FridgeVision: Innovative Food Management for People with Dementia using advanced computer vision techniques

Submitted May 2024, in partial fulfillment of
the conditions for the award of the degree **MSc Computer Science with artificial intelligence.**

Parth Ashwinbhai Bhalodiya
20480979

Supervised by Armaghan Moemeni

School of Computer Science
University of Nottingham

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature Parth Bhalodiya

Date: 31 / 05/ 2024

I hereby declare that I have all necessary rights and consents to publicly distribute this dissertation via the University of Nottingham's e-dissertation archive.

Abstract

Dementia is a progressive condition that gradually impairs cognitive abilities, leading to difficulties in performing daily tasks and maintaining independence. As the global prevalence of dementia rises, there is an urgent need for innovative assistive technologies to support affected individuals and their caregivers. This dissertation presents FridgeVision, a comprehensive food management system designed to assist individuals with dementia in maintaining proper nutrition and independent living. By leveraging state-of-the-art computer vision techniques, such as object detection using YOLOv8, segmentation with SAM, and latent space analysis, FridgeVision accurately identifies, organises, and monitors food items within the refrigerator. The system incorporates user-centric design principles, voice interaction, and personalised recipe recommendations using a large language model to provide intuitive and accessible assistance. A multi-stage development methodology is employed, encompassing data collection, augmentation techniques, model training, and evaluation. The experimental results demonstrate FridgeVision's exceptional performance in object detection, segmentation, and change tracking. User feedback highlights the system's potential to enhance independence, alleviate caregiver burden, and improve the quality of life for individuals with dementia. This interdisciplinary research at the intersection of computer vision, artificial intelligence, and dementia care paves the way for advanced assistive technologies that address real-world challenges faced by this vulnerable population.

Keywords: Dementia, Assistive Technology, Image Processing, Object Detection, Yolo, Image Segmentation, SAM, Latent Space Analysis, Recipe Recommendation, Llama3

Acknowledgements

I want to express my sincere gratitude to a few people, whose advice and knowledge have been crucial to finishing this project. First and foremost, I extend my deepest gratitude to my dissertation supervisor, **Dr Armaghan Moemeni**, for her unwavering support, guidance, and mentorship throughout this research journey. Her expertise, insights, and encouragement have been instrumental in shaping this work and navigating the challenges of interdisciplinary research. I am truly grateful for her dedication, patience, and the countless hours invested in guiding me towards the successful completion of this dissertation.

I sincerely thank **Joshua Goulton**, my fellow MSci student and invaluable collaborator, for his contributions, particularly in developing voice interaction capabilities and expiry date detection functionality, which have greatly enhanced the FridgeVision system. Working alongside Joshua has been a pleasure, and I am grateful for his creativity, technical skills, and the synergy we have achieved as a team.

I am immensely grateful to **Dr Anto Rajamani** from Queen's Medical Centre, whose expertise in dementia-related research has been crucial to the project's success. Professor Rajamani's insights into the unique challenges faced by individuals with dementia and their caregivers have informed the design and development of the FridgeVision system, ensuring our work remains grounded in the real-world needs and perspectives of those affected by dementia.

I express my heartfelt appreciation to **Dominic Price** from the Cobot Maker Space for his invaluable assistance with the setup and ideation phases of this project. Dominic's technical expertise, creative problem-solving, and hands-on support have been essential in bringing the FridgeVision system to life, overcoming practical challenges and transforming our ideas into a tangible and functional prototype.

I thank the University of Nottingham, the School of Computer Science, my family, and my friends for their support, encouragement, and understanding throughout this research journey. To everyone who has contributed to this research, directly or indirectly, I offer my sincere appreciation and gratitude. This dissertation is a testament to the power of collaboration, interdisciplinary research, and the collective effort to impact the lives of individuals affected by dementia positively.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

1.1 What is dementia?

Dementia is a condition that gradually robs individuals of their cognitive abilities - their ability to think clearly, reason, and remember things. It's a progressive condition, meaning it worsens over time. People living with dementia face increasing difficulties with tasks that were once simple and routine for them, like recalling familiar names and faces, processing thoughts coherently, or even accomplishing basic daily activities like dressing or eating.

Dementia is not a single, specific disease itself. Rather, it is an umbrella term that describes a range of brain disorders that lead to cognitive impairment. The most common cause of dementia is Alzheimer's disease, but there are other types too, such as vascular dementia, Lewy body dementia, and frontotemporal dementia. Each type results from different underlying factors affecting the brain.

Living with dementia involves an unraveling of one's memory, thinking skills, and ultimately, their sense of self and independence. As the condition progresses, it poses immense challenges and drastically impacts the quality of life for those affected and their loved ones. The symptoms of dementia cast a long shadow, not just for the individuals suffering from it, but for their entire support network as well.

1.2 Roots of Dementia and Its Challenges

The global prevalence of dementia is rising at an alarming rate, with the number of individuals affected expected to nearly triple from 55 million currently to over 150 million by 2050 [9].

1.2.1 Underlying Factors

This staggering increase can be attributed to several underlying factors:

Aging Population

Advancing age is the greatest risk factor for most forms of dementia, including Alzheimer's disease, vascular dementia, and Lewy body dementia. As life expectancies continue to rise worldwide, the elderly population is growing rapidly, contributing to the escalating dementia burden [32, 39].

Demographic Transitions

Developing nations are undergoing demographic shifts, with declining fertility rates and increased life expectancy. These transitions are leading to a larger proportion of older adults, exacerbating the prevalence of age-related conditions like dementia [9].

Improved Diagnostic Techniques

Advancements in medical imaging and biomarker research have improved the ability to diagnose dementia, particularly in its earlier stages.
This heightened diagnostic accuracy contributes to the increasing recognition and reported cases of dementia globally [11].

1.2.2 Challenges

The challenges posed by dementia are multifaceted and far-reaching, impacting individuals, families, and societies:

Cognitive Impairment

Individuals with dementia experience a progressive decline in cognitive abilities, including memory, reasoning, and problem-solving skills. This impairment can severely impact their ability to perform routine daily activities, such as managing finances, personal care, and household tasks [6, 46].As the condition advances, they may become increasingly reliant on others for even the most basic tasks, further diminishing their sense of autonomy and self-worth.

Nutritional Deficiencies

One of the significant challenges faced by individuals with dementia is maintaining proper nutrition. Difficulties in recognizing, organizing, and preparing food items can lead to potential risks of malnutrition and associated health complications [31, 27].Poor nutrition can further exacerbate cognitive decline, creating a vicious cycle that impairs overall health and well-being.

Caregiver Burden

The debilitating effects of dementia often necessitate constant supervision and assistance from caregivers, whether family members or professional care providers. This care responsibility can lead to heightened stress, emotional strain, and financial burdens on caregivers [26]. Caregivers frequently experience burnout, depression, and social isolation, as the demands of caring for a loved one with dementia can be physically, mentally, and emotionally draining.

Societal Impact

The escalating prevalence of dementia poses significant economic and societal challenges. The direct and indirect costs associated with dementia care, including medical expenses, long-term care facilities, and lost productivity, are substantial and expected to rise further [9].As the population ages and the dementia burden grows, societies will face increasing pressure to allocate resources and develop sustainable support systems for affected indi-

viduals and their families.

Healthcare System Strain

The increasing demand for dementia-related healthcare services, specialized care facilities, and trained professionals can strain healthcare systems, particularly in resource-limited settings [26].This strain can lead to longer wait times, reduced access to quality care, and a further escalation of costs, exacerbating the challenges faced by individuals with dementia and their caregivers.

Addressing the roots of dementia and its multifaceted challenges requires a comprehensive and multidisciplinary approach, encompassing medical research, public health initiatives, social support systems, and technological innovations. The development of assistive technologies, such as the proposed FridgeVision system, holds the potential to alleviate some of the burdens faced by individuals with dementia and their caregivers, particularly in the realm of maintaining proper nutrition and independence.

1.3 Motivation

The motivation behind this research stems from the pressing need to develop assistive technologies that can address the unique challenges faced by individuals with dementia, particularly in the realm of maintaining proper nutrition and independent living. Dementia, a progressive neurodegenerative condition, poses significant obstacles in performing routine daily activities, including recognizing, organizing, and preparing meals [31, 6]. As shown in Figure 1.1, the global prevalence of dementia continues to rise, with projections indicating a nearly threefold increase by 2050 [9], there is an urgent need for innovative solutions that can enhance the quality of life for affected individuals and alleviate the burden on caregivers.

Existing commercial solutions, such as smart refrigerators offered by major manufacturers, often lack specialized functionalities tailored to the unique needs of individuals with cognitive impairments [18, 23, 33, 38]. These products primarily focus on general con-

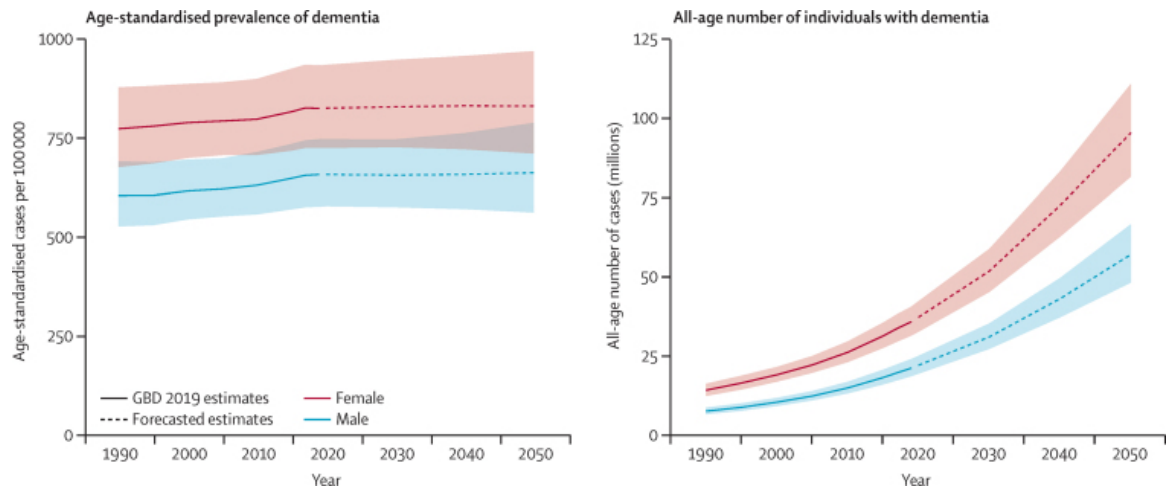

Figure 1.1: Estimated trends in the global age-standardised dementia prevalence (A) and all-age number of cases (B), with 95% uncertainty intervals, 2019–50 GBD=Global Burden of Diseases, Injuries, and Risk Factors Study.[9]

sumer convenience features, failing to address the specific challenges faced by individuals with dementia in recognizing and organizing food items within the refrigerator. While some research efforts have explored adding smart functionalities to existing refrigerators [45, 1], the details on implementation, image capture methods, and accurate food recognition algorithms are often lacking, potentially hindering their effectiveness for the target demographic.

The consequences of improper food management for individuals with dementia can be severe, including nutritional deficiencies, increased caregiver burden, and a diminished quality of life for both patients and their families [27, 47]. Recent advancements in computer vision and deep learning techniques have demonstrated immense potential in addressing these challenges by enabling automated systems that can accurately perceive and understand visual scenes, detect objects, and track information [47]. However, there is a need for solutions that leverage these cutting-edge technologies in a user-friendly and accessible manner, specifically tailored to the unique requirements of individuals with dementia.

The proposed FridgeVision system aims to fill this gap by providing an innovative and affordable food management solution designed explicitly for individuals with dementia. By combining state-of-the-art computer vision techniques, such as object detection and

semantic segmentation, FridgeVision can accurately identify, organize, and monitor food items within the refrigerator, offering personalized assistance to enhance independence, promote proper nutrition, and alleviate caregiver burden.

This interdisciplinary research at the intersection of healthcare and technology is motivated by the potential to contribute significantly to improving the quality of life for individuals affected by dementia and their caregivers. By leveraging the power of artificial intelligence and user-centric design principles, FridgeVision exemplifies the promise of assistive technologies in addressing real-world challenges faced by vulnerable populations, ultimately fostering a more inclusive and compassionate society.

1.4 Aims and Objectives

The primary aim of this research is to develop and evaluate an innovative food management system called FridgeVision, tailored specifically to assist individuals with dementia in maintaining proper nutrition and independent living. By leveraging cutting-edge computer vision techniques, FridgeVision aims to provide personalized assistance in accurately identifying, organizing, and monitoring food items within the refrigerator, ultimately enhancing the quality of life for affected individuals and alleviating the burden on caregivers. To achieve this overarching aim, the following specific objectives have been outlined:

1. Explore and implement state-of-the-art object detection algorithms, such as YOLOv8 [16], to precisely localize and classify various food items within the refrigerator environment.

2. Integrate advanced semantic segmentation techniques, like the Segment Anything Module (SAM) [21], to achieve fine-grained pixel-level labeling and contouring of food items, enabling precise separation and comprehensive understanding of the refrigerator's contents.

3. Develop a latent space analysis framework to efficiently encode and analyze changes in refrigerator contents over time, facilitating automated tracking of food storage and consumption patterns.

4. Construct a comprehensive dataset capturing diverse images of refrigerator environments, food items, and scenarios representative of real-world conditions faced by individuals with dementia.

5. Evaluate the performance of the FridgeVision system in terms of accuracy, robustness, and real-time capabilities, using appropriate metrics and rigorous testing methodologies, such as cross-validation and benchmark comparisons.

6. Investigate potential avenues for improving the system's performance, including data augmentation techniques, semi-supervised learning approaches, and optimizations for efficient deployment on edge devices or mobile platforms.

7. Explore the integration of user-centric design principles and conduct user studies with the target demographic of individuals with dementia and their caregivers, to ensure accessibility, usability, and alignment with their specific needs and preferences.

8. Examine the potential impact of the FridgeVision system on enhancing the daily lives of individuals with dementia, promoting independence, and alleviating the burden on caregivers through interdisciplinary insights from healthcare professionals and subject matter experts.

By achieving these objectives, this research endeavors to contribute to the growing body of knowledge at the intersection of computer vision, artificial intelligence, and healthcare, while also delivering a practical and impactful solution to address the real-world challenges faced by individuals with dementia and their caregivers in managing their nutritional needs.

1.5 Description of the work

The FridgeVision project aims to develop an innovative and user-friendly food management system specifically designed to assist individuals living with dementia in maintaining proper nutrition and independence. By leveraging cutting-edge computer vision and

artificial intelligence techniques, the system seeks to provide personalized assistance in accurately identifying, organizing, and monitoring food items within the refrigerator, ultimately enhancing the quality of life for those affected by dementia and alleviating the burden on their caregivers.

1.5.1 Functional Specification

1. **Food Item Identification and Localization:**

    - Utilize state-of-the-art object detection algorithms, such as YOLOv8 [16], to accurately detect and classify various food items present within the refrigerator.
    - Precisely localize and mark the spatial positions of the identified food items within the captured refrigerator images.

2. **Semantic Segmentation and Contouring:**

    - Employ advanced semantic segmentation techniques, like the Segment Anything Module (SAM) [21], to achieve fine-grained pixel-level labeling and contouring of individual food items.
    - Separate and distinguish between different food items, even in cases of occlusion or close proximity, enabling a comprehensive understanding of the refrigerator's contents.

