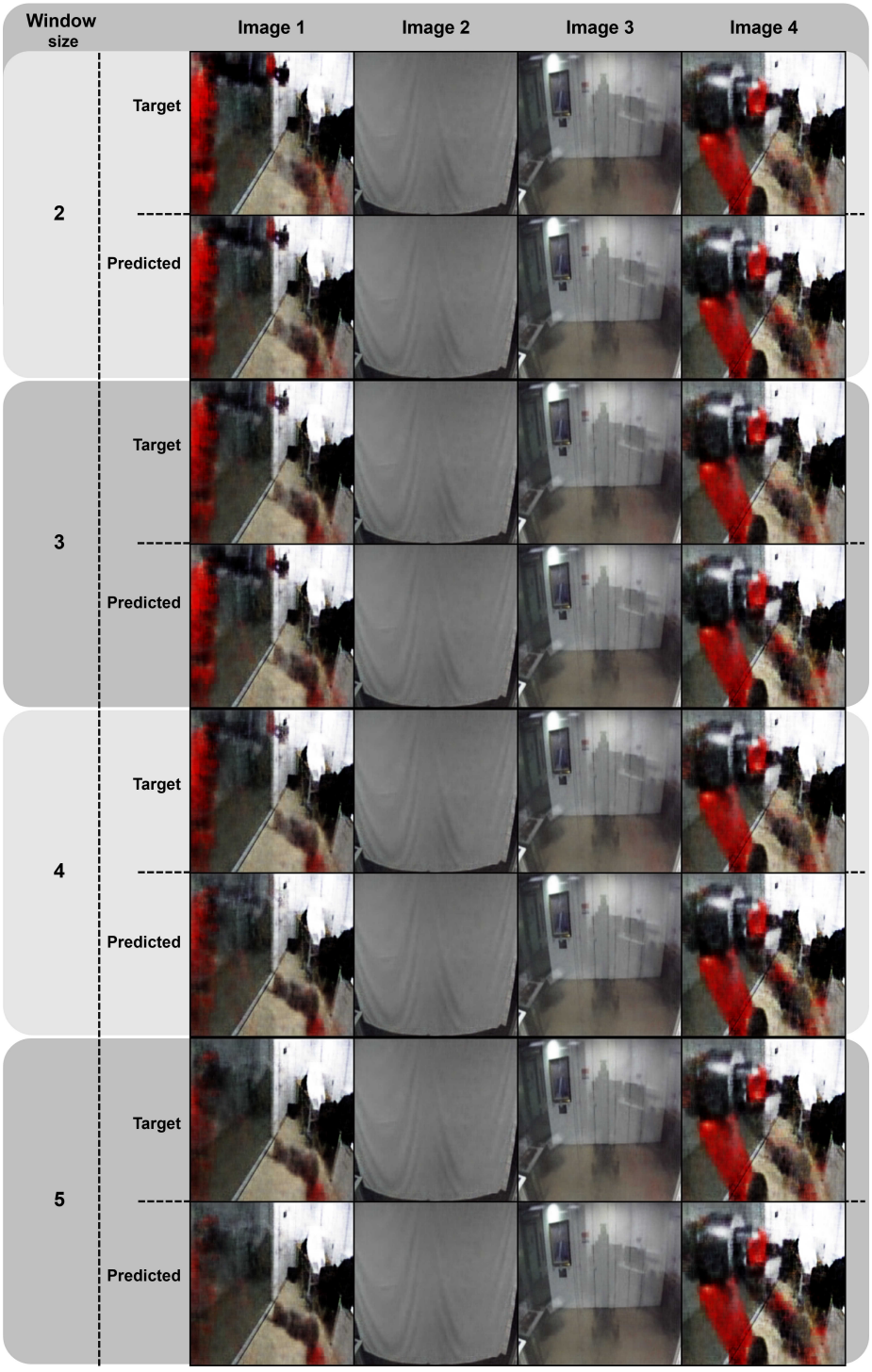

Figure 5.9: Reconstructed examples of four image frames were processed by level-2 LSTM and decoded by level-1 vision decoder on different window sizes. The first row for each window size represents the target frames, and the second row represents their predicted frames.

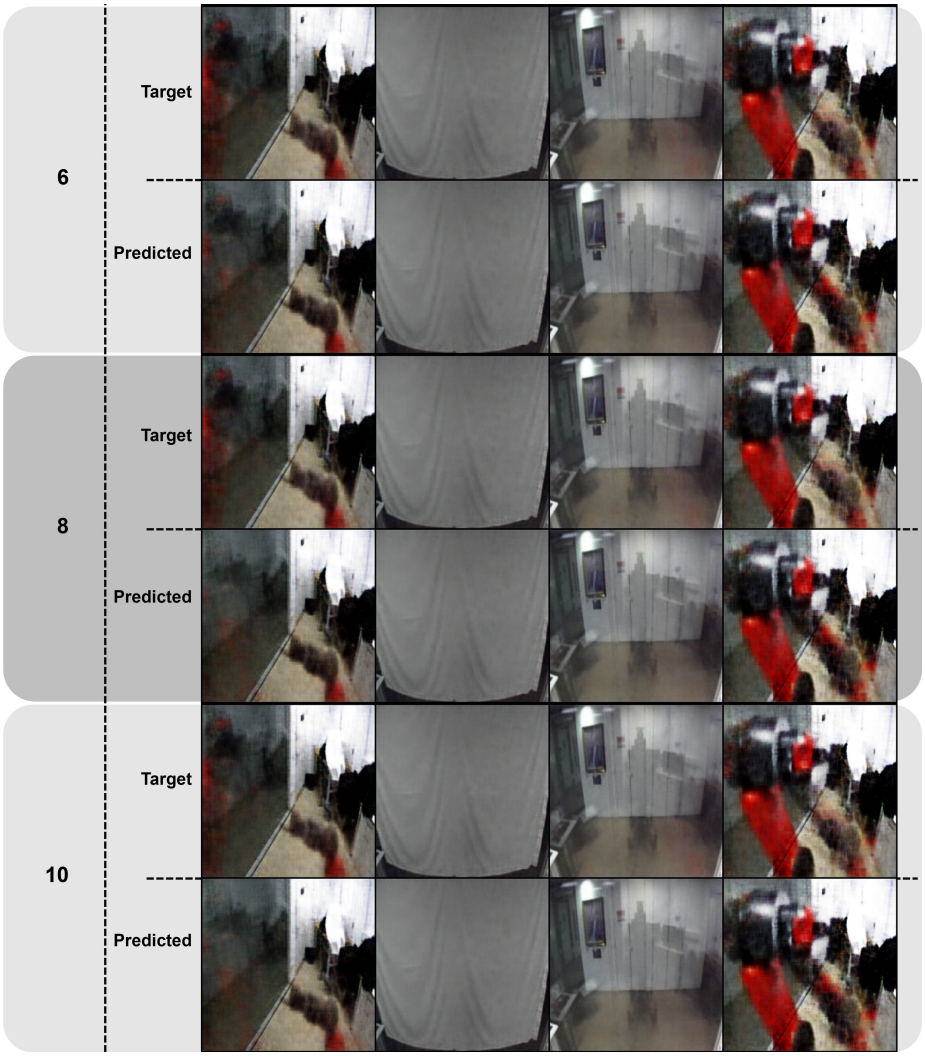

Figure 5.10: Reconstructed examples of four image frames were processed by level-2 LSTM and decoded by level-1 vision decoder on different window sizes. The first row for each window size represents the target frames, and the second represents their predicted frames.

Position Signals Reconstruction

Similarly, the position decoder of level-1 is used to reconstruct the position state signals. A vector of 512 units of the predicted fusion state vector got split into vision and position vectors. The position feature vector consists of 48 concatenated units of 24 Mean and 24 Standard Deviation sampled and decoded by the level-1 position's decoder network. The position signals of the 8-window size are reconstructed. The results in Fig 5.11 show that most of the position signals of target and predicted states are highly aligned together, and the reconstructed position signals are highly matched in some signals. In Fig 5.11, the first two reconstructed position signals trends are highly aligned in the first row. The second trend plot in the second row shows a slight difference between the predicted position points and the original position points. The general reconstruction result of the position network shows a reconstruction match signals between the predicted and the target states.