3. **Food Storage Monitoring and Tracking:**

    - Encode the visual information obtained from object detection and segmentation into a compact latent space representation using an autoencoder framework.
    - Analyze and compare the latent space encodings of previous and current refrigerator contents to quantify changes in food storage and consumption patterns over time.

- Provide automated tracking and visualization of food item additions, removals, and consumption, enabling users to monitor their refrigerator's inventory effortlessly.

4. **User Interface and Interaction:**

   - Develop an intuitive and accessible user interface tailored to the needs of individuals with dementia and their caregivers.
   - Integrate user-centric design principles to ensure ease of use, clear visual representations, and effective communication of relevant information.
   - Provide personalized alerts, reminders, and recommendations based on the monitored food storage and consumption patterns.

5. **Data Collection and Model Training:**

   - Construct a comprehensive dataset capturing diverse images of refrigerator environments, food items, and scenarios representative of real-world conditions faced by individuals with dementia.
   - Employ data augmentation techniques and semi-supervised learning approaches to enhance the robustness and generalization capabilities of the underlying computer vision models.
   - Continuously refine and optimize the models through iterative training and evaluation cycles, leveraging the collected data and user feedback.

6. **Deployment and Integration:**

   - Optimize the system for efficient deployment on edge devices, such as smartphones or dedicated hardware units, enabling real-time performance and accessibility.
   - Explore integration with existing smart home systems or IoT platforms to enhance functionality and provide a seamless user experience.

- Facilitate data synchronization and remote monitoring capabilities, allowing caregivers to stay informed about the food management of their loved ones with dementia.

By achieving these functional specifications, the FridgeVision system aims to provide a comprehensive and personalized solution for individuals with dementia, assisting them in maintaining proper nutrition, promoting independence, and reducing the burden on caregivers. Through continuous refinement, user-centric design, and interdisciplinary collaboration, the project strives to deliver an impactful and practical solution that addresses the real-world challenges faced by this vulnerable population.

Chapter 2

Background and Related Work

The development of the FridgeVision system draws upon a diverse set of research domains and technological advancements, including computer vision, deep learning, assistive technologies, and healthcare informatics. This section provides an in-depth overview of the relevant background and related work that has informed and influenced the design and implementation of the proposed solution.

2.1 Assistive Technologies for Dementia

As the global prevalence of dementia continues to rise, there has been a growing interest in developing assistive technologies to support individuals with cognitive impairments in maintaining their independence and quality of life. These technologies encompass a wide range of solutions, from smart home systems and wearable devices to augmented reality applications and specialized software applications [29, 30, 7, 34].

2.1.1 Smart Home Systems

Smart home technologies have shown promise in assisting individuals with dementia in their daily living activities. A smart home system that leverages depth cameras, sensors and machine learning algorithms to monitor and assist with activities of daily living (ADLs), such as meal preparation, personal hygiene, and medication management, was proposed [42]. The system employs computer vision techniques to track the individual's

movements and gestures, providing context-aware prompts and guidance to support the completion of these tasks.

Similarly, a smart home system was developed that uses sensors and activity recognition algorithms to monitor ADLs and provide reminders and assistance as needed [7].The system can detect potential safety concerns, such as forgetting to turn off the stove or wandering outside the home, and alert caregivers accordingly.

2.1.2 Wearable Assistive Technologies

In addition to smart home systems, wearable devices have been explored as assistive technologies for individuals with dementia. Some developed a wearable system that uses computer vision and activity recognition to provide context-aware prompts and guidance for individuals with cognitive impairments, supporting them in completing daily tasks. The system employs a head-mounted camera and a smartwatch, leveraging computer vision algorithms to analyze the user's environment and activities in real-time.

2.1.3 Augmented Reality (AR) and Virtual Reality (VR) Applications

Augmented reality and virtual reality technologies have also shown potential in supporting individuals with dementia.A virtual reality-based system for cognitive training and rehabilitation, designed to enhance memory, attention, and executive functions, was developed [3]. The immersive virtual environments provide a safe and controlled setting for individuals with dementia to practice cognitive tasks and engage in simulated real-world scenarios.

Similarly, the use of augmented reality for supporting individuals with dementia in their daily activities was explored[34]. Their system leverages AR overlays to provide visual cues and prompts, enhancing the user's understanding and ability to navigate their environment and complete tasks.

While these assistive technologies have shown promising results in supporting individuals with dementia, there is still a need for solutions specifically tailored to address the chal-

lenges associated with maintaining proper nutrition and managing food storage and consumption. The FridgeVision system aims to fill this gap by leveraging cutting-edge computer vision and deep learning techniques to provide a comprehensive and user-friendly solution for food management and nutrition assistance.

By combining advanced object detection, semantic segmentation, and latent space analysis, FridgeVision can accurately identify, organize, and monitor food items within the refrigerator, offering personalized assistance to enhance independence and alleviate caregiver burden. Furthermore, the system emphasizes user-centric design principles, involving the target demographic in the development process to ensure that the solution aligns with their specific needs, preferences, and capabilities.

Through interdisciplinary collaboration with healthcare professionals and subject matter experts, the FridgeVision project aims to deliver a practical and impactful solution that addresses the real-world challenges faced by individuals with dementia and their caregivers in managing their nutritional needs and promoting independent living.

2.2 Smart Refrigerator Solutions

The concept of intelligent refrigerators has garnered increasing attention as a potential assistive technology for individuals with dementia, particularly in addressing challenges related to nutrition and food management. However, existing commercial solutions offered by major manufacturers often lack specialized functionalities tailored to the unique needs of this demographic, primarily focusing on general consumer convenience features [18, 23, 33, 38].

Research efforts in this domain have explored various approaches to augment refrigerators with smart capabilities, leveraging techniques such as sensors, computer vision, and machine learning algorithms. These solutions aim to provide features like automatic inventory management, expiration date tracking, and personalized recommendations to assist users in maintaining proper nutrition and reducing food waste.

2.2.1 Sensor-based Solutions

Several research works have explored the use of various sensor modalities to monitor and track the contents of a refrigerator. A system that combines weight sensors and RFID tags to monitor food items was proposed[1]. The system can detect the addition or removal of items based on weight changes, and RFID tags provide additional information about the item's identity, expiration date, and nutritional content. The authors demonstrated the system's ability to track food consumption and provide notifications about expiring items.

An intelligent fridge system called "Fridge" was developed that uses RFID technology for food management [48]. The system employs RFID tags to identify and track individual food items, enabling accurate monitoring of the fridge contents. Algorithms were implemented to provide food recommendations based on the user's preferences and consumption patterns, as well as alerts for expiring items.

While RFID-based and sensor-based solutions offer valuable insights into the contents of a refrigerator, they have limitations. RFID tags require individual labeling of food items and may not provide detailed information about food quality, while sensor-based solutions often require intrusive modifications and may struggle with accurate identification and tracking of food items in cluttered environments.

2.2.2 Computer Vision-based Solutions

With the advancements in computer vision and deep learning techniques, several researchers have explored the use of camera-based systems for smart refrigerator applications.

A smart refrigerator system that employs computer vision techniques to detect and identify food items was developed[45]. The system uses object detection and classification algorithms to recognize different types of food, enabling automatic inventory management and grocery list generation. However, their work lacks specific details on the implementation, image capture methods, and accuracy of the food recognition algorithms. A computer vision-based system for monitoring refrigerator contents and providing person-

alized dietary recommendations was proposed [2, 49]. The system uses object detection and semantic segmentation models to identify and track food items, while also estimating their quantities based on segmented regions. The authors integrated this system with a recommendation engine that suggests meal plans and recipes based on the user's dietary preferences and available ingredients.

While computer vision-based solutions offer the advantage of non-intrusive monitoring and the potential for accurate food item recognition, they often face challenges related to handling diverse and cluttered refrigerator environments, as well as recognizing a wide range of food items with varying appearances and packaging.

The FridgeVision system aims to address these limitations by leveraging cutting-edge computer vision and deep learning techniques, specifically tailored for the unique requirements of individuals with dementia and their caregivers. By combining advanced object detection, semantic segmentation, and latent space analysis, FridgeVision can accurately identify, organize, and monitor food items within the refrigerator, enabling precise tracking of food storage and consumption patterns.

It is important to note that while previous research efforts have explored various approaches to developing smart refrigerator solutions, many of these works focus on specific aspects or functionalities, often lacking a comprehensive and integrated approach tailored to the unique needs of individuals with dementia. Additionally, user-centric design considerations and accessibility for individuals with cognitive impairments are often overlooked. By addressing these gaps and leveraging the latest advancements in computer vision, deep learning, and user experience design, the FridgeVision project has the potential to make a significant contribution to the field of assistive technologies for dementia care, ultimately enhancing the quality of life for affected individuals and alleviating the burden on their caregivers.

2.3 Computer Vision and Deep Learning

Computer vision is a rapidly evolving field of artificial intelligence that aims to enable machines to perceive, analyze, and interpret visual data from the world around them,

mimicking and extending human vision capabilities. The advent of deep learning, particularly the application of convolutional neural networks (CNNs), has revolutionized computer vision, enabling remarkable advancements in tasks such as object detection, semantic segmentation, image classification, and scene understanding [25, 22].

2.3.1 Object Detection

Object detection refers to the computer vision task of identifying and localizing semantic objects of interest within an image or video frame. It enables a richer understanding of visual scenes by not just classifying the image but also pinpointing spatial occurrences of target categories. Significant progress has been made in advancing object detection capabilities in recent years through deep convolutional neural networks.

Earlier approaches relied on sliding window classifiers and handcrafted features for detection. Algorithms like DPM (Deformable Parts Model) demonstrated successful results by modelling objects through component parts [8]. However, performance was limited due to reliance on low-level features. The breakthrough work on R-CNN (Regions with CNN Features), an augmented object proposals with CNN-based features, substantially improving accuracy [10].

Another influential work is the Faster R-CNN [36] model, which employs a two-stage approach combining a region proposal network (RPN) with a fast R-CNN detector. Faster R-CNN[36] and LiteFCN[49] has demonstrated excellent performance on various benchmark datasets and has been widely adopted in computer vision applications, including food recognition [40].

One of the pioneering works in this domain is the YOLO (You Only Look Once) framework [35]. Figure 2.1 shows a comparison of different YOLO models. YOLO is a one-stage object detection approach that simultaneously predicts bounding boxes and class probabilities in a single evaluation, enabling real-time performance. Its high accuracy and low latency make it well-suited for applications like FridgeVision, which require efficient object localization and recognition. More recently, transformer-based architectures, such as DETR [51] and Deformable DETR [52], have emerged as promising object detection

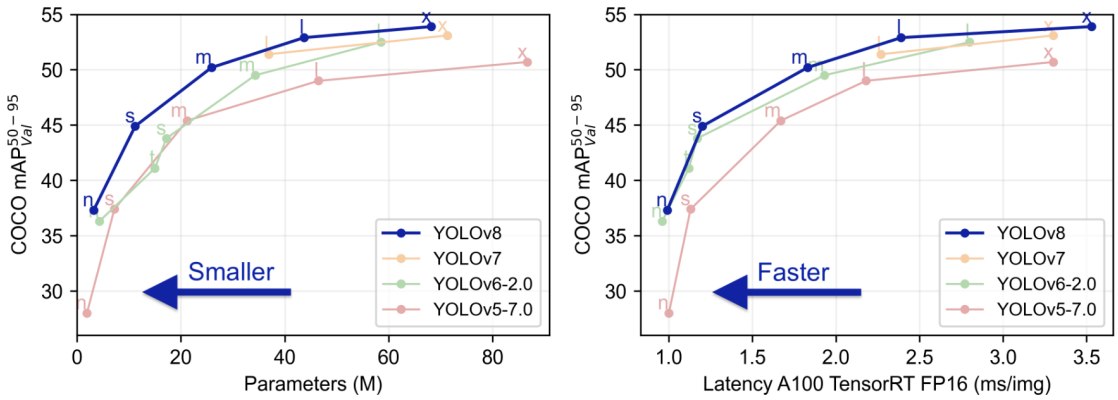

Figure 2.1: Performance comparison of YOLOs[16]

frameworks. These models leverage the power of self-attention mechanisms and transformer encoders, offering improved performance and adaptability to complex scenes.

2.3.2 Semantic Segmentation

While object detection provides bounding box localisations, semantic segmentation extends this task by assigning a class label to every pixel in an image. This capability is particularly valuable for the FridgeVision system, as it enables precise delineation and separation of individual food items, even in cases of occlusion or close proximity. Significant advances in deep learning-based image segmentation have enabled parsing visual scenes at instance, semantic, panoptic and prompt levels. Example is shown in the Figure 2.2.

Mask R-CNN [13] is a popular instance segmentation model that extends the Faster R-CNN framework by adding a parallel branch for predicting segmentation masks for each detected object instance. It has been widely adopted in various computer vision applications, including food item segmentation.

The recent semantic segmentation models, DeepLabV3+ [4] and OCRNet [50], leverage spatial context modelling through atrous convolutions and pyramid pooling to accurately classify image pixels into categories. Its ability to capture fine-grained details makes it a promising candidate for the FridgeVision system. Panoptic segmentation [20] unified both tasks into a sin- gle framework, combining instance segmentation (isolating object

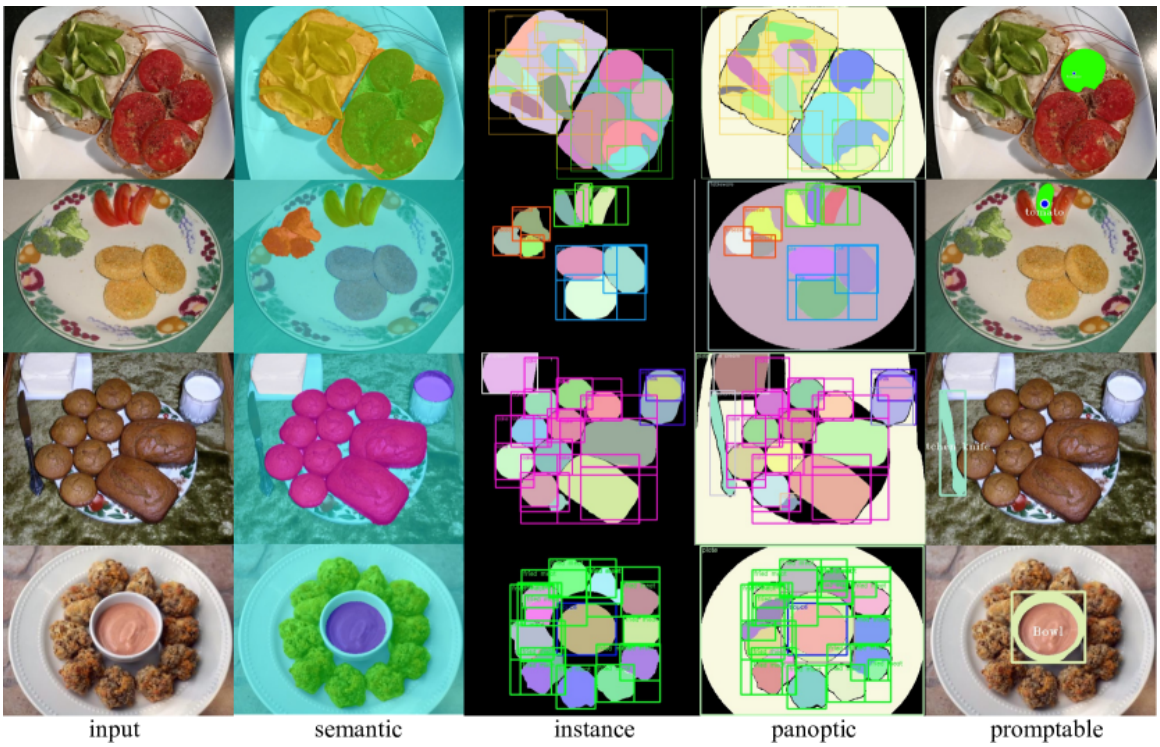

Figure 2.2: showing the difference among semantic segmentation, instance segmentation, panoptic segmentation, prompt segmentation[24]

instances) and semantic segmentation (labelling pixel categories). Axial-DeepLab [44] demonstrates state-of-the-art results on this unified task. The recently proposed Segment Anything Model (SAM) [21] is a transformer-based model that can segment any object or region in an image, given a single prompt point or mask. It emerged as a breakthrough, achieving new benchmarks across multiple datasets at instance, semantic, panoptic and prompt levels. Its flexibility and high accuracy make it an attractive option for the FridgeVision system, as it can potentially handle a wide variety of food items and refrigerator configurations without extensive training data.

2.3.3 Food Recognition and Tracking

While not specific to the domain of dementia care, several research efforts have explored the application of computer vision and deep learning techniques for food recognition and tracking, which is a core component of the FridgeVision system. [40] [49] developed a food recognition system using Faster R-CNN and a custom dataset of food images, achieving promising results in accurately detecting and classifying various food items.

A smart refrigerator system that employs computer vision techniques to detect and identify food items, enabling automatic inventory management and grocery list generation, was developed [45]. However, their work lacks specific details on the algorithms and implementation.

2.3.4 Latent Space Analysis

Latent space analysis and autoencoder frameworks are highly relevant to the FridgeVision system's objective of tracking changes in refrigerator contents over time. By encoding the visual information obtained from object detection and segmentation into a compact latent space representation, the system can efficiently analyze and compare previous and current states to quantify additions, removals, and consumption patterns.

Variational Autoencoders (VAEs) [19] and their variants have been widely used for latent space encoding and generation tasks, offering a powerful framework for learning condensed representations of high-dimensional data.

More recently, diffusion models [15] and their latent space representations have shown promising results in various computer vision tasks, including image generation, inpainting, and manipulation [37]. These techniques could potentially be leveraged in the FridgeVision system for efficient encoding and analysis of refrigerator contents.

Additionally, techniques like contrastive learning [5] and self-supervised learning [12] have emerged as effective methods for learning robust and informative representations from unlabeled data, which could be valuable for the FridgeVision system, particularly in scenarios where labeled data is limited or expensive to obtain.

By leveraging these state-of-the-art computer vision and deep learning techniques, the FridgeVision system aims to push the boundaries of assistive technologies for individuals with dementia, providing a comprehensive and user-friendly solution for food management and nutrition assistance.

It is important to note that while significant progress has been made in these areas, the application of these techniques to the specific domain of assisting individuals with dementia in managing their refrigerator contents and nutrition poses unique challenges.

These include handling diverse and cluttered refrigerator environments, recognizing a wide range of food items with varying appearances, and designing user-friendly interfaces that cater to the cognitive and perceptual abilities of the target demographic.

The FridgeVision project addresses these challenges through a multidisciplinary approach that combines cutting-edge computer vision and deep learning methods with user-centric design principles and insights from healthcare professionals and subject matter experts. By leveraging the latest advancements in artificial intelligence and tailoring the solution to the specific needs of individuals with dementia and their caregivers, FridgeVision has the potential to deliver a practical and impactful solution that enhances independence, promotes proper nutrition, and alleviates caregiver burden.

Chapter 3

Methodology

The FridgeVision system represents an innovative approach to assisting individuals with dementia in managing their food inventory and maintaining proper nutrition by leveraging state-of-the-art computer vision, deep learning, voice interaction, and IoT technologies. This chapter presents the methodology employed in the development and evaluation of the FridgeVision system. This chapter provides a comprehensive overview of the methodology employed in the development and evaluation of the FridgeVision system, encompassing a multi-faceted pipeline that includes system architecture, image processing, object detection using YOLOv8 [16], segmentation with SAM[21], latent comparison using ResNet18 [14] and cosine similarity, data augmentation techniques, LLM-based recipe recommendation, and the data collection process.

The primary objective of the FridgeVision system is to accurately identify and track food items within a refrigerator, providing personalized assistance and recommendations to individuals with dementia and their caregivers. This is achieved through the meticulous design and implementation of a sophisticated pipeline that seamlessly integrates advanced computer vision algorithms, voice interaction capabilities, and IoT technologies. By harnessing these cutting-edge techniques, the FridgeVision system aims to provide a reliable and user-friendly solution to the challenges faced by individuals with dementia in maintaining a healthy and well-organized food inventory. The methodology adopted in this research follows a systematic approach, ensuring the robustness, accuracy, and usability of the FridgeVision system. Each stage of the pipeline has been meticulously designed

and implemented, taking into account the unique challenges and requirements associated with assisting individuals with dementia in food management. From the initial image processing steps to the final recipe recommendations and voice interactions, every component of the system has been carefully crafted to address the specific needs of this vulnerable population.

Throughout this chapter, we will embark on a detailed exploration of each methodological component, delving into the theoretical foundations, implementation specifics, and evaluation metrics employed. The chapter will provide a comprehensive account of the system architecture, highlighting the integration of various technologies such as advanced deep CNNs for object detection and segmentation, latent space analysis, data augmentation techniques, LLM-based recipe recommendation, and the data collection process. It will also provide a detailed overview of the image processing techniques utilised to enhance the quality and clarity of the captured refrigerator images, laying the foundation for subsequent stages of analysis.

In addition, the chapter will discuss the contributions of Joshua Goulton, who has played a crucial role in the development of the FridgeVision system. Goulton's work focuses on the integration of voice interaction capabilities and the implementation of expiry date detection functionality. By leveraging Azure's Speech-to-Text and Text-to-Speech APIs, Goulton has enabled intuitive and accessible interactions for individuals with dementia and their caregivers. Moreover, his approach to expiry date detection, which involves guiding users to capture clear images of expiry dates through voice prompts, ensures accurate monitoring of food freshness and enhances the overall effectiveness of the FridgeVision system.

The insights gained from this comprehensive methodology have the potential to make significant contributions to the broader field of assistive technologies for individuals with dementia. By showcasing the effective integration of computer vision, deep learning, voice interaction, and IoT technologies, the FridgeVision system serves as a testament to the transformative potential of these techniques in addressing real-world challenges faced by this vulnerable population.

In the following sections, we will embark on a detailed exploration of each component of the methodology, starting with the system architecture and progressing through image processing, object detection, segmentation, latent comparison, data augmentation, LLM-based recipe recommendation, and the data collection process. Through this in-depth analysis, we aim to provide a clear and comprehensive understanding of the scientific approach underpinning the development and evaluation of the FridgeVision system, setting the stage for its potential impact on the lives of individuals with dementia and their caregivers.

3.1 System Architecture

The FridgeVision system architecture is designed as a comprehensive and integrated solution for assisting individuals with dementia in managing their food inventory and maintaining proper nutrition. In Figure 3.1, the system architecture encompasses a series of interconnected components that work in harmony to achieve accurate food item detection, segmentation, tracking, personalized recipe recommendation, and intuitive voice interactions tailored to the needs of dementia patients and their caregivers.

At the core of the FridgeVision system lies a modular and scalable architecture that leverages state-of-the-art computer vision, deep learning, and IoT technologies. The system architecture can be broadly divided into three main layers: the input layer, the processing layer, and the output layer. Each layer plays a crucial role in the overall functioning of the system, ensuring efficient data acquisition, analysis, and generation of actionable insights.

The input layer of the FridgeVision system is responsible for capturing high-quality images of the refrigerator interior and facilitating voice-based user interactions. This is achieved through the strategic placement of a camera module inside the refrigerator, providing a clear view of the food items stored within. The camera module is triggered at regular intervals or upon user initiation, capturing images that serve as the primary input for subsequent processing stages. Additionally, the input layer incorporates Joshua Goulton's work on integrating Azure's Speech-to-Text and Text-to-Speech APIs, enabling users to interact with the system verbally, making it accessible and user-friendly for individu-

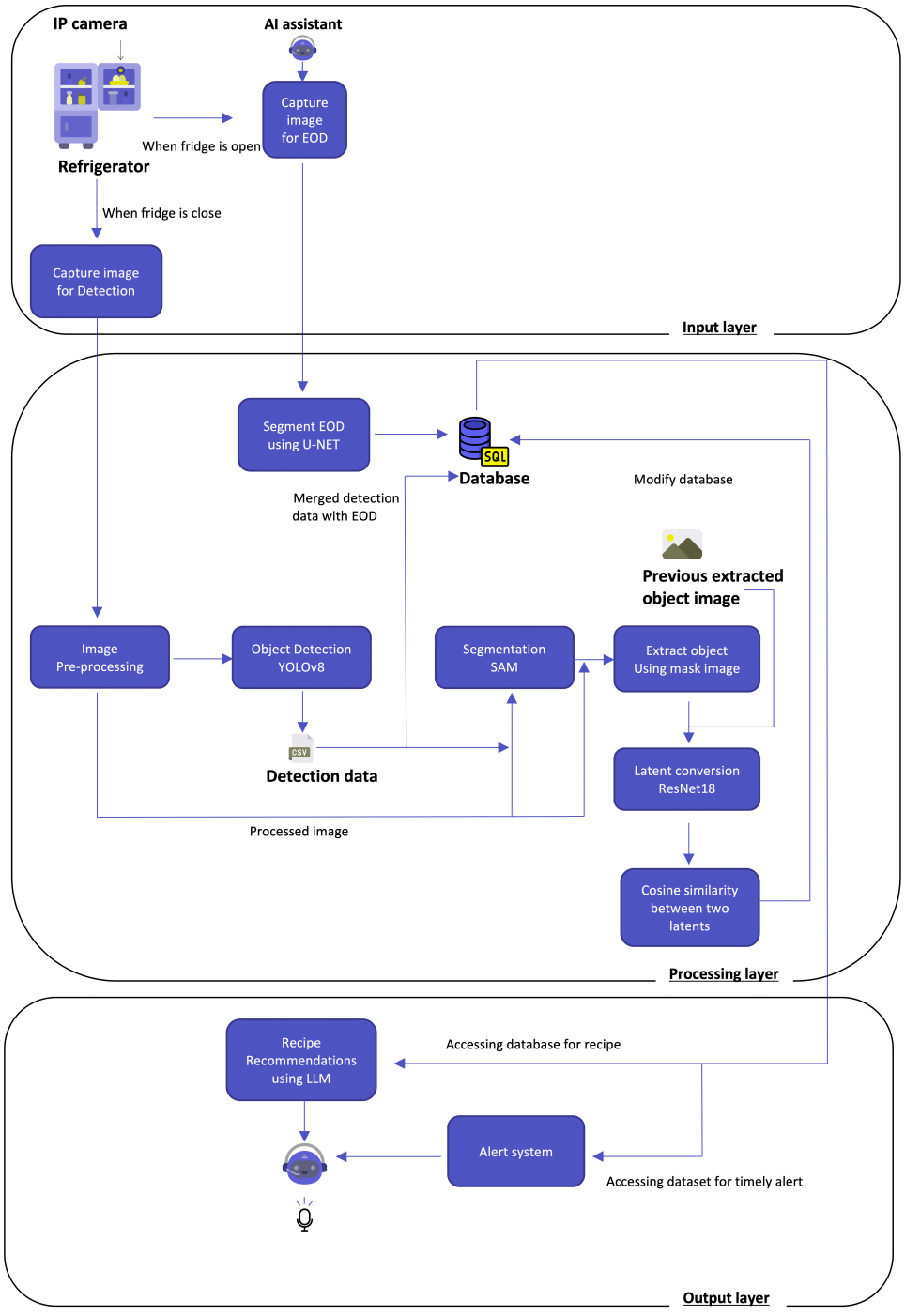

Figure 3.1: Architecture of FridgeVision

als with dementia and their caregivers. Goulton's approach also involves guiding users, through voice prompts, to position products in a way that allows the camera to capture

clear images of the expiry dates when users open the fridge to place new items inside.

Once the refrigerator images are acquired and user voice commands are processed, they are transmitted to the processing layer, which forms the backbone of the FridgeVision system. The processing layer consists of several key components that work in tandem to analyze the captured images, extract meaningful information, and generate personalized recommendations.

The first step in the processing pipeline is image pre-processing, where techniques such as image resizing, normalization, and enhancement are applied to ensure optimal quality and consistency for further analysis. The pre-processed images then undergo object detection using the YOLOv8 algorithm, a state-of-the-art deep learning model renowned for its accuracy and efficiency in identifying and localizing objects within an image. The YOLOv8 model[16] has been specifically trained on a diverse dataset of food items commonly found in refrigerators, enabling it to accurately detect and classify various food categories.

To further refine the understanding of the refrigerator's contents, the FridgeVision system employs the Segment Anything Model (SAM) [21] for precise segmentation of the detected food items. SAM is a cutting-edge segmentation model that excels in delineating the boundaries of objects at a pixel level, allowing for accurate isolation and extraction of individual food items from the image. By combining the outputs of YOLOv8 and SAM, the FridgeVision system achieves a detailed and comprehensive representation of the food items present in the refrigerator.

The segmented food items are then subjected to latent comparison using the ResNet18 model, a deep convolutional neural network architecture known for its robust feature extraction capabilities. The ResNet18 [14] model generates latent representations of the segmented food items, capturing their essential characteristics in a compact and informative format. These latent representations serve as the basis for tracking changes in the refrigerator's contents over time.

To quantify the differences between previous and current refrigerator states, the FridgeVision system employs the cosine similarity metric. By comparing the latent representations of food items across different time points, the system can identify additions, removals,

and consumption patterns, providing valuable insights into the user's food management habits and potential areas for improvement.

The processing layer of the FridgeVision system also incorporates a data augmentation module, which applies techniques such as noise injection, blurring, tilting, and fog simulation to the captured images. These augmentations help to enhance the robustness and generalization capabilities of the object detection and segmentation models, ensuring their effectiveness in handling diverse and challenging real-world scenarios.

Joshua Goulton's work on expiry date detection is seamlessly integrated into the processing layer of the FridgeVision system by leveraging advanced OCR techniques. His approach for capturing image allows the camera to capture clear images of the expiry dates. This interactive process ensures that the system accurately records the expiry date information and stores it in the dataset for further processing and monitoring. His work focuses on robust segmentation of expiry date regions using a U-Net architecture with a ResNet34 backbone. His work enables the FridgeVision system to extract and track expiry dates, providing timely alerts and recommendations to users regarding the freshness and safety of their food items.

The output layer of the FridgeVision system is responsible for generating actionable insights, personalized recommendations, and user alerts based on the analyzed data. A key component of this layer is the Large Language Model (LLM) for recipe recommendation. The LLM takes as input the list of detected food items, their quantities, and expiry dates, and generates tailored recipe suggestions that align with the user's preferences and dietary requirements. By leveraging the power of natural language processing, the LLM can provide creative and diverse recipe ideas, encouraging individuals with dementia to engage in meal preparation activities and maintain a balanced diet.

To ensure the scalability and modularity of the FridgeVision system, the architecture follows a loosely coupled design principle. Each component of the system is developed as an independent module, communicating with other components through well-defined interfaces. This modular approach allows for easy integration of new features, updates, and improvements without disrupting the overall functionality of the system.