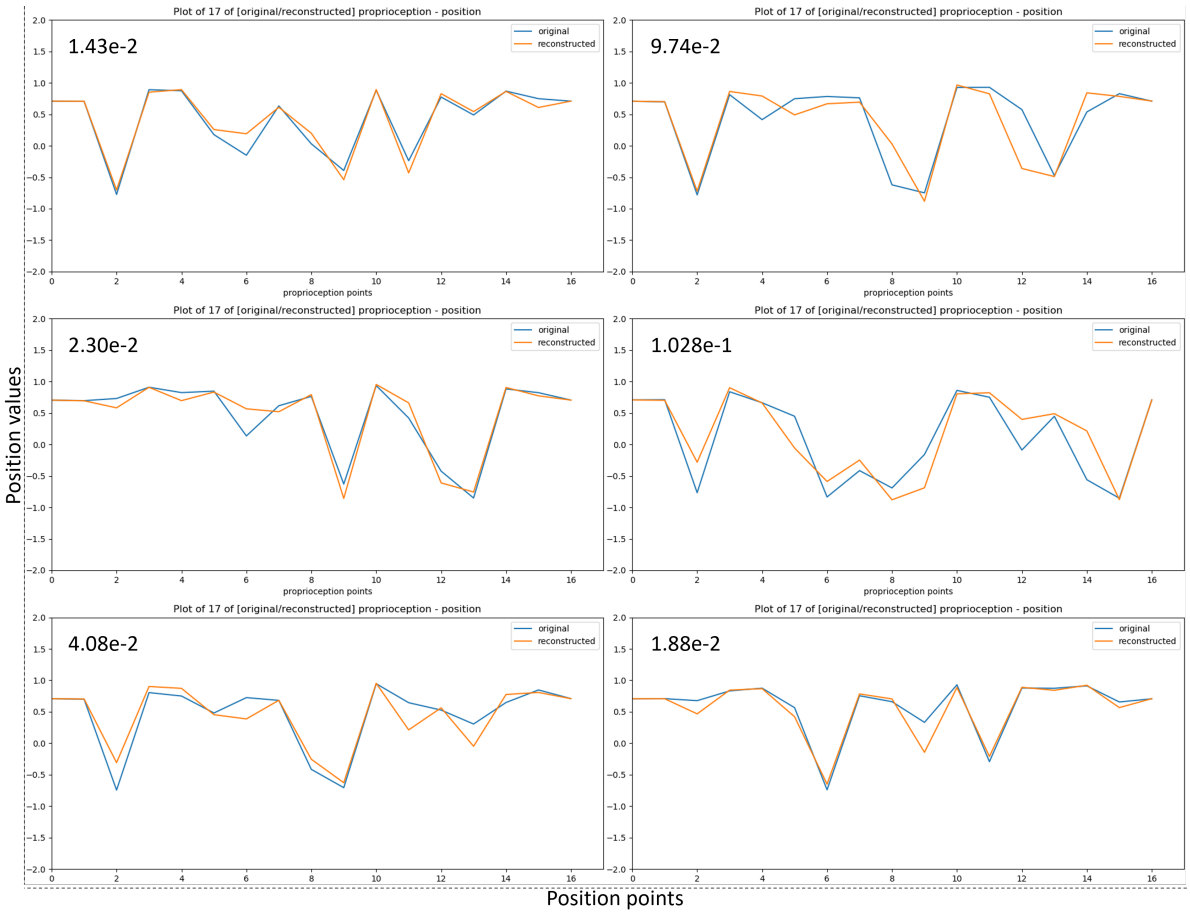

Figure 5.11: The original signal (blue trend) and predicted signal (orange trend) of the position network are plotted; each trend signal consists of 17 units. The majority of the trend signal plots show that the predicted state signal is almost aligned with the original state signal.

5.4.2 Level-2 Classifier Network

The encoded vector of the level-1 fusion network is fed to the level-2 LSTM network. The level-2 LSTM processed and predicted the next state of the fusion vector state. In this section, the classifier is used to classify the level-2 LSTM predicted state by identifying the output of Self or Environment.

Train Classifier Network

The level-2 LSTM-Classifier is trained based on the loss function of equation 5.1, where $loss1$ is the MSE between the predicted states and the target states, and $loss2$ is Cross-Entropy Loss between the predicted states and the target labels. The loss is combined with loss1 and loss2 to fine-tune the network regression guided by classification performance.

$$loss = loss1 + loss2 \tag{5.1}$$

The classifier network is trained for 25 epochs and uses stochastic gradient descent (SGD) optimisation with a learning rate of 1e-3 and a learning scheduler set to 7 step size. The network processes data in 64 batches with shuffling. The level-2 LSTM-Classifier training started by loading and freezing level-1 weights, including vision, position, and fusion networks, and load level-2 LSTM weights (without freezing) that were trained before on the samples dataset of different window frame sizes. The level-2 LSTM were fine-tuned by backpropagating end-to-end training of the level-2 classifier network using the samples dataset of groups and four confounding cases and tested by the leave-one-out as an unseen group.

Validate Classifier Network

The level-2 LSTM-Classifier was validated several times, each with a different sample size of 2, 3, 4, 5, 6, 8, and 10 for each group using the validation dataset. The classification average accuracies for the different window frame sizes experiments are shown in Fig 5.12. The overall average accuracy in Fig. 5.12 is calculated by 5.2

$$WindowSizeOverallAverage = \frac{\sum_{i=1}^{4} group_i}{4} \tag{5.2}$$

Validation dataset					
Window Size	**Group-1** Front Computers	**Group-2** Front towel	**Group-3** In lab	**Group-4** Front Glass	**Overall average**
2	89.62%	88.36%	96.91%	86.97%	**90.46%**
3	89.49%	87.32%	95.83%	87.02%	**89.91%**
4	89.88%	85.59%	96.61%	87.02%	**89.77%**
5	89.62%	85.15%	96.61%	86.97%	**89.58%**
6	89.80%	84.89%	96.74%	86.80%	**89.55%**
8	90.79%	83.50%	96.91%	86.97%	**89.54%**
10	90.36%	84.54%	96.83%	87.19%	**89.73%**

Figure 5.12: The accuracy average of the validation dataset results of different LSTM-Classifier experiments was run on different sample window sizes.

The overall validation average results for all window sizes are slightly similar between 89% and 90.5%. The 2-window size classification accuracy achieved a higher average of 90.46%. Group-3 in all window sizes has an accuracy of 95% and above.

Test Classifier Network

The level-2 LSTM-Classifier was tested by processing different window frame sizes of 2-to-10 encoded fusion vectors. The test involved the unseen test groups dataset samples and the four confounding cases. The test was conducted using the trained models and tested with their leave-one-out samples of an unseen test group of confounding cases.

Fig 5.13 shows the average accuracy of the four groups with different window sizes used from 2-to-10. Also, shows the overall average accuracy derived using equation of 5.2 for the four experiment groups for the different window sizes. The 2-window size classification accuracy achieved the highest overall average of 86.01%, and the 10-window size achieved the lowest overall average of 80.89% over the unseen data groups. The remaining window sizes of 3, 4, 5, 6, 8 overall averages are 83.91%, 84.14%, 81.66%, 81.48%, 80.93%, respectively.

Unseen classification dataset					
Window Size	**Group-1** Front Computers	**Group-2** Front towel	**Group-3** In lab	**Group-4** Front Glass	**Overall average**
2	93.60%	90.64%	69.27%	90.52%	**86.01%**
3	86.92%	90.35%	70.18%	88.20%	**83.91%**
4	88.59%	90.62%	70.75%	86.59%	**84.14%**
5	79.06%	90.79%	71.37%	85.43%	**81.66%**
6	78.62%	90.64%	71.57%	85.08%	**81.48%**
8	75.47%	91.04%	71.94%	85.28%	**80.93%**
10	74.42%	90.94%	71.97%	86.21%	**80.89%**

Figure 5.13: The results of classification averages of the unseen dataset using samples of different window sizes.

Confusion matrices are used to get more details about the unseen average results. Therefore, a confusion matrix of 2x2 is plotted for each group, along with different window sizes. In Fig 5.14 the unseen datasets distributed into matrices, the data show that in groups such as Group-1 and Group-3, most of the misclassification occurs to the environment class where it is classified as self.