Furthermore, the FridgeVision system architecture incorporates robust data management and storage mechanisms to handle the large volumes of image data, extracted information, and generated insights. The system employs secure and efficient data storage solutions, such as containerized databases within the IoT environment, to ensure the integrity, accessibility, and confidentiality of the processed data. This containerized approach, hosted on a dedicated IoT platform like openHAB, provides an additional layer of security and privacy, mitigating the risks associated with cloud-based storage and ensuring that sensitive user data remains within the local network.

As part of the future development plans for the FridgeVision system, the architecture will include a user interface component, which will serve as the primary point of interaction between the user and the system. The user interface will present the analyzed data, generated insights, and personalized recommendations in a clear and intuitive manner, utilizing visual elements such as graphs, charts, and images to facilitate easy comprehension. The interface will also allow users to provide feedback, set preferences, and access additional resources related to food management and nutrition. This future enhancement will further improve the usability and accessibility of the FridgeVision system for individuals with dementia and their caregivers.

The architecture's input, processing, and output layers work seamlessly together to acquire high-quality refrigerator images, analyze the contents using advanced object detection and segmentation models, track changes over time, and generate personalized recipe recommendations and user alerts. The integration of Joshua Goulton's work on expiry date detection and voice interaction further enhances the system's usability and accessibility for individuals with dementia and their caregivers.

Through its robust data management mechanisms and integration with IoT platforms, the FridgeVision system provides a reliable, secure, and accessible tool for individuals with dementia and their caregivers. The proposed future development of a user-centric interface will further enhance the system's usability and accessibility, ultimately promoting independence, well-being, and quality of life for individuals with dementia.

3.2 Image Pre-processing

Image pre-processing is a crucial step in the FridgeVision system, as it ensures that the captured images are of optimal quality and consistency for subsequent analysis. The pre-processing pipeline employs various techniques to enhance the visual characteristics of the images, making them more suitable for object detection, segmentation, and latent comparison. This section will delve into the details of the image pre-processing techniques implemented in the FridgeVision system.

To facilitate the integration of the FridgeVision system with IP cameras, the function `get_image_from_ip_camera` is implemented. This function takes the URL of an IP camera as input and retrieves the image data from the camera using HTTP requests. The retrieved data is then decoded into an image array using OpenCV's imdecode function. This functionality allows the FridgeVision system to seamlessly integrate with various IP camera models and capture images remotely. The code snippet below demonstrates the implementation of the `get_image_from_ip_camera` function:

Function `GetImageFromIpCamera(`url`):`
  $response \leftarrow$ requests.get(url);
  $image\_array \leftarrow$ np.array(bytearray($response.content$), dtype=np.uint8);
  $image \leftarrow$ cv2.imdecode($image\_array, -1$);
  return $image$;

Algorithm 1: Get Image from IP Camera

The first step in the preprocessing pipeline is image resizing. The captured images from the refrigerator's camera module may have varying dimensions, depending on the camera's resolution and settings. To ensure consistent input for the subsequent computer vision models, the images are resized to a fixed dimension of 640x640 pixels. The `resize_image` function is implemented to perform this resizing operation, as shown in the code snippets below:

Function
`ResizeImage(`$image, width, height, interpolation = cv2.INTER\_LINEAR$`):`
  $resized\_image \leftarrow cv2.resize(image, (width, height), interpolation = interpolation)$;
  return $resized\_image$;

Algorithm 2: Resize Image

The `resize_image` function takes an input image along with the desired width and height parameters, which are set to 640 for both dimensions in the FridgeVision system. It utilizes the cv2.resize function from the OpenCV library to resize the image to the specified dimensions. The resizing operation is performed using interpolation techniques, such as bilinear or bicubic interpolation, to maintain the aspect ratio and minimize distortion.

After resizing, the preprocessed images undergo various enhancement techniques to improve their overall quality and visual appeal. These techniques include adjusting brightness, contrast, shadows, and highlights. The `adjust_brightness_contrast` function is implemented to modify the pixel intensities of the image based on the provided brightness and contrast adjustment values. The code snippet below demonstrates the implementation of the brightness and contrast adjustment:

Function `AdjustBrightnessContrast`$(image, brightness = 30, contrast = 30)$**:**
  if $brightness \neq 0$ **then**
    if $brightness > 0$ **then**
      $shadow \leftarrow brightness$;
      $highlight \leftarrow 255$;
    else
      $shadow \leftarrow 0$;
      $highlight \leftarrow 255 + brightness$;
    end
    $alpha\_b \leftarrow (highlight - shadow)/255$;
    $gamma\_b \leftarrow shadow$;
    $image \leftarrow cv2.addWeighted(image, alpha\_b, image, 0, gamma\_b)$;
  end
  if $contrast \neq 0$ **then**
    $f \leftarrow \frac{131 \times (contrast + 127)}{127 \times (131 - contrast)}$;
    $alpha\_c \leftarrow f$;
    $gamma\_c \leftarrow 127 \times (1 - f)$;
    $image \leftarrow cv2.addWeighted(image, alpha\_c, image, 0, gamma\_c)$;
  end
  return *image*;

Algorithm 3: Adjust Brightness and Contrast

The `adjust_brightness_contrast` function is a Python function used to adjust the brightness and contrast of an input image. It takes three parameters: image, which is the input image, brightness, which controls the brightness adjustment, and contrast, which controls the contrast adjustment.

Inside the function, it first checks if the brightness parameter is non-zero. If it is, it computes the shadow and highlight values based on the brightness level provided. Then, it calculates the alpha and gamma values required for brightness adjustment using these shadow and highlight values. The function then applies the brightness adjustment to the image using the cv2.addWeighted function from the OpenCV library. Next, the function checks if the contrast parameter is non-zero. If it is, it computes the factors required for contrast adjustment based on the contrast level provided. It then applies the contrast adjustment to the image using the same cv2.addWeighted function.

The `enhance_image` function serves as a wrapper for additional image enhancement techniques. It converts the input image from the BGR color space to RGB, applies desired enhancements using the PIL (Python Imaging Library) library, and then converts the image back to BGR format. This function provides flexibility for incorporating further enhancement techniques based on specific requirements. The code snippet below shows the implementation of the `enhance_image` function:

Function `EnhanceImage`(*image*)**:**
  image_pil ←
   Image.fromarray(cv2.cvtColor(image, cv2.COLOR_BGR2RGB));
  image_enhanced ←
   cv2.cvtColor(np.array(image_pil), cv2.COLOR_RGB2BGR);
  return *image_enhanced*;

Algorithm 4: Enhance Image

Figure 3.2: Visual example of Image Pre-processing

The pre-processed images, resized to a fixed dimension of 640x640 pixels and enhanced

using the aforementioned techniques, undergo further analysis in the subsequent stages of the FridgeVision system. The YOLOv8 object detection model and the Segment Anything Model (SAM) rely on these pre-processed images to accurately detect and segment food items within the refrigerator. Moreover, Joshua Goulton's work on expiry date detection benefits from the pre-processing techniques, as they improve the clarity and contrast of the expiry date regions, facilitating accurate OCR using TesseractOCR. Visual example of Image Pre-processing is shown in Figure 3.2.

It is important to note that the pre-processing techniques applied in the FridgeVision system are not limited to the ones mentioned above. Depending on the specific requirements and characteristics of the captured images, additional pre-processing techniques can be incorporated. For example, noise reduction techniques like Gaussian blurring or median filtering can be applied to mitigate the effects of image noise. Color space transformations, such as converting images to gray-scale or applying color normalization, can be employed to enhance the robustness of the computer vision models.

In summary, image pre-processing plays a vital role in the FridgeVision system by enhancing the quality and consistency of the captured images. The modular design of the FridgeVision system's pre-processing pipeline allows for easy integration of new techniques and customization based on the unique needs of individuals with dementia and their caregivers. The pre-processed images serve as input for the subsequent computer vision models, including YOLOv8 for object detection, SAM for segmentation, and Joshua Goulton's work on expiry date detection using a U-Net architecture with a ResNet34 backbone and TesseractOCR. By continuously refining and adapting the pre-processing techniques, the FridgeVision system ensures optimal performance and usability for individuals with dementia.

3.3 Object Detection

Object detection is a fundamental component of the FridgeVision system, enabling the accurate identification and localization of food items within the refrigerator. The system employs the state-of-the-art YOLOv8 (You Only Look Once version 8)[16] object

detection model, renowned for its high accuracy and real-time performance. This section will provide an in-depth discussion of the object detection process implemented in the FridgeVision system, with a special focus on the YOLOv8 model and its architecture.
YOLOv8 is the latest iteration of the YOLO family of object detection models, which have revolutionized the field of computer vision. YOLO models are known for their ability to perform object detection in real-time while maintaining high accuracy. The key advantage of YOLO models is their ability to frame object detection as a regression problem, directly predicting bounding box coordinates and class probabilities from the input image in a single forward pass.

3.3.1 YOLOv8 Architecture

The architecture of YOLOv8 [16] builds upon the success of its predecessors while introducing several enhancements to improve performance and efficiency. The backbone of YOLOv8 is a convolutional neural network (CNN) that extracts features from the input image at multiple scales. Each of the convolutional layers outputs a specific 'feature map' or 'activation map'. These maps represent the presence of various high-level and low-level features throughout the image. Notably, YOLOv8 replaces the initial 6x6 convolutional layer with a 3x3 convolutional layer for improved feature extraction. The backbone network is designed to capture both low-level and high-level features, enabling the model to detect objects of varying sizes and complexities.
YOLOv8 introduces the C2f module, which efficiently combines high-level features with contextual information. This is accomplished by concatenating the outputs of bottleneck blocks, which are composed of two 3x3 convolutions with residual connections. This architectural improvement is intended to improve feature representation. One of the significant improvements in YOLOv8 is the introduction of the anchor-free detection head. Unlike previous YOLO versions that relied on predefined anchor boxes, YOLOv8 directly predicts the bounding box coordinates without the need for anchor priors.
Another key feature of YOLOv8 is the use of a novel loss function called the Complete Intersection over Union (CIoU) loss. YOLOv8 also incorporates a feature pyramid network

(FPN) to enhance its multi-scale detection capabilities. The FPN allows the model to detect objects at different scales by fusing features from multiple layers of the backbone network. This multi-scale representation enables YOLOv8 to handle objects of varying sizes effectively, from small items to large ones. Figure 3.3 shows the architecture of YOLOv8. The architecture of YOLOv8 consists of several key components:

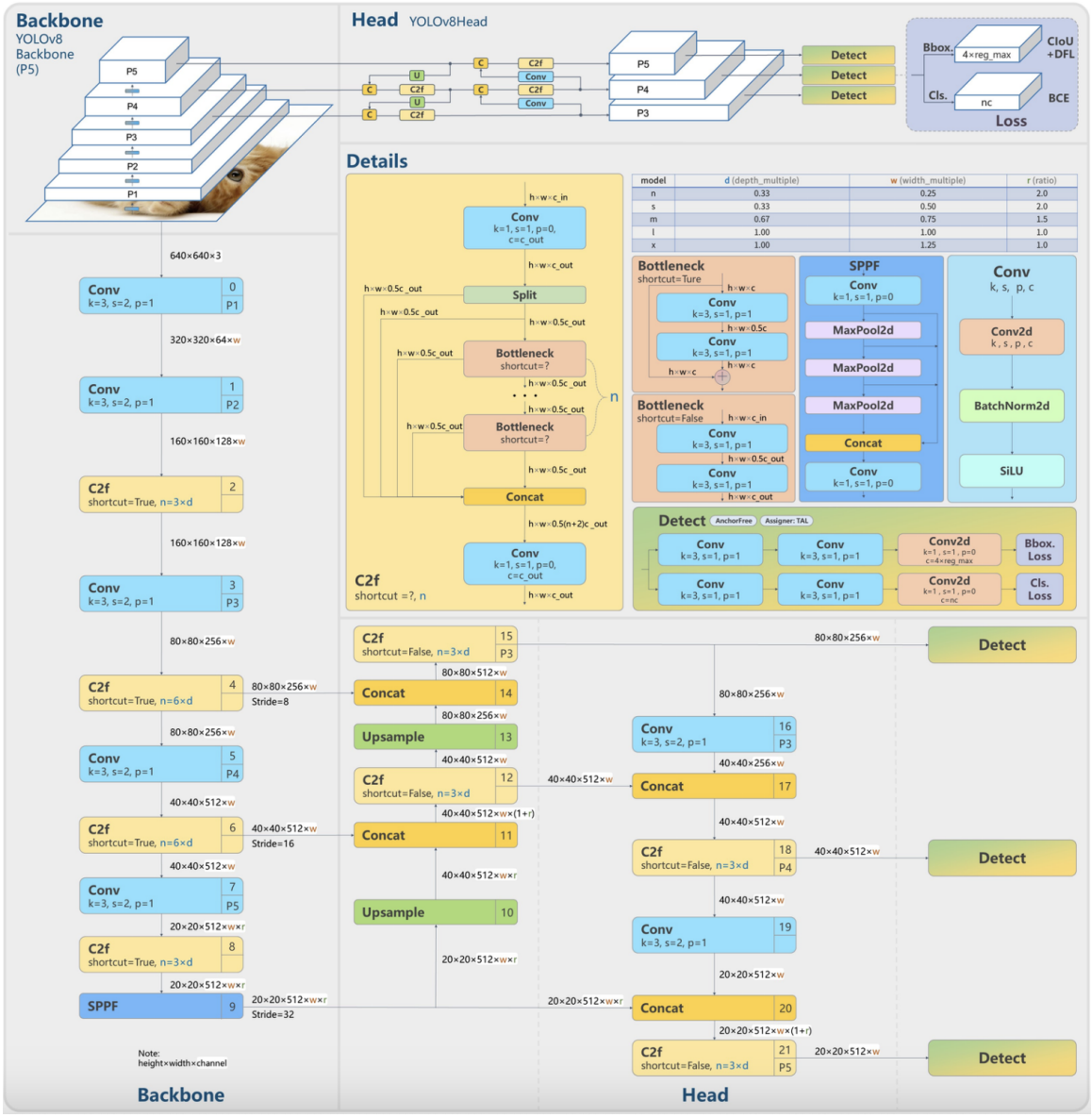

Figure 3.3: Architecture of YOLOv8 [16]

1. Backbone: The backbone of YOLOv8 is a CNN that extracts features from the input image. It typically uses a pre-trained network, such as CSPDarknet or EfficientNet, which has been proven to be effective in capturing rich and discriminative features.

The backbone network is designed to have a good balance between accuracy and computational efficiency.

2. Neck: The neck of YOLOv8 is responsible for aggregating and refining the features extracted by the backbone. It often employs techniques like feature pyramid networks (FPN) and path aggregation network (PAN) to enhance the multi-scale representation of features. The neck helps in improving the model's ability to detect objects at different scales and resolutions.

3. Head: The head of YOLOv8 is the part of the network that performs the actual object detection. It consists of several detection layers that predict bounding box coordinates, class probabilities, and objectness scores. YOLOv8 introduces an anchor-free detection head, which directly predicts the bounding box coordinates without relying on predefined anchor boxes. This simplifies the training process and reduces the computational overhead.

4. Independent Branches:YOLOv8 employs a decoupled head method, separating objectness, classification, and regression tasks into distinct branches. This architecture enables each branch to focus on a specialised job, improving detection accuracy.