		Group-1 Front		Group-2 Front towel		Group-3 In lab		Group-4 Front Glass	
		Predicted Self	Predicted Env.	Predicted Self	Predicted Env.	Predicted Self	Predicted Env.	Predicted Self	Predicted Env.
Unseen Test Groups Size: 2	True Self	21.65%	3.45%	21.00%	4.00%	21.48%	3.57%	17.49%	7.54%
	True Env.	2.95%	71.95%	5.35%	69.65%	27.16%	47.79%	1.94%	73.03%
Unseen Test Groups Size: 3	True Self	22.00%	2.95%	20.97%	4.03%	21.21%	3.87%	18.20%	6.80%
	True Env.	10.12%	64.93%	5.62%	69.38%	25.94%	48.98%	4.99%	70.01%
Unseen Test Groups Size: 4	True Self	21.55%	3.55%	21.17%	3.83%	21.16%	3.89%	18.52%	6.50%
	True Env.	7.86%	67.04%	5.54%	69.46%	25.35%	49.60%	6.91%	68.07%
Unseen Test Groups Size: 5	True Self	20.41%	4.69%	21.31%	3.69%	21.23%	3.82%	18.67%	6.35%
	True Env.	16.25%	58.66%	5.52%	69.48%	24.80%	50.15%	8.22%	66.76%
Unseen Test Groups Size: 6	True Self	19.99%	5.11%	21.29%	3.71%	21.38%	3.70%	19.05%	5.97%
	True Env.	16.27%	58.63%	5.64%	69.36%	24.73%	50.20%	8.95%	66.03%
Unseen Test Groups Size: 8	True Self	19.59%	5.36%	21.22%	3.78%	21.43%	3.62%	19.51%	5.52%
	True Env.	19.17%	55.88%	5.18%	69.82%	24.43%	50.52%	9.20%	65.78%
Unseen Test Groups Size: 10	True Self	19.49%	5.61%	20.92%	4.08%	21.55%	3.52%	19.56%	5.47%
	True Env.	19.97%	54.94%	4.98%	70.02%	24.50%	50.42%	8.32%	66.66%

Figure 5.14: The confusion matrix results of the classification of the unseen dataset by samples of different window sizes.

5.5 Discussion

5.5.1 Level-2 LSTM

The validation datasets results show that increasing the window size is helping to achieve a better Mean Square Error (MSE) average between the predicted vector state and the target state. The MSE of the 10-window of frame size is 1.60e-4, less than 15.6% of the MSE associated with the 2-window frame size of 1.86e-4. Moreover, the reconstruction MSE average for the first 64 batches of fusion latent space shows better MSE average results, which indicates that processing a bigger sample timestep on level-2 LSTM helps to get a better reconstruction. And because in this level-2 LSTM stage, the model compared the predicted state with its target, the minimum

reconstruction MSE average between them shows the model's capability of predicting the next state of a robot. Furthermore, processing a sample of an optimal windows size participates in the reconstruction, mainly they become clearer and near to the target, as in Fig 5.8, after separation and reconstruction, in the second row, the last rightmost predicted image is representing a clearer vision state. The latent space gets split and reconstructed. Thus, the latent space prediction is important, represented by a distribution reconstructed by level-1 networks. In addition, a good MSE reconstruction in level-2 can predict the temporal robot state.

5.5.2 Level-2 Temporal Classification

In comparing the average classification accuracy for the different window frame sizes over the unseen groups and confounding cases, the result shows that the minimum window frame sizes of 2, 3, and 4 yields a higher overall average classification accuracy of 86.01%, 83.92%, and 84.14% respectively. Thus, the level-2 LSTM classifier model can process two to four times steps of an encoded fusion vector to predict self or env. While the best overall average result is from the 2-window frame size of 86.01%, the lowest overall average accuracy result is from the 10-window frame size with 80.89%. The results in Fig 5.13 show that the level-2 LSTM classifier performs better by processing the predicted time step produced from the minimum steps of the encoded fusion vector.

The most cluttered (complex) unseen test groups are Front Computers and In Lab, in contrast to the less complex (semi cluttered) unseen test groups of Front Towel and Front Glass. The slight changes among results for the same unseen group within different sample sizes have a positive indication, which means the predicted state is consistent, such as in Front Towel, In Lab, and Front Glass. However, in Front Computers, it is not consistent, and they decreased in their average accuracy when considering higher sample sizes. The consistent results prediction in the classification shows that level-2 can process a different sample of window size as expected, and the model is not generating noise. The Front Computers: Group-1 confusion metrics result in Fig 5.14, show that the environment class mainly is participating in the misclassification results, where the false positive data that represent the environment is decreasing noticeably from 71.95% in the 2-window size to 54.94% in the 10-window size, and the false negative to increase gradually for the same window size ranges. This show that some of the environment features are classified as self when that features get processed within samples of bigger window size. On the other hand, the group of Front Computers has performed better with consistency within the validation dataset, as in Fig 5.12.

The changes in sample window size show variance between the tested unseen groups. The true positive percentages of the self class show consistency over the unseen groups throughout the samples with different window sizes, except the Front Glass group-4, which show slight increases in the true positive percentages. The Front Glass show better classification related to the window size, where the true positive average accuracy increases gradually from 17.49%

using 2-window size to 19.56% using 10-window size.

In comparing groups' average results from the window size point of view of the reconstruction Fig 5.1 and the classification Fig 5.13, comparing both trends of the reconstruction losses and the overall averages makes it noticeable that the former averages are increasing where the latter is decreasing. Based on a task, the reconstruction average is important in level-2 to get a firm and close belief in what a robot predicts to see and feel. On the other hand, a robot must confirm whether the observed scene is self or environment; thus, classifying the class faster by processing fewer frames is recommended within the classification task.

5.5.3 Level-2 and Level-1 Comparison

In chapter 4, the model of artificial self-awareness level-1 got a lower classification average accuracy of 80.7% compared with the level-2 best classification average accuracy of 86.01%. In level-2, we achieve a higher average accuracy of self-certainty by 5.31% for the same dataset.

In comparing both results of level-1 in chapter 4 Fig 4.16 and level-2 Fig 5.13 classification from the experiment group perspective, it is noticeable that all experimental groups perform better in level-2, especially in group-1 (FC), where in level-2, the accuracy advanced by 9.7% with over all of 93.6%. In group-3 (IL), both level-1 and level-2 achieved a classification accuracy of 67.35% and 69.27%, respectively, showing a slightly better average achieved by level-2. The overall indication that level-2 has advanced the results in all the groups is due to temporal confirmation of level-2 model that helps a robot to recognise itself by predicting its next state of information.

5.6 Level-2 Statistical Analysis

This section aims to further perform a statistical significance analysis for the results of the level-2 fusion MSEs predicted features. To show that the result is statistically significant, it is required to show enough evidence provided to reject a null hypothesis. The null hypothesis (h0) [88] is a statistical hypothesis that assumes the null hypothesis is correct until enough evidence exists to support rejecting this hypothesis. In level-2 analysis, h0 is: There is no relation between the windows sizes and the fusion MSE error variables (Windows size is not effective towards the predicted fusion signal).

Moreover, a P-value is used to determine if the changes in the sample indicate an effect on the population or if it could have occurred simply by chance or sampling error; thus, P-value is used to decide if something is statistically significant. In addition, the Alpha value used for this experiment is 0.05 (5%), which helps to decide whether to reject or accept the above Null hypothesis.

A correlation test was selected to determine the correlation between two samples; a Pearson correlation was performed to test if there is a relationship between the window sizes and the

MSE average of fusion predictions reconstruction. The Pearson correlation results indicated a significant positive association between window size and fusion reconstruction. The calculated coefficient is 0.921871105 (Max=1), and the P-value is 0.003140937, which is highly less than the Alpha value of 0.05. This supports a relationship between the window sizes and the predicted fusion signal and shows that the result produced from the level-2 experiment do not occur by chance.

5.7 Conclusion

This chapter presented level-2 of the artificial self-awareness framework. The proposed modular artificial self-awareness framework allows a robot to get a sense of itself and situate it within a different environment. Therefore, devising this framework in a robot allows it to recognise itself within a different environment. This chapter shows that a higher degree of self-awareness is captured in level-2 of artificial self-awareness, and a higher level yields higher self-confidence. Also, it shows that the experimental hypothesis of level-2 holds for the experimented framework; the predicted states of what a robot is intrinsically believed are meet the robot's original state. Also, the predicted state of the fused modalities of position and vision inputs are successfully reconstructed and comparable with their original state of position and vision. Moreover, higher average classification accuracy is achieved by 5.71% for level-2 over level-1.

This chapter verifies one of the hypothesises of this thesis:

- A temporal perception of the robot's dynamic movement increases the self-certainty of the robot in a different environment if the predicted reconstructed signals for vision and proprioception are statistically significant and classification accuracy is above level-1 for the same dataset.

In future work, Artificial self-awareness in a task context will be introduced, which is applied to a safety application while performing a simple, specific task.

Chapter 6

Conclusion and Future Work

This thesis presents the artificial self-awareness work, allowing a robot to acquire self-knowledge to sight and sense itself in different environments. The artificial self-awareness is devised based on levels where each level shows a higher capacity for self-awareness. These levels of artificial self-awareness are inspired by the cognitive study of human self-awareness development levels by Rochat's [1]. The minimal self that represents the first two levels (level-1 and level-2) of self-development is considered in this thesis. These levels depend on the robot's sensory inputs, mainly proprioception and vision. The senses fusion is achieved to represent multi-modal features in a robot that allows state it self in different environments.

This chapter includes the artificial self-awareness study achievements and conclusion. In addition, includes the recommendation for future work.

6.1 Summary of Contributions

This thesis is involved in achieving the artificial self-awareness study objectives, answering the research questions of each experiment, and validating the corresponding hypothesis. The summary of the level-1 and level-2 artificial self-awareness achievements are as follows:

6.1.1 Enabling the Sense of Self in a Dual-Arm Robot

This subsection summarises the achievements of chapter 3 by revisiting the objectives, the research questions, and the hypothesis and stating their achievements.

Objectives

- Understand self-awareness level-1 requirements and devise it artificially for a robot. Devise multisensory data to enable a robot to associate its movements with its physical body. Initiate a pilot model architecture of artificial self-awareness that enables the sense of self, using a deep neural network and simulated data. Develop an extended model architecture

of artificial self-awareness that enables the sense of self, using real robotic data. Demonstrate the artificial self-awareness level-1 acquisition with a physical robot.

The level-1 requirements are investigated by consulting different self-robotic studies and identifies the gaps mentioned in the chapter 1, section 1.3, also cognitive studies participate in shaping the self-awareness levels of this thesis research specifically Rochat's [1] research of self-awareness development in humans. Accordingly, the artificial self-awareness depicted two implicit levels of self-awareness (level-1 and level-2) to define the artificial self in a robot. Moreover, theses described a multimodal dataset of a simulated robot and devised an artificial sense of self in a robot using a pilot model architecture consisting of deep neural networks. Furthermore, chapter 3 objectives are achieved by devising the sense of self in a robot using a real multimodal dataset with an extended model architecture of a deep neural network. The final objective is achieved by allowing a real physical dual-arm robot to identify itself or its environment.

Research Questions

- Pilot model RQ1: Is it possible to initiate a sense of self in a robot by processing simulated visual and internal velocity feedback?
- Pilot model RQ2: Is it possible to associate simulated visual and internal velocity using DNN to enable a robot to acquire a sense of self?
- Extended model RQ1: Does developing the sense of self allow a robot to know itself?
- Extended model RQ2: Is the robot able to associate the observed and the felt movements in different environments with high accuracy?

Hypothesis

- The sense of self can be enabled in a robot using a Deep Neural Network model architecture and the association of simulated multimodal data.
- Level-1 for artificial self-awareness in the robot increases its self-certainty in an unseen environment.

The pilot model architecture was the first experiment to acquire a sense of self using simulated data of vision and proprioception. The pilot model was built based on basic deep neural architecture as shown in section 3.3.4. The model's average classification accuracy is 93.05% and the robot was capable of sensing itself within a simulated environment. This result shows that the pilot model RQ1 associated with the possibility of initiating a sense of self in a robot

using multimodal of simulated visual and internal velocity data is answered successfully. Moreover, the proposed model answers the RQ2 because the pilot model enabled the sense of self by using the deep neural network architecture, which shows the capability of DNN to encode and learn the data associations to define the sense of self.

The extended model architecture in 3.4 introduced to increase the perceptual capabilities of the robot and investigate the sense of self acquisition. A real dataset of robot vision and proprioception was prepared for this model, including four different environments of different complexity, each with confounding signals to test the sense of self in a robot. The extended model RQ1 is achieved by developing artificial self-awareness architecture that acquires the sense of self, which enables a robot to identify itself. This is supported by the extended model RQ2 which is also answered by the result of robot ability to relate its observed and sense movements in different unseen environments with an average classification accuracy of 88.7%. Moreover, the extended model classification results backed by a saliency map show insight into which parts of the image the network focused on to predict the underlying output. Also, the proposed model shows that its mutual information while being trained with different data showed less variability in the underlying information learned across different environments groups.

Based on the above results and achievements the hypothesis are validated, because the sense of self is enabled in a robot using deep neural network architectures using simulated and real environments. Also, a robot was able to achieve self-certainty in different unseen environments.

6.1.2 The Artificial Self-Awareness Level-1

This subsection summarises the achievements of chapter 4 by revisiting the objectives, the research questions, and the hypothesis and stating their achievements.

Objectives

- Reconstruct a scalable and interpretable level-1 of artificial self-awareness architecture. Building architecture processes multimodal sensory data. Building an architecture that can fuse the vision and, position data into a low-level data representation. Construct an output state of integrated vision and position vector to be processed by the next level of artificial self-awareness (level-2).

In chapter 4, corresponding to the above Artificial Self-awareness level-1 objectives, two deep neural network architectures were built based on generative models, representing the level-1 baseline model and the level-1 VAE architectures. Also, a supervised method was used to classify the output of the two unsupervised network architectures. The level-1 VAE has achieved the objectives of processing multimodal sensory data, fusing the multimodal data into a low-level data representation, and constructing an output representing a vision and position feature that is later used for level-2 of artificial self-awareness.

Research Questions

- RQ1: Can a robot initiate a sense of self by associating its multimodal data using an unsupervised learning method?
- RQ2: Can a robot's sense of self enable it to interpret what it sees and sense by reconstructing its fused multimodal features?

Hypothesis

- A robot's self-awareness sense of self can be enabled using an unsupervised learning method to associate and reconstruct the fused real multimodal data.

The baseline model architecture that consisted of autoencoder networks was able to answer the above RQ1. By initiating the sense of self using unsupervised learning, this chapter introduced a method that can associate multimodal data and represent into a latent vector. The latent vector is classified with an average classification of 77.55% in distinguishing the self class from the environment class over four experimental groups using unseen test datasets. While the level-1 baseline model answered only the RQ1, the level-1 VAE was able to answer both RQ1 and RQ2. The level-1 VAE architecture was built with two variational autoencoder networks to process the vision and position data and an autoencoder as a fusion network. The classification result showed a robot's capability to sense self using the multimodal latent vector with a classification score of 80.3% as average over the four experimental groups using unseen test datasets. Also, level-1 VAE offered the ability to interpret the network's output by reconstructing the fusion data into the corresponding original vision and position data which answers RQ2.

The result answered the above research question and verified its chapter's hypothesis, which validates that self-awareness acquisition is possible by encoding and associating with multimodal data to differentiate the robot from the environment. Also, the reconstruction results validate that unsupervised methods help a robot interpret its vision and sense signals.

6.1.3 The Artificial Self-Awareness Level-2

This subsection summarises the achievements of chapter 5 by revisiting the objectives, the research questions, and the hypothesis and stating their achievements.

Objectives

- Considering an advanced level of self-awareness such as level-2 gives a robot the ability to go beyond getting a sense of self by allowing it to utilise its higher self-confidence and integrate it to control its task. Devise level-2 of artificial self-awareness to adapt self over time and enhance a robotic task. Applied the framework of different environments

with the confounding case and demonstrate that level-2 substantially improve robot self-recognition (higher self-certainty).

The artificial self-awareness level-2 presented in chapter 5 associated with the above objectives was archived by confirming its ability to process the level-1 fused multimodal features. The fused multimodal data contains the robot visual and sense timesteps states processed by the level-2 using a long short term memory architecture and produced the next predicted timestep successfully, where the predicted state gets reconstructed and compared with its original target data, which showed close mean square error average to their originals. Also, level-2 of artificial awareness has achieved another objective by showing a higher self-certainty in comparing its classification accuracy with level-1 of artificial self-awareness.

Research Questions

- RQ1: Is a robot able to relate its arm movements with its body?
- RQ2: Is a robot able to build an inter-modal link between what is seen and felt?

Hypothesis

- A temporal perception of the robot's dynamic movement increases the self-certainty of the robot in a different environment if the predicted reconstructed signals for vision and proprioception are statistically significant and classification accuracy is above level-1 for the same dataset.

The above RQ1 is answered successfully as the predicted timestep state meets the original next state, intrinsically allowing the robot's beliefs based on the predicted timestep to meet the original state. Moreover, the RQ2 is answered as a robot was able to learn an inter-modal relation and predict the next timestep. This was verified by separating the modalities from the predicted timestep fusion vector and comparing their reconstruction signals with their target signals. The result shows a successful reconstruction that is comparable with their target state of position and vision.

While the above research questions are achieved, the hypothesis was validated by achieving a P-value of 0.0031, which supports that the relationship between the window sizes and predicted fusion signal is not by chance. Moreover, the classification of the predicted timestep of the same dataset that was used for level-1 and level-2 yielded an increase of 5.71%, which showed that with level-2 of artificial self-awareness has a higher self-certainty.