5. Loss Function: YOLOv8 optimises the model using CIoU (Complete Intersection over Union) and DFL (Dynamic Focal Loss) loss functions for bound- ing box regression and binary cross-entropy for classification. These loss functions improve object detection, especially for tiny objects. The CIoU loss combines the benefits of both the Intersection over Union (IoU) loss and the Distance-IoU (DIoU) loss.

6. Activation Functions: YOLOv8 uses activation functions like the Sigmoid function for the objectness score and the Softmax function for class probabilities. These activation functions ensure that the predicted values are within the appropriate range and can be interpreted as probabilities.

7. Hyperparameters: YOLOv8 includes various hyperparameters that can be tuned to optimize the model's performance. Proper selection and tuning of these hyperpa-

rameters are crucial for achieving optimal detection accuracy and efficiency.

3.3.2 Implementation

In the FridgeVision system, the YOLOv8 model has been trained using a custom food item dataset. The specific model used is YOLOv8n.pt, which is a pre-trained model that has been further fine-tuned on the custom dataset. The custom dataset consists of a diverse collection of food items commonly found in refrigerators, along with their corresponding annotations. The fine-tuning process involves training the YOLOv8n.pt model on this dataset, allowing it to learn and adapt to the specific characteristics and visual patterns of the food items.

To improve model performance, YOLOv8 incorporates novel training strategies such as mosaic augmentation. During training, YOLOv8 combines four photos to encourage the model to learn item contexts in diverse positions and against varying backdrops. However, this augmentation is disabled during the final ten training sessions to avoid any performance loss. Also, to enhance model learning, we have transferred pre- trained model weights that were trained on the coco dataset using transfer learning by freezing some initial layers. Hyperparameter tuning was carried out by considering several parameters and determining which ones produced the greatest results, as shown in Table 3.1.

Table 3.1: Hyperparameter Settings

Hyperparameter	**Value**
Epochs	100
Batch Size	16
Learning Rate	0.01
Weight Decay	0.0005
Optimizer	Adam
Momentum	0.937

Justification on hyperparameter:

1. Epochs = 100: The number of epochs determines how many times the model iterates over the entire training dataset during the training process. Setting the number of epochs to 100 allows the model to have sufficient exposure to the training data, enabling it to learn and capture the relevant features and patterns. A higher number

of epochs helps the model converge and reach a stable state, where it can effectively generalize to unseen data. However, it's important to strike a balance, as an excessively high number of epochs can lead to overfitting, where the model becomes too specialized to the training data and performs poorly on new data.

2. Batch Size = 16: The batch size determines the number of training samples processed in each iteration before updating the model's weights. A batch size of 16 is a commonly used value that provides a good balance between computational efficiency and model convergence. It allows the model to process multiple samples simultaneously, leveraging parallelism and utilizing the available hardware resources efficiently. A larger batch size can speed up the training process but may require more memory. Conversely, a smaller batch size may result in slower training but can be beneficial for memory-constrained environments.

3. Learning Rate = 0.01: The learning rate controls the step size at which the model's weights are updated during the optimization process. A learning rate of 0.01 is a reasonable starting point that allows the model to learn and adapt its weights effectively. It strikes a balance between the model's ability to converge towards the optimal solution and the risk of overshooting or diverging. If the learning rate is set too high, the model may take large steps and miss the optimal solution. On the other hand, if it is set too low, the model may converge slowly or get stuck in suboptimal regions.

4. Weight Decay = 0.0005: Weight decay is a regularization technique that adds a penalty term to the model's loss function, discouraging large weight values. A weight decay value of 0.0005 helps prevent overfitting by regularizing the model's weights and promoting simpler and more generalized solutions. It encourages the model to learn more robust and meaningful features rather than relying on specific noise or irrelevant patterns in the training data. The chosen value of 0.0005 is a commonly used default that has been found to work well in various computer vision tasks.

5. Optimizer = Adam: Adam (Adaptive Moment Estimation) is a popular optimization algorithm that adapts the learning rate for each model parameter based on its historical gradients. It combines the benefits of two other optimization algorithms: AdaGrad and RMSProp. Adam is known for its efficiency, fast convergence, and ability to handle sparse gradients and noisy data. It is well-suited for training deep learning models like YOLOv8, as it can effectively navigate complex loss landscapes and find good solutions.

6. Momentum = 0.937: Momentum is a technique used in optimization algorithms to accelerate convergence and overcome local minima. A momentum value of 0.937 is a commonly used default that has been found to work well in practice. It helps the optimizer maintain a certain level of velocity in the direction of the gradients, allowing it to overcome small bumps and plateaus in the loss landscape. The chosen momentum value balances the trade-off between the optimizer's responsiveness to new gradients and its ability to maintain a consistent direction

The `detect_objects` function (Algorithm 5) in the FridgeVision system leverages the power of YOLOv8 for accurate and real-time object detection. The function begins by loading the pre-trained YOLOv8 model using the specified `model_path`. The pre-trained model has been trained on a diverse dataset of food items commonly found in refrigerators, enabling it to detect and classify a wide range of food categories. It then retrieves the image from the IP camera using the `get_image_from_ip_camera` function, which utilizes HTTP requests to capture the image data from the provided URL.

Before feeding the image to the YOLOv8 model, several preprocessing steps are applied to ensure optimal detection performance. The image is first resized to a fixed dimension of 640x640 pixels using the `resize_image` function. This standardization of input size is crucial for maintaining consistency and compatibility with the YOLOv8 model's architecture. Next, the image undergoes brightness and contrast adjustments using the `adjust_brightness_contrast`function, which enhances the visual quality and clarity of the image. Finally, the `enhance_image` function is applied to perform additional image enhancements, such as color space conversions or further adjustments, based on specific

Function `DetectObjects`($url, model\_path, csv\_file\_path$)**:**

model ← YOLO(model_path);
image ← GetImageFromIpCamera(url);
frame ← ResizeImage(image, 640, 640);
frame ← AdjustBrightnessContrast(frame);
frame ← EnhanceImage(frame);
results ← model(frame, stream=True);
`// Open CSV file for writing`
open(csv_file_path, mode='w', newline='') writer ← csv.writer(file);
writer.writerow(['Item Name', 'X', 'Y', 'Width', 'Height', 'Confidence']);
foreach r *in* $results$ **do**
  boxes ← r.boxes;
  foreach box *in* $boxes$ **do**
    $x1, y1, x2, y2 \leftarrow box.xyxy[0].tolist()$;
    $height, width, \_ \leftarrow frame.shape$;
    $x1\_norm \leftarrow x1/width$;
    $y1\_norm \leftarrow y1/height$;
    $x2\_norm \leftarrow x2/width$;
    $y2\_norm \leftarrow y2/height$;
    print("Normalized Coordinates —¿",
     $[x1\_norm, y1\_norm, x2\_norm, y2\_norm]$);
    cv2.rectangle($frame$, $(int(x1), int(y1))$, $(int(x2), int(y2))$,
     $(255, 0, 255)$, 3);
    $confidence \leftarrow math.ceil((box.conf[0] * 100))/100$;
    print("Confidence —¿", $confidence$);
    $cls \leftarrow int(box.cls[0])$;
    print("Class name –¿", $classNames[cls]$);
    open(csv_file_path, mode='a', newline='') writer ← csv.writer(file);
    $class\_name \leftarrow classNames[cls]$;
    $x\_norm \leftarrow (x1\_norm + x2\_norm)/2$;
    `// Normalize center x-coordinate`
    $y\_norm \leftarrow (y1\_norm + y2\_norm)/2$;
    `// Normalize center y-coordinate`
    $w\_norm \leftarrow x2\_norm - x1\_norm$;
    $h\_norm \leftarrow y2\_norm - y1\_norm$;
    writer.writerow([$class\_name$,x_norm, y_norm,w_norm,
     h_norm,confidence]);
  end
end
print(f'Object detection results saved to csv_file_path");
return $frame$;

Algorithm 5: Detect Objects

requirements.

With the preprocessed image ready, it is passed to the YOLOv8 model for object detection. The model object is called with the frame as input, and the stream=True parameter is set to enable real-time detection. The YOLOv8 model processes the image and returns a list of detection results, which include bounding box coordinates, class labels, and confidence scores for each detected object.

The detection results are then iteratively processed to extract relevant information and save it to a CSV file. For each detected object, the bounding box coordinates (x1, y1, x2, y2) are obtained from the box.xyxy[0] tensor and converted to a list. These coordinates represent the top-left and bottom-right corners of the bounding box. To normalize the coordinates, the height and width of the input frame are retrieved, and the coordinates are divided by the respective dimensions. The normalized coordinates are then printed for debugging purposes.

To visualize the detected objects, the cv2.rectangle function is used to draw bounding boxes on the input frame. The function takes the frame, the top-left and bottom-right coordinates of the bounding box, the color of the rectangle (in this case, purple with RGB values (255, 0, 255)), and the thickness of the rectangle (set to 3 pixels).

The confidence score of each detection is obtained from box.conf[0] and rounded to two decimal places using the math.ceil function. The confidence score represents the model's level of certainty in the detection and is printed for informational purposes.

The class label of the detected object is obtained from box.cls[0], which is an integer value corresponding to the index of the class name in the classNames list. The classNames list contains the names of the object classes that the YOLOv8 model is trained to detect. The class name is then printed to provide a human-readable representation of the detected object.

Finally, the detection results are saved to the specified CSV file. The file is opened in append mode (mode='a') to allow multiple detections to be written to the same file. The class name, normalized center coordinates (x_norm, y_norm), normalized width (w_norm), normalized height (h_norm), and confidence score are written as a row in the CSV file

using the csv.writer object. The successful saving of the detection results is confirmed by printing a message indicating the file path.

The `detect_objects` function returns the input frame with the detected objects visualized using bounding boxes. This allows for further processing or display of the detection results in subsequent stages of the FridgeVision system.

The modular design of the object detection pipeline allows for easy integration with other components of the FridgeVision system, such as Joshua Goultan's work on expiry date detection using a U-Net architecture with a ResNet34 backbone. The detected objects serve as input for these subsequent stages, providing a solid foundation for further analysis and processing.

In summary, the object detection component of the FridgeVision system utilizes the state-of-the-art YOLOv8 model to accurately identify and localize food items within the refrigerator. YOLOv8 introduces several architectural enhancements, such as anchor-free detection, CIoU loss, and a feature pyramid network, to improve detection accuracy and efficiency. The architecture of YOLOv8 consists of a backbone network for feature extraction, a neck for feature aggregation and refinement, a head for object detection, and a novel loss function for improved localization accuracy. The `detect_objects` function encompasses the entire object detection pipeline, including image capture from an IP camera, preprocessing techniques like resizing and enhancement, YOLOv8 model inference, and saving the detection results to a CSV file. The detected objects are visualized using bounding boxes on the input frame, providing a visual representation of the refrigerator's contents. The object detection results form the basis for subsequent stages of the FridgeVision system, enabling personalized recommendations, expiry date tracking, and enhanced user interaction. By continuously refining and adapting the object detection component, the FridgeVision system ensures accurate and reliable performance in assisting individuals with dementia and their caregivers in managing their food inventory.

3.4 Segmentation

Segmentation is a critical component of the FridgeVision system, enabling the precise delineation and extraction of individual food items within the refrigerator. The system employs the state-of-the-art Segment Anything Model (SAM) [21]for this task, leveraging its powerful segmentation capabilities to achieve fine-grained pixel-level labeling and contouring of food items. This section will provide an in-depth discussion of the segmentation process implemented in the FridgeVision system, with a focus on the SAM architecture and its integration with the object detection results.

3.4.1 SAM Architecture

The Segment Anything Model (SAM) [21] is a cutting-edge segmentation model developed by Meta AI. It is designed to segment any object or region in an image, given a single prompt point or mask. As shown in Figure 3.4, SAM's architecture consists of three main components: an image encoder, a prompt encoder, and a mask decoder.

3.4.2 Architecture of SAM

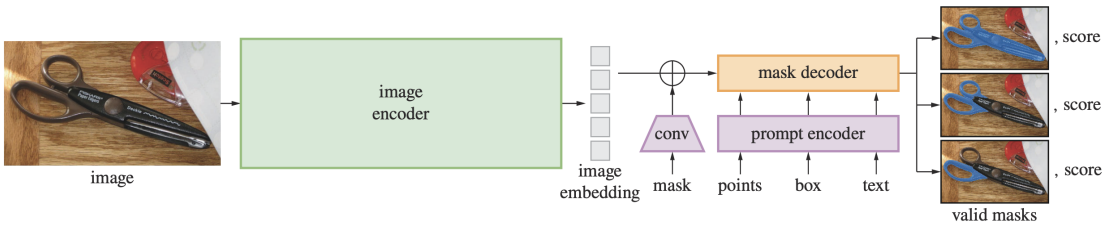

Figure 3.4: Architecture of SAM [21]

1. Image Encoder: The image encoder is responsible for extracting rich feature representations from the input image. It typically utilizes a transformer-based architecture, such as a Vision Transformer (ViT), to capture both local and global context. The image encoder takes the input image and outputs a dense feature map that encodes the visual information at different scales and resolutions.

2. Prompt Encoder: The prompt encoder takes the user-provided prompts, such as point coordinates or masks, and encodes them into a format compatible with the image features. It learns to interpret the prompts and generate a set of query embeddings that guide the segmentation process. The prompt encoder allows SAM to adapt to various types of user inputs and enables flexible and interactive segmentation.

3. Mask Decoder: The mask decoder takes the encoded image features and the prompt embeddings and generates the final segmentation masks. It typically employs a convolutional decoder architecture that gradually upsamples the feature maps while incorporating the prompt information. The mask decoder outputs a binary mask for each object or region specified by the prompts.

SAM's architecture is designed to be highly flexible and adaptable to different segmentation scenarios. It can handle a wide range of object categories and can generate accurate segmentation masks with just a single prompt point or mask. The model is trained on a large-scale dataset of annotated images, enabling it to learn rich feature representations and generalize well to unseen objects and scenes.

3.4.3 Implementation

In the FridgeVision system, the segmentation process is performed by the ObjectSegmenter class. The ObjectSegmenter takes an image path, a CSV file containing the YOLO object detection coordinates, and the path to the SAM model checkpoint as input. The class initializes by loading the image and the SAM model into memory.

The `load_image` (Algorithm 6) method reads the image from the specified path using OpenCV and converts it from the BGR color space to RGB.

The `load_image` method loads the SAM model from the provided checkpoint. It uses the `sam_model_registry` to obtain the default SAM model architecture and initializes it with the checkpoint weights. The model is then moved to the appropriate device (GPU if available, otherwise CPU) for efficient inference.

Function `Init`($image\_path, csv\_file\_path, model\_checkpoint$)**:**
Input: $image\_path$, csv_file_path, $model\_checkpoint$;
Output: Initialization of class attributes
$self.image\_path \leftarrow image\_path$;
$self.csv\_file\_path \leftarrow csv\_file\_path$;
$self.model\_checkpoint \leftarrow model\_checkpoint$;
$self.img \leftarrow$ `LoadImage()`;
if *torch.cuda.is_available()* **then**
  $self.device \leftarrow$ "*cuda*";
else
  $self.device \leftarrow$ "*cpu*";
end
$self.sam \leftarrow$ `LoadModel()`;

Algorithm 6: Initialize Segmentation Class

The `read_yolo_coordinates` method reads the object detection coordinates from the CSV file. It assumes that the CSV file contains columns named 'X', 'Y', 'Width', and 'Height', representing the normalized center coordinates and dimensions of the detected objects. The method returns a list of tuples containing the YOLO coordinates for each object.