6.2 Summary

Human self-awareness is an essential element that is developed in every human in the early stages of their lives, which allows individuals to know and recognise themselves. Researching and devising a prototype of artificial self-awareness into a robot will make a robot recognise itself. This research is inspired by the development stages of self-awareness described by Rochat [7]. With the advancement of artificial neural networks, the levels of artificial self-awareness are structured and trained using deep neural modules. Each level results will present a milestone in the development of artificial self-awareness. Integrating the developed modules will allow the robot to acquire self-awareness.

Moreover, the robot's autonomy degree increases by integrating the artificial self-awareness feature. The result of the first level of the artificial self-awareness model shows that a robot achieved a sense of self and allows it to recognise itself and can differentiate itself within a different environment. The result of the second level of artificial self-awareness helped a robot observe and confirm the temporal sensory changes of its arms, which helped to situate the robot in a different environment.

6.3 Future Work

The work conducted in this theses opens new opportunities to further enhance a robot as described on what follows.

6.3.1 Artificial Self-Awareness Level-3

The modular approach followed in this thesis allows easy integration of other levels of self-awareness. Thus, the third level of self-awareness, based on Rochat's self-awareness development, represents the identification level. In this level, a robot utilises the predicate feature output and processes it to identify itself within a scene by segmenting its arm based on its vision and proprioceptive features. Level-3 of artificial self-awareness will help identify the referred self arm with an overlay segmentation in a scene. This will increase the robot's self-certainty, as a robot will be disentangled from the scene, giving it more ability to use the self and adapt to different environments.

6.3.2 Artificial Self-Awareness in Application Context

Artificial self-awareness is an important ability that does not exist in robotic applications. Integrating an artificial self-awareness ability in a robot will help a robot adapt itself to a task's environment. This because a self-aware robot can distinguish itself from the environment, which helps a robot perform its specific tasks effectively. This future work includes the experiment that

reflects the developed self-aware models within an experimental context. That is, a standalone experiment integrated with the developed artificial self-awareness system.

Use Case

Many robotic applications can be supported by artificial self-awareness. Furthermore, a safety robotic application is implemented to show a use case for the robotic artificial self-awareness framework research. This research integrates artificial self-awareness into a robotic task of pick and placing. The aim is to test the safety application during a pick-and-place task.

Workspace

A robot's workspace depends on a set of all positions a robot can reach. For a task to be reached by a robot, it needs to be within a robot workspace area. This workspace area is a danger zone, which might cause a hazard to humans sharing the same area or disturb the task. The self-aware robot can provide safe zones within its workspace area based on the arms reaction when identifying itself from the environment.

The pick-and-place task in Fig 6.1 is designed to be in front of a robot and defined by two points, P1, where the cube initially resides and P2, the placing point. In addition, point P0 represents an initial point that resides within robot space. The artificial self-awareness uses the predefined points in the task workspace to safely manage to pick and place the cube into P2.

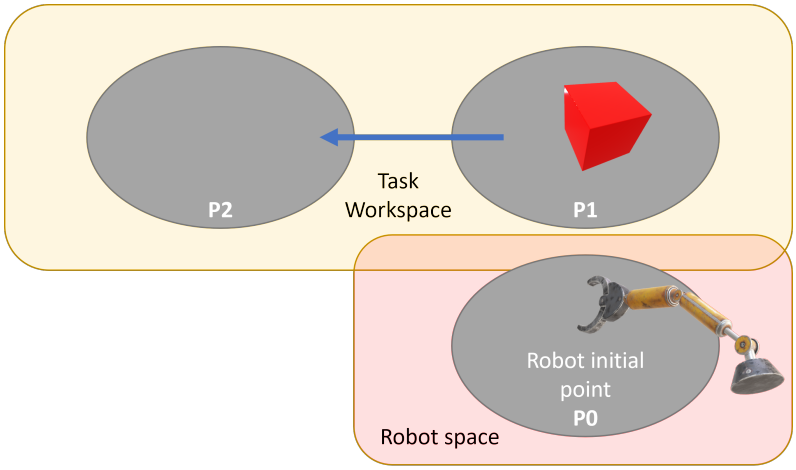

Figure 6.1: The workspace consists of the task workspace with two points, P1 and P2, and the robot space consists of point P0.

Safety Application

In a robotic safety application, a robot must not harm a human or the environment within its working space, and a robot must be safe during the pick and place task. The artificial self-awareness framework can potentially help a robot keep its workspace safe for itself and the task

during implementation. In Fig 6.2, the self-awareness decision thresholds (T) output is utilised to override conditions that help a robot control itself during incidents.

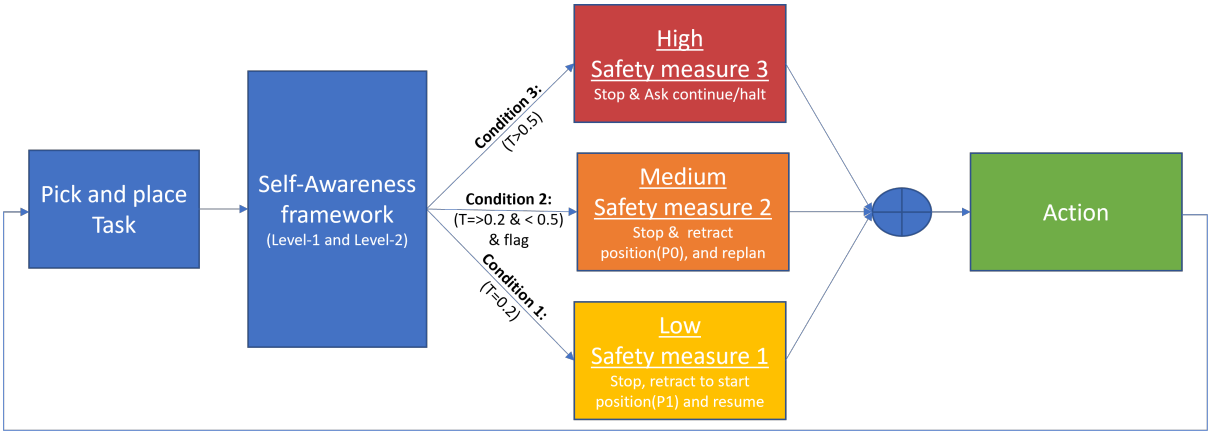

Figure 6.2: The subsumption conditions activate by a threshold(T) that outputs from the artificial self-awareness framework.

The safety application supported with artificial self-awareness must be responsive to external forces that might affect robotic task execution. Thus, a robot will detect the interference that might happen based on the current state information and the next predicted state information. Moreover, as long as an aware robot can predict the next state of its movements integrated with the fusion of its vision and position data successfully, it can also raise a safety concern if something else interferes with its movements.

Our approach in the artificial self-awareness framework is modular based on level-1 and level-2 of artificial self-awareness. In level-1, a robot gets a sense of self by takes a snapshot of his sensory inputs to distinguish itself from the environment. Adding the implementation of self-awareness level-2, a robot can temporally situate itself to distinguish itself from the environment.

The aim of this experiment is, therefore, to integrate artificial self-awareness into a task to show self-awareness benefit in a use case. Furthermore, the experiment aims to show that a robot with artificial self-awareness is safer by detecting itself and controlling itself by, for example, stopping/retracting its arm in a detected incident. This shows that a robot with artificial self-awareness can safely control its task by recognising itself within the environment. Moreover, this shows the usefulness of defining artificial self-awareness and devising it in a robot.

The Experiment Hypothesis

- In a robotic pick and place task execution, an artificial self-aware robot can safely react to stop an unwanted behaviour cussed by an external force such as a push by a human hand within the environment, with a success rate of reacting average response of 70+%.

The Experiment Research Question

- Is the robot supported with self-awareness measure a low safety state and take action by stopping, retracting to position(P1), and resuming the task.
- Is the robot supported with self-awareness measure a medium safety state and take action by stopping, retracting to position (P0) and re-plan the task.
- Is Baxter supported with self-awareness measure a high safety state and take action by immediately stopping and asking an operator to resume the task.

Embedding The Self-Awareness Framework

For this experiment, the Baxter robot can be used to demonstrate the artificial self-awareness framework for a safety use case in a pick-and-place task. During task execution, Baxter, the self-aware robot, can safely select from a prepared action to stop any uncertainty or unwanted result because of external forces applied from the environment. Also, Baxter can prepare to adapt to some situations and continue working on the task safely. The Artificial self-awareness framework embedded in Baxter is as follows:

- Define the pick and place task within the Baxter environment.
- Artificial self-aware framework ROS node: contains a model of level-1 and level-2 of artificial self-awareness.
- Load the training weights of the models level-1 and level-2 of artificial self-awareness.
- Initial checking the input and the predictions by testing different forces and getting the estimated results differences.
- Baxter performs the pick and place task: approaching a certain point for picking an object.
- Checking the fusion MSE: between the predicted state and the current of the fusion vector.
- An action will be decided based on the Fusion MSE changes: if the MSE is high: state "not me", and the action taken is to halt or to retract to the initial point. Otherwise, the MSE is low: state "me" and continue working on the task.
- Safety breaker: an external force will apply to the Baxter's arm while performing a task, using a human arm or an object from the environment.

The Experiment

Baxter attempts to reach, pick, and place an object between two defined points, point A to point B. Firstly, Baxter shows that it can reach the objects properly without interfering. Later, a barrier and external force is applied to measure Baxter's behaviour and check if Baxter can retract and keeps performing the task to complete it safely based on the following three experiment scenarios.

- Scenario 1: Static barrier - A fixed barrier in between start point (P1) and end point (P2).
- Scenario 2: Variable Force - Force applied while performing pick-and-place task from start point (P1) and end point (P2).
- Scenario 3: Fixed Force - Force applied using rob that tide with a fixed mass, during performing pick-and-place task from start point (P1) and end point (P2).

The three scenarios mentioned above would run for two cases with and without the artificial self-awareness framework, and Baxter's reactions are recorded.

The experiment can be quantified using success measures. The success measures are calculated based on the retract response, reaction responsiveness, and completion of the task. The variable associated with the experiment are described in Table 6.1.

Table 6.1: The variable list that associated with the pick-and-place task.

Variable List		
Variable	Fixed	Description
One robot arm	Y	Only one robot arm used during the experiment.
P0	Y	The initial point for robot arm position.
P1	Y	The picking point.
P2	Y	The placing point.
P1-P2 distance	Y	The distance are fixed between the P1 and P2.
Arm Speed	Y	The arm speed uses the default factory.
Attempts	Y	The number of attempt for each scenario is X.
One object	Y	The picking and placing for one cube.
Workspace	Y	The same work space described in Fig. 6.1 is used for all attempts.
Barrier	Y	The same barrier between P1 and P2 is used for all attempts.
Thresholds limit	Y	A threshold limit specified for the three conditions.
Interruption force	Y	Constant force applied that associated with a constant mass.
Human force	N	Force applied by human hand.

The Experiment Evaluation

The arm deviation decision will be based on a robot's belief and its predicted state. A threshold is set to perform an action when reached, ranging from a low safety measure by stopping and

retracting to the position of P1 to a high safety measure by alerting a controller. The threshold is set based on an optimal limit of MSE between the original and predicted frames. The reaction responsive sensitivity is determined based on the best reaction time calculated to activate the retract action.

To evaluate the experiment, a number of experiments can be performed without interfering and another experiment runs with an intervention from the environment.

For statistical analysis, one of the following approaches can be selected:

- 1: The number of experiments (n) incrementally performed with different window sizes or a different result of the mean (b). Then a stability check of the estimate is run to check if the value of b estimated is enough. If not, an additional experiment can be considered until the result shows stability, and that shows that a proper size of experiments is reached.

- 2: Confidence interval: Measure the precision of the estimation of the difference in the two means.

A number of attempts based on the above method can be implemented and the T-Test is calculated for the results from all the experiment attempts to check if they are statistically significant.

Bibliography

[1] P. Rochat, "Five levels of self-awareness as they unfold early in life," *Consciousness and Cognition*, vol. 12, no. 4, pp. 717–731, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053810003000813

[2] K. Gold and B. Scassellati, "A Bayesian Robot That Distinguishes "Self" from "Other"," *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 29, no. 29, 2007. [Online]. Available: https://escholarship.org/uc/item/5z03g2b6

[3] C. Sancaktar, M. A. J. van Gerven, and P. Lanillos, "End-to-End Pixel-Based Deep Active Inference for Body Perception and Action," in *2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, Oct. 2020, pp. 1–8, iSSN: 2161-9484.

[4] P. Rochat, "Self-Unity as Ground Zero of Learning and Development," *Frontiers in Psychology*, vol. 10, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00414

[5] L. Legrain, A. Cleeremans, and A. Destrebecqz, "Distinguishing three levels in explicit self-awareness," *Consciousness and Cognition*, vol. 20, no. 3, pp. 578–585, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053810010001984

[6] R. Van Gulick, "Consciousness," in *The Stanford Encyclopedia of Philosophy*, winter 2021 ed., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2021. [Online]. Available: https://plato.stanford.edu/archives/win2021/entries/consciousness/

[7] P. Rochat, "The self as phenotype," *Consciousness and Cognition*, vol. 20, no. 1, pp. 109–119, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053810010001789

[8] C. Torras, "From the Turing Test to Science Fiction: The Challenges of Social Robotics," *Artificial Intelligence Research and Development*, pp. 5–7, 2013, publisher: IOS Press. [Online]. Available: https://ebooks.iospress.nl/doi/10.3233/978-1-61499-320-9-5

[9] J. M. Jordan, *Robots*, ser. The MIT Press Essential Knowledge series. MIT Press, Oct. 2016, google-Books-ID: yQ9DDQAAQBAJ. [Online]. Available: https://books.google.com.kw/books?id=yQ9DDQAAQBAJ

[10] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015, _eprint: https://www.science.org/doi/pdf/10.1126/science.aab3050. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aab3050

[11] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and Brain Sciences*, vol. 40, p. e253, 2017, publisher: Cambridge University Press.

[12] A. A. Aly and J. B. Dugan, "Experiential robot learning with deep neural networks," in *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, Sep. 2017, pp. 356–361, iSSN: 2161-9484.

[13] V. V. Hafner, P. Loviken, A. Pico Villalpando, and G. Schillaci, "Prerequisites for an Artificial Self," *Frontiers in Neurorobotics*, vol. 14, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnbot.2020.00005

[14] A. Agostini, C. Torras, and F. Woergoeotter, "Integrating Task Planning and Interactive Learning for Robots to Work in Human Environments," p. 6, 2011. [Online]. Available: https://www.aaai.org/ocs/index.php/IJCAI/IJCAI11/paper/view/3200/3705

[15] J.-J. Park, H.-S. Kim, and J.-B. Song, "Safe robot arm with safe joint mechanism using nonlinear spring system for collision safety," in *2009 IEEE International Conference on Robotics and Automation*, May 2009, pp. 3371–3376, iSSN: 1050-4729.

[16] E. Colgate, A. Bicchi, M. A. Peshkin, and J. E. Colgate, "Safety for Physical Human-Robot Interaction," in *Springer Handbook of Robotics*. Springer, 2008, pp. 1335–1348.

[17] Rethink Robotics, "Collaborative Robot Safety and Compliance | Rethink Robotics," Nov. 2018. [Online]. Available: https://web.archive.org/web/20181114181419/https://www.rethinkrobotics.com/safety-compliance (accessed Jul. 13, 2022).

[18] M. Vasic and A. Billard, "Safety issues in human-robot interactions," in *2013 IEEE International Conference on Robotics and Automation*, May 2013, pp. 197–204, iSSN: 1050-4729.

[19] S. Gallagher, "Philosophical conceptions of the self: implications for cognitive science," *Trends in Cognitive Sciences*, vol. 4, no. 1, pp. 14–21, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364661399014175

[20] P. Lanillos, E. Dean-Leon, and G. Cheng, "Yielding Self-Perception in Robots Through Sensorimotor Contingencies," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 9, no. 2, pp. 100–112, Jun. 2017.

[21] P. Lanillos and G. Cheng, "Active inference with function learning for robot body perception," in *International Workshop on Continual Unsupervised Sensorimotor Learning, IEEE Developmental Learning and Epigenetic Robotics (ICDL-Epirob)*, 2018.

[22] J. W. Hart and B. Scassellati, "A robotic model of the Ecological Self," in *2011 11th IEEE-RAS International Conference on Humanoid Robots*, Oct. 2011, pp. 682–688, iSSN: 2164-0580.

[23] Y. Nagai, Y. Kawai, and M. Asada, "Emergence of mirror neuron system: Immature vision leads to self-other correspondence," in *2011 IEEE International Conference on Development and Learning (ICDL)*, vol. 2, Aug. 2011, pp. 1–6, iSSN: 2161-9476.

[24] B. Amos, L. Dinh, S. Cabi, T. Rothörl, S. G. Colmenarejo, A. Muldal, T. Erez, Y. Tassa, N. d. Freitas, and M. Denil, "Learning Awareness Models," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=r1HhRfWRZ

[25] R. Kwiatkowski and H. Lipson, "Task-agnostic self-modeling machines," *Science Robotics*, vol. 4, no. 26, p. eaau9354, 2019, _eprint: https://www.science.org/doi/pdf/10.1126/scirobotics.aau9354. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.aau9354

[26] I. Rodriguez, J. M. Martínez-Otzeta, E. Lazkano, T. Ruiz, and B. Sierra, "On how self-body awareness improves autonomy in social robots," in *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec. 2017, pp. 1688–1693.