The `yolo_to_absolute` method converts the normalized YOLO coordinates to absolute pixel coordinates. It takes the image dimensions (height and width) and the YOLO coordinates as input and returns a list of tuples containing the absolute coordinates of the object bounding boxes.

The `segment_objects` (Algorithm 7) method performs the actual segmentation using the SAM model. It initializes a SamPredictor with the loaded SAM model and sets the input image. It then reads the YOLO coordinates from the CSV file and converts them to absolute coordinates using the `yolo_to_absolute` method.

For each object bounding box, the `segment_objects` method calculates the center coordinates and uses them as a prompt point for the SAM model. It calls the predict method of the SamPredictor with the center coordinates and a point label of 1, indicating a positive object. The `multimask_output` parameter is set to True to obtain multiple segmentation masks for each prompt point.

The method then selects the best segmentation mask based on the highest score returned by the SAM model. The best mask for each object is appended to a list of masks.

Function `SegmentObjects(`img, sam`):`
- $predictor \leftarrow$ SamPredictor(sam);
- $predictor$.`SetImage(`img`)`;
- $yolo\_coords \leftarrow$ `ReadYoloCoordinates()`;
- $abs\_coords \leftarrow$ `YoloToAbsolute(`$yolo\_coords$`)`;
- $masks \leftarrow [\,]$;
- **foreach** $(x\_min, y\_min, width, height)$ *in* abs_coords **do**
  - $center\_x \leftarrow x\_min + \frac{width}{2}$;
  - $center\_y \leftarrow y\_min + \frac{height}{2}$;
  - $(predicted\_masks, scores, \_) \leftarrow predictor$.`Predict(`
    point_coords=np.array([center_x, center_y]), point_labels=np.array([1]),
    multimask_output=True`)` top_score $\leftarrow 0$;
  - $best\_mask \leftarrow None$;
  - **foreach** $(score, mask)$ *in* $(scores, predicted\_masks)$ **do**
    - **if** $score > top\_score$ **then**
      - $top\_score \leftarrow score$;
      - $best\_mask \leftarrow mask$;
    - **end**
  - **end**
  - **if** $best\_mask \neq None$ **then**
    - $masks$.append($best\_mask$);
  - **end**
- **end**
- $final\_image \leftarrow$ `ApplyMasksToImage(`$img, masks$`)`;
- `ShowImage(`$final\_image$`)`;
- $cv2.imwrite('final\_image.png', cv2.cvtColor(final\_image, cv2.COLOR\_RGB2BGR))$;

Algorithm 7: Segment objects method

Function `ApplyMasksToImage(`$img, masks$`):`
- $final\_image \leftarrow$ array of zeros with the same shape as img;
- **foreach** $mask$ *in* $masks$ **do**
  - $object\_mask \leftarrow$ stack mask along the last axis to match RGB channels;
  - $color \leftarrow$ random RGB color;
  - $object\_image \leftarrow$ element-wise product of $object\_mask$ and $color$;
  - $final\_image \leftarrow$ element-wise maximum of $object\_image$ and $final\_image$;
- **end**
- **return** $final\_image$;

Algorithm 8: Method for Applying Masks to Image

After obtaining the segmentation masks for all objects, the `apply_masks_to_image` method (Algorithm 8) is called to visualize the segmented objects on the input image. It creates a new image with the same dimensions as the input image and iterates over the segmentation masks. For each mask, it assigns a random color and applies the mask to the corresponding region in the image. The resulting segmented objects are overlaid on the original image.

Finally, the `show_image` method is used to display the segmented image, and the cv2.imwrite function is used to save the segmented image to a file named `'final_image.png'`. To further process the segmented objects and extract them from the background, the ObjectExtracter class is introduced. The ObjectExtracter takes an image path and a mask path as input. It loads the image and the corresponding segmentation mask obtained from the ObjectSegmenter.

The `segment_objects_with_precomputed_masks` method (Algorithm 9) in the ObjectExtracter class performs the object extraction using the precomputed segmentation masks. It loads the input image and the corresponding mask image, and then calls the `apply_masks_to_image` method to extract the objects.

The `apply_masks_to_image` method in the ObjectExtracter class takes the input image and the segmentation mask as input. It identifies the unique grayscale values in the mask, each representing a different object. For each unique value (except for the background value of 0), it creates a binary mask for the corresponding object and applies it to the input image. The resulting object image is assigned a random color and overlaid on a new image with a black background. This process is repeated for all objects, creating a final image that contains only the extracted objects with a black background.

The resulting image, containing only the extracted objects with a black background, is displayed using the `show_image` method and saved to a file named `'final_image.png'`.

The integration of the ObjectExtracter class in the FridgeVision system allows for the extraction of individual food items from the segmented image, providing a cleaner representation of the objects for further analysis and processing. This step is particularly useful for tasks such as latent space analysis, where the focus is on the individual objects

Function `SegmentObjectsWithPrecomputedMasks`($image\_path, mask\_path$)**:**
  $img \leftarrow cv2.imread(image\_path)$;
  if *img is None* **then**
    return Error: Image not found at the specified path;
  end
  $img \leftarrow cv2.cvtColor(img, cv2.COLOR\_BGR2RGB)$;
  $mask \leftarrow cv2.imread(mask\_path, cv2.IMREAD\_GRAYSCALE)$;
  if *mask is None* **then**
    return Error: Mask image not found at the specified path;
  end
  $final\_image \leftarrow$ `ApplyMasksToImage`($img, mask$);
  `ShowImage`($final\_image$);
  $cv2.imwrite('final\_image.png', cv2.cvtColor(final\_image, cv2.COLOR\_RGB2BGR))$;

Algorithm 9: Segment Objects with Precomputed Masks

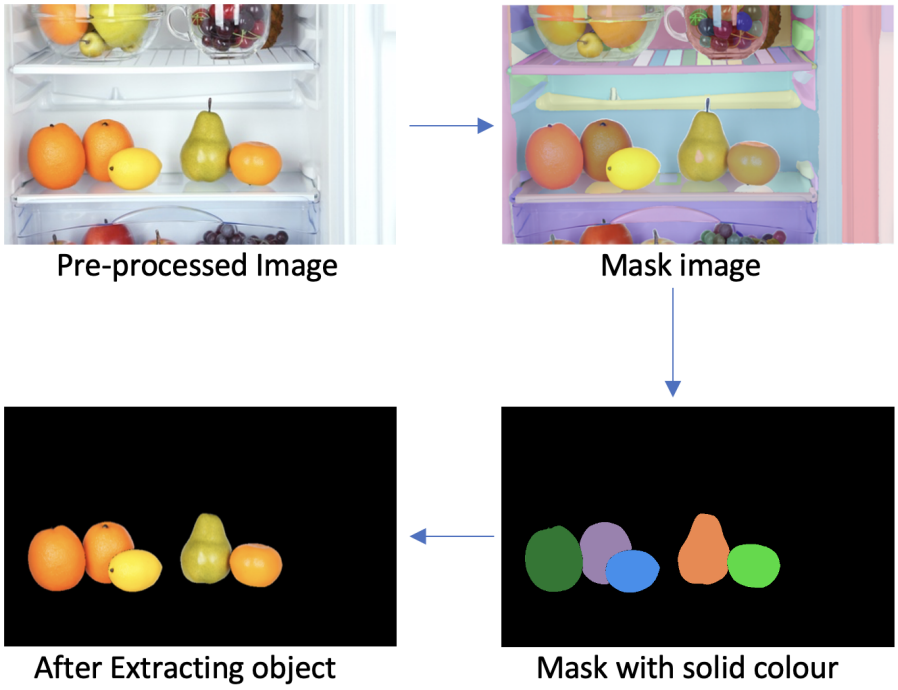

Figure 3.5: Visual Interpretation of segmentation process

rather than the entire refrigerator image. Figure 3.5 shows the visual interpretation of segmentation process.

In summary, the segmentation component of the FridgeVision system utilizes the SAM to achieve precise and fine-grained segmentation of food items within the refrigerator. The ObjectSegmenter class integrates SAM with the object detection results from YOLOv8,

focusing the segmentation on the regions of interest. The ObjectExtracter class further processes the segmented objects, extracting them from the background and providing a cleaner representation for subsequent analysis. The segmentation and extraction results form the basis for personalized recommendations and enhanced user interaction in the FridgeVision system. By continuously refining and adapting these components, the system ensures accurate and reliable delineation and isolation of food items, assisting individuals with dementia and their caregivers in managing their food inventory.

3.5 Latent Space Analysis

Latent space analysis is a crucial component of the FridgeVision system, enabling the tracking of changes in refrigerator contents over time. By comparing the latent representations of food items across different time points, the system can identify additions, removals, and consumption patterns, providing valuable insights into the user's food management habits. This section will delve into the latent space analysis methodology employed in the FridgeVision system, with a focus on the ResNet18 [14] architecture for image encoding and the use of cosine similarity for comparing latent representations.
The latent space analysis in the FridgeVision system is performed by the `ImageComparator` class. The `ImageComparator` takes an optional `model_checkpoint` parameter, allowing the use of a pre-trained model for image encoding. If no checkpoint is provided, the class initializes a ResNet18 model with pre-trained weights.

3.5.1 Architecture of ResNet18

ResNet18 [14] is a deep convolutional neural network architecture that has been widely used for various computer vision tasks, including image classification and feature extraction. In Figure 3.6, the architecture consists of 18 layers, including convolutional layers, batch normalization layers, and fully connected layers.
The key innovation of ResNet18 is the introduction of residual connections, which allow the network to learn residual functions and facilitate the training of deeper networks.

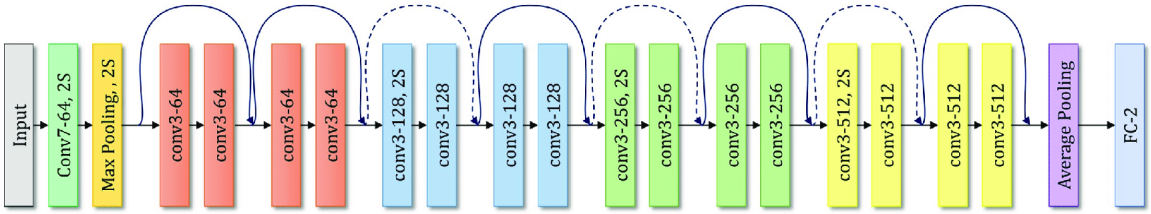

Figure 3.6: Architecture of ResNet18 [14]

Residual connections enable the network to learn the difference between the input and the desired output, rather than learning the entire mapping from scratch. This helps alleviate the vanishing gradient problem and allows for the training of much deeper networks.
The architecture of ResNet18 can be summarized as follows:

1. Convolutional Layer:The input image is passed through an initial convolutional layer with 64 filters of size 7x7 and a stride of 2. This layer learns to extract low-level features from the image.

2. Max Pooling Layer:The output of the convolutional layer is then passed through a max pooling layer with a kernel size of 3x3 and a stride of 2. This layer reduces the spatial dimensions of the feature maps.

3. Residual Blocks:The network then consists of four residual blocks, each containing two convolutional layers with 3x3 filters. The number of filters in each block is 64, 128, 256, and 512, respectively. These residual blocks learn to extract higher-level features from the input.

4. Global Average Pooling: After the residual blocks, a global average pooling layer is applied to reduce the spatial dimensions of the feature maps to a fixed size.

5. Fully Connected Layer:Finally, a fully connected layer is used to map the features

to the desired output classes.

The choice of ResNet18 for latent space analysis in the FridgeVision system is motivated by its ability to extract rich and discriminative features from images. The pre-trained weights of ResNet18, obtained from training on a large-scale dataset like ImageNet, enable the network to capture meaningful representations of objects and their characteristics [14]. The initialization of the ResNet18 model in the ImageComparator class is done as follows:

Listing 3.1: Initialization of model

```
1 if model_checkpoint:
2     self.model = torch.load(model_checkpoint)
3 else:
4     self.model = models.resnet18(pretrained=True)
5 self.model.eval()
```

This code snippet checks if a `model_checkpoint` is provided. If it is, the pre-trained model is loaded from the checkpoint. Otherwise, a new ResNet18 model with pre-trained weights is initialized using `models.resnet18(pretrained=True)`. The eval() function is called to set the model to evaluation mode, which is necessary for inference.

3.5.2 Implementation

The latent space analysis in the FridgeVision system utilizes the output images generated by the `ObjectExtracter` class. The `ObjectExtracter` class segments the objects from the refrigerator images and extracts them with a black background. These extracted object images serve as the input for the latent space analysis.

The `ImageComparator` class provides methods for loading images, cropping objects based on their coordinates, preprocessing images, and encoding images using the ResNet18 model. The `compare_objects` method is the main function for performing latent space analysis. It takes the paths to two `ObjectExtracter` output images (previous and current) and the path to the CSV file containing the object information. The method reads the object information from the CSV file and loads the two images.

For each object in the CSV file, the method crops the corresponding regions from both images using the `crop_object` method. The cropped images are then preprocessed using the `preprocess_image` method and encoded using the `encode_image` method, resulting in latent representations for each object.

To compare the latent representations of an object across the two images, the `compare_objects` method computes the cosine similarity between the encoded features. Cosine similarity measures the cosine of the angle between two vectors, providing a measure of their similarity. A cosine similarity of 1 indicates that the vectors are identical, while a similarity of -1 indicates that they are completely dissimilar.

The computation of cosine similarity is performed using the following function:

Listing 3.2: Cosine similarity

```
similarity = torch.nn.functional.cosine_similarity(feature1,
    feature2)
```

This code snippet computes the cosine similarity between the latent representations feature1 and feature2 of an object in the previous and current images, respectively. The computed cosine similarity for each object is compared against a specified threshold (e.g., 0.9) to determine if a significant change has occurred. If the cosine similarity falls below the threshold, the object coordinates are added to a list of changed areas. After comparing all objects, if any significant changes are detected, the `display_images_with_highlights` method is called to visualize the changes. This method displays the two images side by side and highlights the regions where significant changes have been identified using rectangular bounding boxes.

Once the latent space analysis process is completed, the previous ObjectExtracter output image is replaced with the current image. This allows the system to continuously track changes in the refrigerator contents over time by comparing the current image with the previous image from the previous run.

The latent space analysis component of the FridgeVision system plays a crucial role in providing insights into the user's food management habits and consumption patterns. By identifying additions, removals, and changes in food items, the system can generate

personalized recommendations and alerts, assisting individuals with dementia and their caregivers in maintaining a well-organized and properly managed food inventory.

It is important to note that while the ResNet18 model used in the FridgeVision system is pre-trained on a large-scale dataset, fine-tuning the model on a dataset specific to refrigerator environments and food items could further improve its performance and adaptability. Additionally, the choice of similarity threshold for detecting significant changes may need to be adjusted based on the specific requirements and characteristics of the user's refrigerator.

The modular design of the latent space analysis pipeline allows for easy integration with other components of the FridgeVision system, such as object detection and segmentation. The detected and segmented objects serve as input for the latent space analysis, enabling the system to focus on specific regions of interest and track changes at a granular level.