[27] K. Gold and B. Scassellati, "Using probabilistic reasoning over time to self-recognize," *Robotics and Autonomous Systems*, vol. 57, no. 4, pp. 384–392, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889008001085

[28] J. W. Hart and B. Scassellati, "Mirror Perspective-Taking with a Humanoid Robot," in *Twenty-Sixth AAAI Conference on Artificial Intelligence*, Jul. 2012. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5177

[29] P. Lanillos, J. Pages, and G. Cheng, "Robot self/other distinction: active inference meets neural networks learning in a mirror," *Santiago de Compostela*, p. 7, 2020.

[30] M. Hoffmann, H. Marques, A. Arieta, H. Sumioka, M. Lungarella, and R. Pfeifer, "Body Schema in Robotics: A Review," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 4, pp. 304–324, Dec. 2010.

[31] J. P. Vasconez, G. A. Kantor, and F. A. A. Cheein, "Human–robot interaction in agriculture: A survey and current challenges," *Biosystems Engineering*, vol. 179, pp. 35–48, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1537511017309625

[32] D. L. Parnas, "On the Criteria to Be Used in Decomposing Systems into Modules," in *Pioneers and Their Contributions to Software Engineering: sd&m Conference on Software Pioneers, Bonn, June 28/29, 2001, Original Historic Contributions*, M. Broy and E. Denert, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 479–498. [Online]. Available: https://doi.org/10.1007/978-3-642-48354-7_20

[33] A. AlQallaf and G. Aragon-Camarasa, "Enabling the Sense of Self in a Dual-Arm Robot," in *2021 IEEE International Conference on Development and Learning (ICDL)*, Aug. 2021, pp. 1–7.

[34] IEEE Spectrum, "What Is a Robot? - ROBOTS: Your Guide to the World of Robotics." [Online]. Available: https://robots.ieee.org/learn/what-is-a-robot (accessed Jun. 17, 2022).

[35] M. Mataric, *The Robotics Primer*, ser. Intelligent Robotics and Autonomous Agents series. MIT Press, 2007. [Online]. Available: https://books.google.co.uk/books?id=WWJPjgz-jgEC

[36] A. Chen, R. Yin, L. Cao, C. Yuan, H. Ding, and W. Zhang, "Soft robotics: Definition and research issues," in *2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, Nov. 2017, pp. 366–370.

[37] Open Robotics, "Documentation - ROS Wiki." [Online]. Available: https://wiki.ros.org (accessed Jun. 19, 2022).

[38] ——, "Gazebo." [Online]. Available: https://gazebosim.org/home (accessed Jun. 26, 2022).

[39] PickNik Robotics, "PickNik Robotics Software R&D." [Online]. Available: https://picknik.ai (accessed Jun. 30, 2022).

[40] Rethink Robotics, "Hardware Specifications - sdk-wiki." [Online]. Available: https://sdk.rethinkrobotics.com/wiki/Hardware_Specifications (accessed Jul. 23, 2022).

[41] Stereolabs Inc, "Stereolabs - Capture the World in 3D." [Online]. Available: https://www.stereolabs.com (accessed Jul. 07, 2022).

[42] R. Alterovitz, S. Koenig, and M. Likhachev, "Robot Planning in the Real World: Research Challenges and Opportunities," *AI Magazine*, vol. 37, no. 2, pp. 76–84, Jul. 2016. [Online]. Available: https://ojs.aaai.org/index.php/aimagazine/article/view/2651

[43] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. [Online]. Available: https://doi.org/10.1038/nature14539

[44] P. R. Kumar and E. B. K. Manash, "Deep learning: a branch of machine learning," *Journal of Physics: Conference Series*, vol. 1228, 2019.

[45] "PyTorch." [Online]. Available: https://www.pytorch.org

[46] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231216315533

[47] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[49] N. Adaloglou, "Intuitive Explanation of Skip Connections in Deep Learning," Mar. 2020. [Online]. Available: https://theaisummer.com/skip-connections/

[50] M. Welling and D. P. Kingma, "Auto-encoding variational bayes," *ICLR*, 2014.

[51] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019. [Online]. Available: http://dx.doi.org/10.1561/2200000056

[52] I. Csiszar, "I-Divergence Geometry of Probability Distributions and Minimization Problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146 – 158, 1975, publisher: Institute of Mathematical Statistics. [Online]. Available: https://doi.org/10.1214/aop/1176996454

[53] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[54] colah, "Understanding LSTM Networks – colah's blog." [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[55] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5929–5955, Dec. 2020. [Online]. Available: https://doi.org/10.1007/s10462-020-09838-1

[56] V. Radu, C. Tong, S. Bhattacharya, N. D. Lane, C. Mascolo, M. K. Marina, and F. Kawsar, "Multimodal Deep Learning for Activity and Context Recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 4, Jan. 2018, place: New York, NY, USA Publisher: Association for Computing Machinery. [Online]. Available: https://doi.org/10.1145/3161174

[57] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," in *ICML*, 2011, pp. 689–696. [Online]. Available: https://icml.cc/2011/papers/399_icmlpaper.pdf

[58] S. Münzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. Dürichen, "CNN-Based Sensor Fusion Techniques for Multimodal Human Activity Recognition," in *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, ser. ISWC '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 158–165, event-place: Maui, Hawaii. [Online]. Available: https://doi.org/10.1145/3123021.3123046

[59] R. Chatila, E. Renaudo, M. Andries, R.-O. Chavez-Garcia, P. Luce-Vayrac, R. Gottstein, R. Alami, A. Clodic, S. Devin, B. Girard, and M. Khamassi, "Toward Self-Aware Robots," *Frontiers in Robotics and AI*, vol. 5, 2018. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2018.00088

[60] D. Claveau and S. Force, "Lurch: The Social Robot that can Wait," in *2018 Second IEEE International Conference on Robotic Computing (IRC)*, Jan. 2018, pp. 177–178.

[61] A. Andriella, G. Alenyà, J. Hernández-Farigola, and C. Torras, "Deciding the different robot roles for patient cognitive training," *International Journal of Human-Computer Studies*, vol. 117, pp. 20–29, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581918300958

[62] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, Feb. 2010, number: 2 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nrn2787

[63] C. Finn and S. Levine, "Deep visual foresight for planning robot motion," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 2786–2793.

[64] P. Marion, P. R. Florence, L. Manuelli, and R. Tedrake, "Label Fusion: A Pipeline for Generating Ground Truth Labels for Real RGBD Data of Cluttered Scenes," in *2018 IEEE*

International Conference on Robotics and Automation (ICRA), May 2018, pp. 3235–3242, iSSN: 2577-087X.

[65] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, S. Levine, and V. Vanhoucke, "Using Simulation and Domain Adaptation to Improve Efficiency of Deep Robotic Grasping," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 4243–4250, iSSN: 2577-087X.

[66] G. Riva, "The neuroscience of body memory: From the self through the space to the others," *Cortex*, vol. 104, pp. 241–260, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010945217302381

[67] N. Haber, D. Mrowca, S. Wang, L. F. Fei-Fei, and D. L. Yamins, "Learning to Play With Intrinsically-Motivated, Self-Aware Agents," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/71e63ef5b7249cfc60852f0e0f5bf4c8-Paper.pdf

[68] G. G. Gallup Jr., "Self-awareness and the emergence of mind in primates," *American Journal of Primatology*, vol. 2, no. 3, pp. 237–248, 1982, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajp.1350020302. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ajp.1350020302

[69] J. Sturm, C. Plagemann, and W. Burgard, "Body schema learning for robotic manipulators from visual self-perception," *Journal of Physiology-Paris*, vol. 103, no. 3, pp. 220–231, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0928425709000461

[70] M. S. Mahler, F. Pine, and A. Bergman, *The Psychological Birth of the Human Infant: Symbiosis and Individuation*. London: Routledge, 1975.

[71] A. Stoytchev, "Self-detection in robots: a method based on detecting temporal contingencies," *Robotica*, vol. 29, no. 1, pp. 1–21, 2011, publisher: Cambridge University Press.

[72] J. S. Watson, "Detection of self: The perfect algorithm," in *Self-Awareness in Animals and Humans: Developmental Perspectives*, S. T. Parker, R. W. Mitchell, and M. L. Boccia, Eds. Cambridge University Press, 1994, pp. 131–148.

[73] G. G. Gallup, "Chimpanzees: Self-Recognition," *Science*, vol. 167, no. 3914, pp. 86–87, 1970, _eprint: https://www.science.org/doi/pdf/10.1126/science.167.3914.86. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.167.3914.86

[74] J. K. O'Regan and A. Noë, "A sensorimotor account of vision and visual consciousness," *Behavioral and Brain Sciences*, vol. 24, no. 5, pp. 939–973, 2001, publisher: Cambridge University Press.

[75] K. E. Twomey and G. Westermann, "Curiosity-based learning in infants: a neurocomputational approach," *Developmental Science*, vol. 21, no. 4, p. e12629, 2018, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/desc.12629. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/desc.12629

[76] R. L. Fantz, "Visual Experience in Infants: Decreased Attention to Familiar Patterns Relative to Novel Ones," *Science*, vol. 146, no. 3644, pp. 668–670, 1964, _eprint: https://www.science.org/doi/pdf/10.1126/science.146.3644.668. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.146.3644.668

[77] C. Lang, G. Schillaci, and V. V. Hafner, "A Deep Convolutional Neural Network Model for Sense of Agency and Object Permanence in Robots," in *2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, Sep. 2018, pp. 257–262, iSSN: 2161-9484.

[78] P. Lanillos and G. Cheng, "Adaptive Robot Body Learning and Estimation Through Predictive Coding," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2018, pp. 4083–4090, iSSN: 2153-0866.

[79] K. Friston, "Hierarchical Models in the Brain," *PLOS Computational Biology*, vol. 4, no. 11, pp. 1–24, Nov. 2008, publisher: Public Library of Science. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1000211

[80] I. Guyon, "A Scaling Law for the Validation-Set Training-Set Size Ratio," in *AT & T Bell Laboratories*, 1997.

[81] M. Ogura and vainaijr, "Misaogura/flashtorch: 0.1.1," Sep. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3461737

[82] H. Bahng, S. Chun, S. Yun, J. Choo, and S. J. Oh, "Learning De-biased Representations with Biased Representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, Jul. 2020, pp. 528–539. [Online]. Available: https://proceedings.mlr.press/v119/bahng20a.html

[83] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning Not to Learn: Training Deep Neural Networks With Biased Data," 2019, pp. 9012–9020. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Kim_Learning_Not_to_Learn_Training_Deep_Neural_Networks_With_Biased_CVPR_2019_paper.html

[84] H. Fang, V. Wang, and M. Yamaguchi, "Dissecting Deep Learning Networks—Visualizing Mutual Information," *Entropy*, vol. 20, no. 11, 2018. [Online]. Available: https://www.mdpi.com/1099-4300/20/11/823

[85] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[86] I. Shafkat, "Intuitively Understanding Variational Autoencoders," Oct. 2021. [Online]. Available: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

[87] L. v. d. Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[88] D. I. MacKenzie, J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines, "Chapter 3 - Fundamental Principals of Statistical Inference," in *Occupancy Estimation and Modeling (Second Edition)*, second edition ed., D. I. MacKenzie, J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines, Eds. Boston: Academic Press, 2018, pp. 71–111. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780124071971000041

Appendix A

Proprioception

The proprioception used in this thesis represents robot kinematics of three main elements Joint Position, Joint Velocity, and Joint Torque. The below information presents a raw sample of Baxter's proprioception. The proprioception information contains the three main elements used across different experiments in this thesis.

```
header:
  seq: 35460
  stamp:
    secs: 709
    nsecs: 763000000
  frame_id: ''
name: [head_pan, l_gripper_l_finger_joint, l_gripper_r_finger_joint, left_e0, left_e1,
left_s0, left_s1, left_w0, left_w1, left_w2, r_gripper_l_finger_joint, r_gripper_r_finger_joint,
  right_e0, right_e1, right_s0, right_s1, right_w0, right_w1, right_w2]
position: [-4.475188283059595e-06, 0.020833011036351917, 3.3320567407784883e-08,
-0.015949463633992522, 0.49432022128339437, 0.19248171594511465,
 1.0470000321181505, -0.19779870973147418, 0.02677042894793935,
-0.015231248174336187, 0.020833001863019673, 1.0446947801077943e-09,
-0.020360697388244198, 0.49534278445302604, -0.2726283000337206,
1.0470000132350048, -0.10529881357510984, 0.028133736675437504,
0.029579277836319307]
velocity: [7.085345183049303e-08, -1.461892986511863e-06, -7.439729175244792e-06,
 -2.1975473107955685e-06, 8.038361891775635e-05, -1.0266391258739978e-06,
  -3.177062681697951e-05, -2.975473849317068e-05, -4.880370497319846e-05,
   2.536788324258493e-06, 7.082878728610074e-08, 7.074817405863714e-08,
    1.6367102806456997e-06, -4.311219372261315e-07, 4.318769952535339e-08,
        2.737705208600908e-09, -3.4153545793299156e-05, 2.458852064911225e-07,
4.730965224763903e-06]
effort: [0.0, 0.0, 16.88, 6.176, 0.092, -9.3, 0.052, -1.848, -0.336, 11.864, 13.22, -0.372,
-12.772, -0.364, -1.532, -0.204, -20.48, 0.092, 0.050]
```

Figure A.1: A raw data sample was captured from Baxter's proprioception information.

Appendix B

Saliency Map

The saliency map for a scaled up sample images for clarity. The sample represent an image from the "In Lab" group input discussed by Chapter 3, in subsection 3.4.7 under the Saliency Map.

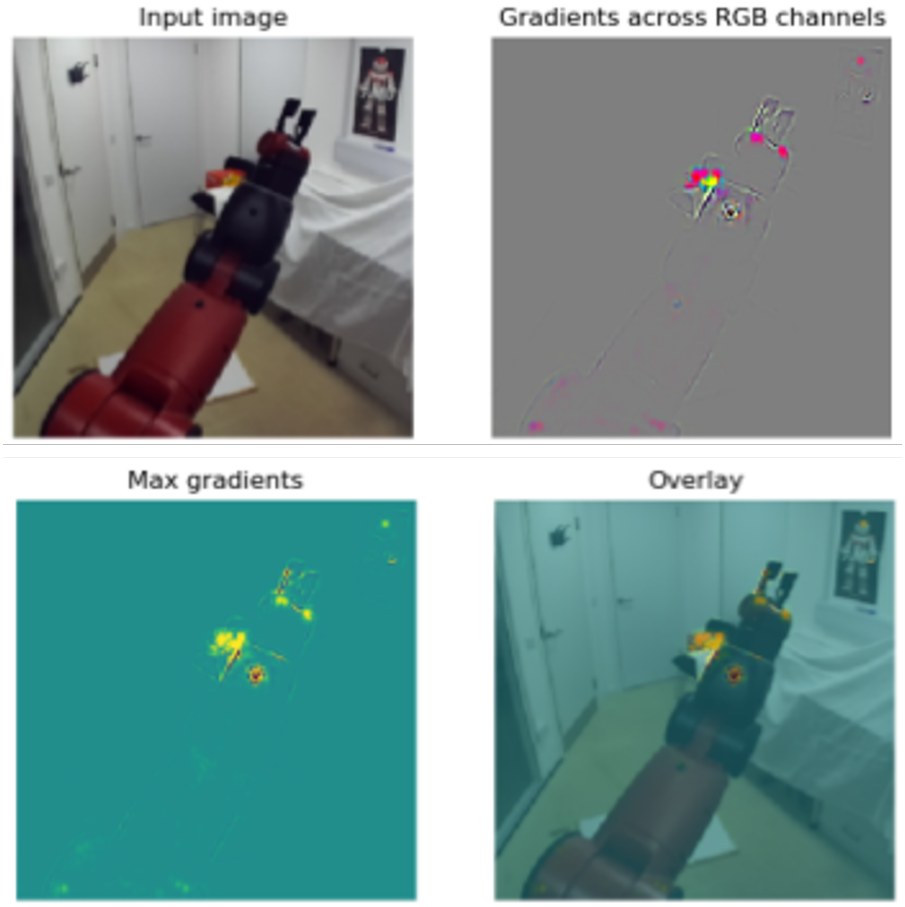

Figure B.1: The first left image in the first raw represents an untouched input sample from the "In Lab" group input image. The three remaining images represent the saliency maps of the input image based on Gradient cross RGB, Max gradients, and the overlay of the gradients with the original image. They show that the level-1 module mainly focuses on the robot's arm related pixels and edges.

Appendix C

Level-2 Vision Reconstruction

The level-2 reconstructions of vision features are processed by level-2 LSTM and then decoded by the vision decoder. This reconstruction scaled up sample images for clarity that are presented and discussed in Chapter 5, section 5.4.1 under the subsection of "Vision Frames Reconstruction".

Figure C.1: Sample of the vision reconstruction results of the target state (the upper row) and the predicted states (the lower row). The values (yellow text) on the second row represent the error difference between the predicted state and its target.