In summary, the latent space analysis component of the FridgeVision system utilizes the ResNet18 architecture for image encoding and cosine similarity for comparing latent representations. The `ImageComparator` class operates on the `ObjectExtracter` output images, which contain the segmented objects with a black background. By identifying significant changes in food items and visualizing them through highlighted regions, the system offers valuable insights into the user's food management habits. After each run, the previous `ObjectExtracter` output image is replaced with the current image, allowing for continuous tracking of changes over time. The latent space analysis pipeline forms an integral part of the FridgeVision system, contributing to personalized recommendations, alerts, and enhanced user interaction. Through continuous refinement and adaptation, the latent space analysis component ensures accurate and reliable tracking of changes in refrigerator contents, ultimately assisting individuals with dementia and their caregivers in maintaining a well-organized and properly managed food inventory.

3.6 Data Collection

The success of any machine learning-based system heavily relies on the quality and diversity of the data used for training and evaluation. In the FridgeVision system, a com-

prehensive and representative dataset is crucial for developing robust object detection, segmentation, and latent space analysis models. This section will provide an overview of the data collection methodology employed in the FridgeVision project, which involves a combination of web scraping, photographing fridge environments, and leveraging the Roboflow platform.

3.6.1 Web Scraping

Web scraping is a technique used to extract data from websites automatically. In the context of the FridgeVision system, web scraping was employed to gather a diverse collection of food item images from various online sources. By leveraging the power of web scraping, we were able to efficiently collect a large number of images spanning different food categories, brands, and packaging styles.

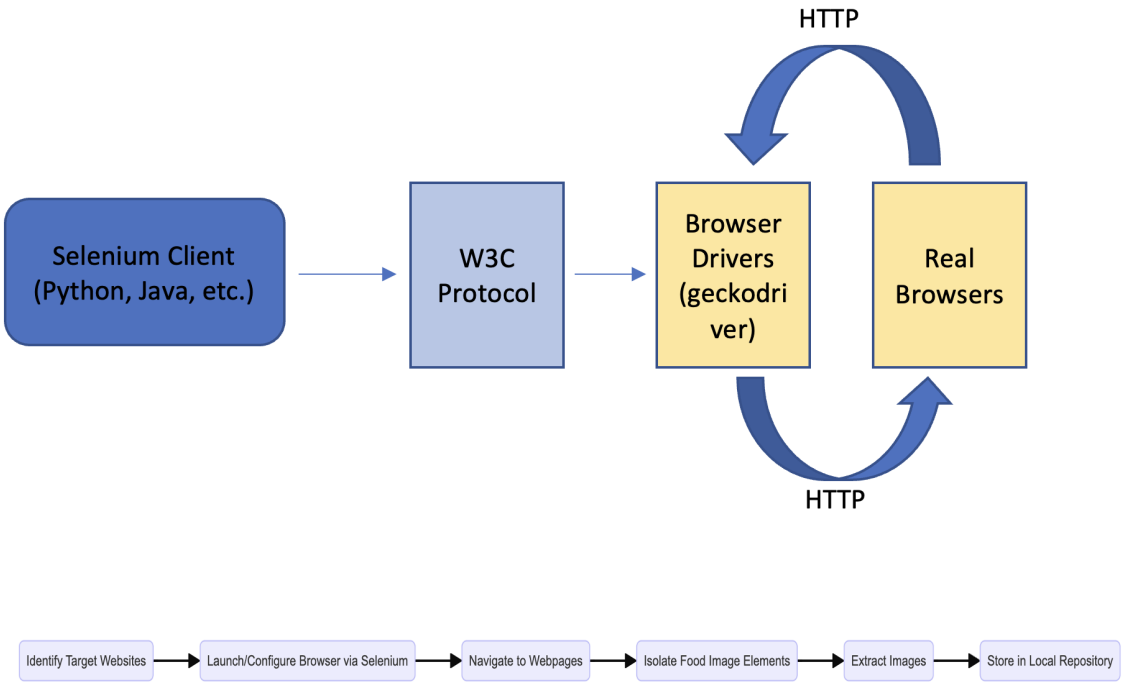

Figure 3.7: Web Scraping using Selenium

The web scraping process was implemented using Python and the Selenium library. Selenium is a popular web automation tool that allows for programmatic interaction with web pages, enabling the extraction of desired data. The web scraping script was designed

to navigate through targeted websites, locate relevant food item images, and download them along with their associated metadata, such as product names and categories. Figure 3.7 shows the process of scraping image using selenium

To ensure the quality and relevance of the scraped images, several data cleaning and filtering steps were applied. These steps included removing duplicate images, filtering out irrelevant or low-quality images, and organizing the collected data into appropriate categories. The resulting dataset obtained through web scraping provided a rich and diverse collection of food item images that could be used for training and testing the object detection and segmentation models in the FridgeVision system.

3.6.2 Photographing Fridge Environments

While web scraping provides a valuable source of food item images, it is equally important to capture the real-world context in which these items are stored—the fridge environment. To address this, a significant portion of the FridgeVision dataset was collected by photographing actual fridge environments.

The fridge environment images were captured using various camera devices, including smartphones, digital cameras, and even the camera modules intended for integration into the FridgeVision system. This diversity in image capture devices helped to simulate real-world scenarios and ensured the robustness of the developed models across different imaging conditions.

In addition to capturing the fridge environment images, the open-source tool RectLabel used for annotation. These annotations included bounding box coordinates, object labels, and any relevant metadata. The annotated fridge environment images served as valuable ground truth data for training and evaluating the object detection and segmentation models.

3.6.3 Roboflow Platform

To further enhance the dataset and streamline the data annotation process, the FridgeVision project leveraged the Roboflow platform. Roboflow is a cloud-based platform that

provides tools and services for managing, annotating, and augmenting computer vision datasets.

The fridge environment images collected through web scraping and photography were uploaded to the Roboflow platform. Roboflow's intuitive interface and annotation tools allowed for efficient labeling of objects within the images. The platform provided features such as bounding box drawing, polygon annotation, and label management, making the annotation process more streamlined and collaborative.

One of the key advantages of using Roboflow was its data augmentation capabilities. Roboflow also provided seamless integration with popular machine learning frameworks and tools, such as TensorFlow and PyTorch. This integration allowed for easy export of the annotated dataset in various formats, making it convenient to use the collected data in the model training and evaluation pipelines.

3.6.4 Dataset Composition and Statistics

The final FridgeVision dataset consisted of a combination of images obtained through web scraping, photographing fridge environments, and the Roboflow platform. The dataset encompassed a wide range of food items commonly found in refrigerators, including fruits, vegetables, dairy products, beverages, and packaged goods.

The dataset was carefully curated to ensure a balanced representation of different food categories and brands. Efforts were made to include images with varying levels of occlusion, lighting conditions, and object scales to simulate realistic fridge scenarios.

In total, the FridgeVision dataset comprised approximately 8840 images, with 34 unique food item categories. The dataset was split into training, validation, and testing subsets, following a ratio of 85:10:5. This split allowed for proper evaluation and fine-tuning of the models during the development phase.

3.6.5 Data Privacy and Ethics

Given the sensitive nature of fridge contents and the potential for personal information to be captured in the images, data privacy and ethical considerations were paramount

throughout the data collection process. Steps were taken to ensure the anonymity of individuals and households involved in the fridge environment photography.

The FridgeVision project adhered to the principles of responsible data collection and usage, ensuring that the collected data was used solely for the purpose of developing and evaluating the system. Access to the dataset was restricted to authorized project members and collaborators, and appropriate security measures were implemented to protect the data from unauthorized access or misuse.

3.6.6 Continuous Data Collection and Expansion

The initial FridgeVision dataset served as a strong foundation for developing the system. However, it is important to recognize that the dataset is not exhaustive and may not cover all possible food items or fridge configurations. To address this limitation and ensure the long-term robustness of the FridgeVision system, a continuous data collection and expansion strategy was adopted.

As the FridgeVision system is deployed and used by individuals with dementia and their caregivers, there is an opportunity to collect additional data from real-world usage scenarios. With appropriate consent and privacy measures in place, the system can gather new fridge images and annotations, which can be used to further refine and expand the dataset.

This continuous data collection approach allows for the incorporation of a wider range of food items, packaging variations, and fridge environments over time. It enables the FridgeVision system to adapt and improve its performance based on real-world data, ensuring its effectiveness and reliability in assisting individuals with dementia.

Moreover, the expanded dataset can be leveraged to fine-tune and update the object detection, segmentation, and latent space analysis models periodically. This iterative refinement process helps to maintain the system's accuracy and robustness as new food products and packaging designs emerge in the market.

In summary, the data collection methodology employed in the FridgeVision project involved a combination of web scraping, photographing fridge environments, and leveraging

the Roboflow platform. The resulting dataset comprised a diverse and representative collection of food item images, along with their corresponding annotations. The dataset was carefully curated, considering data privacy and ethical concerns, and served as the foundation for training and evaluating the machine learning models in the FridgeVision system. A continuous data collection and expansion strategy was adopted to ensure the system's long-term robustness and adaptability to real-world scenarios. By actively involving individuals with dementia and their caregivers in the data collection process, the FridgeVision project aims to create a comprehensive and evolving dataset that supports the development of effective assistive technologies for food management and independent living.

3.7 Data Augmentation

Data augmentation is a useful technique in computer vision that involves applying various transformations to the existing dataset to increase its size and diversity. By creating modified versions of the original images, data augmentation helps to improve the robustness and generalization capabilities of the trained models. In the FridgeVision system, data augmentation techniques were employed to enhance the performance of the YOLOv8 object detection model. This section will discuss the specific data augmentation methods applied to the image dataset used for training the YOLOv8 model. As discussed in last section, Roboflow offered a range of augmentation techniques, including image rotation, flipping, scaling, and color adjustments. By applying these augmentations to the collected images, we were able to expand the dataset and introduce variability, which helped to improve the generalization abilities of the trained models.

3.7.1 Blur Augmentation

Blurring is a common data augmentation technique that simulates the effect of camera blur or motion blur in images. In the FridgeVision system, blur augmentation was applied to the training images to make the YOLOv8 model more resilient to blurry inputs. The

blur augmentation was implemented using a Gaussian blur filter with a maximum kernel size of 1.5 pixels.

The Gaussian blur filter works by convolving the image with a Gaussian kernel, which is a matrix of weights that follows a Gaussian distribution. The kernel size determines the extent of the blur effect, with larger kernel sizes resulting in more significant blurring. By randomly applying Gaussian blur to the training images, the YOLOv8 model learns to detect objects even in the presence of slight blurriness.

Blur augmentation helps to simulate real-world scenarios where the fridge images may be captured under suboptimal conditions, such as low light or motion. By training the model on blurred images, it becomes more robust and capable of accurately detecting objects even when the input images are not perfectly sharp.

3.7.2 Noise Augmentation

Noise augmentation involves adding random noise to the training images to simulate the effect of sensor noise or image compression artifacts. In the FridgeVision system, noise augmentation was applied to a maximum of 0.94% of the pixels in each training image. The noise augmentation technique used was additive Gaussian noise, which adds random values drawn from a Gaussian distribution to the pixel intensities of the image. The amount of noise added to each pixel is controlled by the standard deviation of the Gaussian distribution. By randomly selecting a small percentage of pixels and adding Gaussian noise to them, the YOLOv8 model learns to handle noisy inputs and becomes more resistant to image degradation.

Noise augmentation helps to improve the model's robustness by simulating the imperfections and variations that can occur during image acquisition or transmission. By training the model on noisy images, it becomes better equipped to handle real-world scenarios where the fridge images may contain some level of noise or artifacts.

3.7.3 Mosaic Data Augmentation

Mosaic data augmentation is a technique specific to the YOLO family of object detection models, including YOLOv8. It involves combining multiple training images into a single mosaic image, where each image occupies a random quadrant of the mosaic. The mosaic image is then used as a single training sample for the model.

In the FridgeVision system, mosaic data augmentation was applied during the training of the YOLOv8 model. The mosaic images were generated by randomly selecting four training images and arranging them in a 2x2 grid. The bounding box annotations of the objects in each image were adjusted accordingly to maintain their relative positions within the mosaic.

Mosaic data augmentation offers several benefits. Firstly, it increases the effective batch size during training, as each mosaic image contains information from multiple original images. This allows the model to learn from a more diverse set of objects and their spatial relationships within a single training iteration.

Secondly, mosaic augmentation helps to improve the model's ability to handle object scale variations. By combining images with different object scales in a single mosaic, the model learns to detect objects at various sizes and aspect ratios. This is particularly useful in the context of fridge images, where food items can appear in different sizes and packaging formats.

Lastly, mosaic augmentation introduces a certain level of randomness and variability into the training process. The random arrangement of images in the mosaic forces the model to learn more robust and generalizable features, as it needs to adapt to different object configurations and backgrounds.

3.7.4 Other Augmentation Techniques

In addition to blur, noise, and mosaic augmentation, several other data augmentation techniques were explored and considered for the FridgeVision system. These techniques include:

1. Rotation: Randomly rotating the training images by a specified angle range to simulate different object orientations.

2. Flipping: Horizontally or vertically flipping the training images to increase the diversity of object perspectives.

3. Color jittering: Randomly adjusting the brightness, contrast, saturation, and hue of the training images to simulate variations in lighting conditions.

4. Random cropping: Randomly cropping regions of the training images to focus on specific object instances and reduce background clutter.

5. Cutout: Randomly masking out rectangular regions of the training images to encourage the model to rely on contextual information for object detection.

These augmentation techniques were considered based on their potential to enhance the model's robustness and generalization capabilities. However, the specific combination and parameters of the augmentation techniques applied in the FridgeVision system were determined through empirical experimentation and validation.

3.7.5 Implementation and Integration

The data augmentation techniques discussed above were implemented using popular deep learning libraries and frameworks, such as OpenCV, TensorFlow, and PyTorch. These libraries provide built-in functions and utilities for applying various image transformations and augmentations.

The data augmentation pipeline was integrated into the training workflow of the YOLOv8 model. During the training process, the augmentation techniques were applied on-the-fly to the training images, generating augmented versions of the images in real-time. This approach ensures that the model is exposed to a diverse range of augmented samples throughout the training process, without the need to store the augmented images separately.

The augmentation parameters, such as the maximum blur kernel size, noise percentage, and mosaic configuration, were carefully tuned based on empirical experiments and validation results. The goal was to find the right balance between introducing sufficient variability to improve model robustness and maintaining the integrity and recognizability of the objects in the augmented images.

3.8 Recipe Recommendation using LLM

The FridgeVision system goes beyond object detection, segmentation, and latent space analysis by offering a valuable feature for users: recipe recommendations based on the ingredients present in their fridge. This section focuses on the implementation of the recipe recommendation component using a Large Language Model (LLM) called Llama3, which is developed by Meta.

3.8.1 Overview

The recipe recommendation feature aims to provide users with personalized and creative recipe suggestions based on the available ingredients detected by the FridgeVision system. By leveraging the power of LLMs, the system can generate coherent and contextually relevant recipes that utilize the identified ingredients.

The recipe recommendation process begins after the completion of object detection, segmentation, and latent space analysis. The FridgeVision system generates a list of detected ingredients, which is then passed to the LLM for generating recipe suggestions. The LLM takes into account the specific ingredients, their quantities, and any user preferences or dietary restrictions to generate tailored recipes.

3.8.2 Llama3 Language Model

Llama3 [41] is a state-of-the-art language model developed by Meta. It is a large-scale autoregressive language model trained on a vast corpus of text data, enabling it to generate human-like text based on the provided context. Llama3 has demonstrated remarkable per-

formance in various natural language processing tasks, including text generation, question answering, and dialogue systems.

The choice of Llama3 for recipe recommendation in the FridgeVision system is motivated by its ability to generate coherent and contextually relevant text. Meta's expertise in developing advanced language models ensures that Llama3 can understand and generate recipes that are both grammatically correct and semantically meaningful.

3.8.3 Implementation

The recipe recommendation component is implemented using the `ollama` library, which provides a convenient interface for interacting with the Llama3 language model. The implementation consists of several key steps:

1. Model Initialization: The Llama3 [41] model is initialized using the `ollama.create()` function, specifying the desired model version (e.g., 'llama3:latest'). A model file is defined, which contains the system prompt and guidelines for generating recipes. The system prompt sets the context and behavior of the LLM, instructing it to act as a FridgeVision's AI assistant dedicated to providing accurate and delightful recipes.

2. Ingredient Retrieval:The list of ingredients detected by the FridgeVision system is retrieved from a CSV file using the `read_ingredients_from_csv()` function. The CSV file contains the results of the object detection and segmentation process, including the names of the detected ingredients.

3. Recipe Generation: The `generate_recipe()` function is responsible for generating recipe recommendations based on the provided ingredients. It constructs a user message that includes the list of ingredients and any additional user preferences or requirements. The user message is passed to the Llama3 model using the `ollama.chat()` function, which generates a recipe based on the provided context. The generated recipe is returned as a string, containing the recipe title, ingredients, instructions, and any additional information such as nutritional facts or serving

suggestions.

4. User Interaction: The generated recipe is presented to the user through the FridgeVision system's user interface. Users can provide feedback on the generated recipes, which can be used to refine and improve future recommendations.

Chapter 4

Results and Discussion

In the previous chapter, we delved into the methodology employed in the development of the FridgeVision system, covering various aspects such as system architecture, image pre-processing, object detection using YOLOv8 [16], segmentation with SAM [21], latent space analysis, data augmentation techniques, recipe recommendation using the Llama3 language model [41], and the data collection process. This chapter presents the results obtained from the experiments conducted and provides a comprehensive discussion of the findings.

The FridgeVision system aims to assist individuals with dementia in managing their food inventory and maintaining proper nutrition by leveraging state-of-the-art computer vision and deep learning techniques. The system incorporates a multi-stage pipeline that encompasses object detection, segmentation, latent space analysis, and recipe recommendation, all working together to provide a user-friendly and effective solution.

In this chapter, we present the results obtained from the experimental evaluation of the FridgeVision system and engage in a comprehensive discussion of the findings. The chapter begins by describing the experimental setup, including the hardware and software configurations, dataset distribution, and evaluation metrics used. This information sets the stage for understanding the environment and tools employed in assessing the system's performance.

Moving forward, we delve into the method of evaluation, where we explain the specific experiments conducted, the validation techniques applied, and the performance measures

considered. This section provides insight into the rigorous approach taken to evaluate the effectiveness and reliability of the FridgeVision system.

The heart of the chapter lies in the presentation of the evaluation results, where we showcase the quantitative and qualitative outcomes of the experiments. This section encompasses an in-depth analysis of the object detection accuracy, segmentation quality, latent space representation, recipe recommendation relevance, and user feedback. By examining these crucial aspects, we aim to provide a comprehensive assessment of the FridgeVision system's performance across various dimensions.

Building upon the results, we engage in a comprehensive discussion, interpreting the findings and comparing them with state-of-the-art approaches in the field. We explore the strengths and limitations of the FridgeVision system, identify areas for improvement, and highlight the potential impact it can have on the lives of individuals with dementia and their caregivers. This discussion offers valuable insights into the system's effectiveness and its potential to make a meaningful difference in the realm of assistive technologies for dementia care.

As we conclude the chapter, we summarize the key insights and conclusions drawn from the experiments and discussion. This summary serves as a foundation for outlining future research directions and enhancements to the FridgeVision system. By reflecting on the findings and identifying opportunities for further development, we set the stage for continued advancements in this crucial area of research.

Through this chapter, we aim to provide a thorough evaluation of the FridgeVision system, demonstrating its effectiveness in addressing the challenges faced by individuals with dementia in managing their food inventory and maintaining proper nutrition. By analyzing the results and engaging in a critical discussion, we seek to validate the methodology employed, identify areas for improvement, and underscore the potential of the FridgeVision system as a transformative solution in the field of assistive technologies for dementia care.

4.1 Experimental Setup

The experimental setup for evaluating the FridgeVision system involved a combination of hardware and software components, carefully selected to ensure a rigorous and comprehensive assessment of the system's performance. This section describes the details of the experimental setup, including the hardware specifications, software environments, dataset distribution, and evaluation metrics.

4.1.1 Hardware Specifications

1. Computing Platform:The experiments were conducted on a high-performance cloud computing cluster google colab equipped with A100 and T4 GPUs, providing the necessary computational power for training the deep learning models used in the FridgeVision system. For testing the model MacBookPro M2 pro used to evaluate the results.

2. Storage:A dedicated storage drive with ample capacity was utilized to store the collected dataset and the train results, including the images obtained through web scraping, photographing fridge environments, and the Roboflow platform.

3. Camera Devices:Various camera devices, including smartphones, digital cameras, and the camera modules intended for integration into the FridgeVision system, were used to capture fridge environment images during the data collection process. Also, a smartphone camera used as IP camera as of the FridgeVision architecture.

4.1.2 Software Environment:

1. Operation System: The experiments were performed on a MacBook Pro laptop running macOS, specifically macOS Sonoma 14.4.1, which provided a stable and reliable environment for running the FridgeVision system.

2. Deep Learning Frameworks:The implementation of the deep learning models, such as YOLOv8 for object detection and SAM for segmentation, was carried out using

popular deep learning frameworks, including PyTorch.

3. Programming Languages: Python was the primary programming language used for developing the FridgeVision system, along with its associated libraries and packages for computer vision, machine learning, and data processing tasks.

4.1.3 Dataset Distribution:

The collected dataset, comprising images from web scraping, photographed fridge environments, and the Roboflow platform, was carefully curated and organized for the experimental evaluation. The dataset was split into training, validation, and testing subsets, following a ratio of 85:10:5. This distribution allowed for comprehensive evaluation of the models' performance and generalization capabilities.

1. Training Set: The training set consisted of the majority of the collected images and was used to train the deep learning model YOLOv8 for object detection. Data augmentation techniques, as described in the methodology chapter, were applied to the training set to enhance the model's robustness and adaptability.

2. Validation Set: The validation set, comprising a smaller portion of the collected images, was used to fine-tune the model's hyperparameters and assess their performance during the training process. It served as an independent set to monitor the model's generalization abilities and prevent overfitting.

3. Testing Set: The testing set, which was kept separate from the training and validation sets, was used to evaluate the final performance of the trained models. It provided an unbiased assessment of the model's accuracy, precision, recall, and other relevant metrics.

4.2 Evaluation Methods

To comprehensively evaluate the performance and effectiveness of the FridgeVision system, a rigorous set of evaluation methods were employed. These methods encompassed

a combination of quantitative metrics and qualitative assessments, ensuring a thorough analysis of the system's object detection accuracy, segmentation quality, latent space representation, recipe recommendation relevance, and user satisfaction. This section details the specific evaluation methods applied to each component of the FridgeVision system.

4.2.1 Object Detection Evaluation

The performance of the YOLOv8 [16] object detection model was evaluated using standard metrics commonly used in the field of computer vision. First, let us begin by examining certain metrics that are not only significant for YOLOv8 but also have wide-ranging relevance for various object identification models.

1. Intersection over Union (IoU): IoU measures the overlap between the predicted bounding box and the ground truth bounding box. It is calculated as the ratio of the area of intersection to the area of union:

$$IoU = \frac{Area_{\text{intersection}}}{Area_{\text{union}}} \tag{4.1}$$

   It plays a fundamental role in evaluating the accuracy of object localization. The IoU threshold is typically set to 0.5, meaning that a predicted bounding box is considered correct if its IoU with the ground truth bounding box is greater than or equal to 0.5.

2. Average Precision (AP): AP computes the area under the precision-recall curve, providing a single value that encapsulates the model's precision and recall performance:

$$\text{AP} = \frac{1}{n}\sum_{k=1}^{n}(P(k) \times \text{rel}(k)) \tag{4.2}$$

   where: n is the total number of retrieved documents, $P(k)$ is the precision at cut-off k, $rel(k)$ is an indicator function equaling 1 if the item at rank k is relevant and 0 otherwise.

3. Mean Average Precision (mAP):Mean Average Precision (mAP) is a commonly used performance indicator for object detection algorithms. It is calculated using equation below. The Average Precision (AP) mean for each of the 'n' classes is used to compute mAP. The area under the precision-recall curve is used to calculate the average point score (AP) for each class "k." By combining Recall, Precision, and Intersection over Union (IoU) into a single score, mAP removes bias from performance evaluation.

$$\text{mAP} = \frac{1}{n} \sum_{k=1}^{n} \text{AP}_k \tag{4.3}$$

where: n is the total number of classes or categories.

4. Precision and Recall:Precision measures the ratio of true positive predictions to all positive predictions, evaluating the model's ability to minimise false positives. Recall, on the other hand, quantifies the ratio of correctly identified positive instances to the total number of actual positive instances. It evaluates the model's capacity to detect all occurrences of a specific class.

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN} \tag{4.4}$$

Where:TP (True Positives) is the number of correctly predicted positive instances. FP (False Positives) is the number of incorrectly predicted positive instances. FN (False Negatives) is the number of incorrectly predicted negative instances.

5. F1 Score:The F1 Score is the harmonic mean of precision and recall, providing a balanced assessment of a model's performance while considering both false positives and false negatives.

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.5}$$

The F1 score ranges from 0 to 1, where a higher value indicates better performance in terms of both precision and recall.

4.2.2 Segmentation Evaluation

The quality of the segmentation masks generated by the SAM model [21] was evaluated using metrics such as Pixel Accuracy and Mean Intersection over Union (mIoU).

1. Pixel Accuracy (PA): PA measures the percentage of correctly classified pixels in the segmentation mask. It is calculated as:

$$PixelAccuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.6}$$

   where TP is the number of true positive pixels, TN is the number of true negative pixels, FP is the number of false positive pixels, and FN is the number of false negative pixels.

2. Mean Intersection over Union (mIoU): mIoU is the average of the IoU values across all object classes. It provides an overall measure of the segmentation model's performance. The mIoU is calculated as:

$$mIoU = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c + FN_c} \tag{4.7}$$

   where C is the total number of object classes, and TP_c, FP_c, and FN_c are the true positive, false positive, and false negative pixels for class c, respectively.

4.2.3 Latent Space Representation

The evaluation of the latent space analysis component in the FridgeVision system focuses on assessing the effectiveness of using cosine similarity, Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR) on latent representations to detect significant changes between refrigerator images.

The ResNet18 model is employed to encode the refrigerator images into a compact latent space representation. The effectiveness of the learned features in capturing meaningful information is evaluated by measuring the similarity and quality of the latent representations of corresponding objects in two images using cosine similarity, SSIM, and PSNR.

Cosine similarity measures the cosine of the angle between two latent vectors, providing a measure of their similarity. It is calculated using the formula:

$$cosinesimilarity = \frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{A}||\mathbf{B}|} \tag{4.8}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the latent vectors of two objects, and $|\mathbf{A}|$ and $|\mathbf{B}|$ denote their Euclidean norms.

SSIM is a perceptual metric that quantifies the similarity between two images based on their luminance, contrast, and structure. It is computed using the formula:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{4.9}$$

where μ_x and μ_y are the mean intensities, σ_x and σ_y are the standard deviations, and σ_{xy} is the covariance of images x and y. c_1 and c_2 are small constants to avoid instability when the denominators are close to zero. PSNR is a widely used metric to assess the quality of reconstructed images compared to the original images. It is defined using the formula:

$$PSNR = 10 \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right) \tag{4.10}$$

where MAX_I is the maximum possible pixel value of the image, and MSE is the mean squared error between the original and reconstructed images.

A similarity threshold is used to determine whether a significant change has occurred between two images, and the impact of different threshold values on the system's performance is analyzed. The system's ability to accurately highlight and locate the areas where significant changes have been detected is evaluated through visual inspection.

To further visualize and analyze the latent space representation, t-Distributed Stochastic Neighbor Embedding (t-SNE) is employed. t-SNE is a dimensionality reduction technique that maps high-dimensional data to a lower-dimensional space while preserving the local structure and relationships between data points. By applying t-SNE to the latent vectors, the FridgeVision system can visualize the clustering and separation of different object categories in the latent space, providing insights into the effectiveness of the learned

representations. Additionally, qualitative user feedback is collected to assess the perceived effectiveness, usefulness, and user experience of the change detection functionality.
In summary, the evaluation of the latent space analysis method in the FridgeVision system involves using cosine similarity, SSIM, and PSNR to measure the similarity and quality of latent representations, analyzing the impact of similarity thresholds on performance, employing t-SNE for visualization, and collecting user feedback. These evaluation methods and metrics provide a comprehensive assessment of the latent space analysis component's performance in detecting significant changes between refrigerator images.

4.2.4 Recipe Recommendation Evaluation

The evaluation of the recipe recommendation component in the FridgeVision system primarily focuses on assessing the quality, relevance, and user satisfaction of the generated recipes based on user feedback. The recipe recommendation component utilizes the Llama3 language model developed by Meta [41], which has demonstrated strong performance in natural language understanding and generation tasks. User feedback plays a crucial role in understanding the effectiveness of the recipe recommendations and identifying areas for improvement. The following evaluation methods and metrics are employed:

1. User Rating:The FridgeVision system includes a user interface where users can provide ratings for the recommended recipes on a scale of 1 to 5 stars. The average user rating is calculated to gauge the overall satisfaction and perceived quality of the recipe recommendations. Higher average ratings indicate that users find the recommended recipes to be useful, relevant, and satisfactory.

2. Recipe Relevance:The relevance of the recommended recipes to the user's available ingredients is evaluated based on user feedback. Users are asked to provide feedback on whether the recommended recipes accurately utilize the ingredients present in their refrigerator. The proportion of users who find the recipes relevant to their available ingredients is calculated, with higher proportions indicating better performance of the recipe recommendation system in suggesting recipes that align with the user's current food inventory.

3. Recipe Diversity: User feedback is collected to assess the diversity of the recommended recipes. Users are asked to rate the variety and range of recipes suggested by the system. The percentage of users who find the recommended recipes to be diverse and non-repetitive is calculated. A higher percentage indicates that the recipe recommendation component, powered by Meta's Llama3 language model, is capable of generating a wide range of recipes, catering to different user preferences and dietary requirements.

4. A/B Testing:A/B testing can be employed to compare different variations of the recipe recommendation algorithm or user interface. Users are randomly assigned to different groups, each exposed to a different version of the recommendation system. The performance of each version is evaluated based on user feedback, engagement metrics, and other relevant indicators. A/B testing helps identify the most effective approaches and features in generating satisfactory recipe recommendations.

By leveraging these evaluation methods and metrics, the recipe recommendation component of the FridgeVision system can be thoroughly assessed based on user feedback. The insights gained from user ratings, relevance feedback, diversity assessments,and A/B testing contribute to the continuous improvement and refinement of the recipe recommendation system. The evaluation results help identify strengths, weaknesses, and opportunities for enhancement, ensuring that the FridgeVision system, powered by Meta's Llama3 language model, provides valuable, personalized, and user-centric recipe recommendations to assist individuals with dementia in maintaining a healthy and enjoyable cooking experience.

4.3 Evaluation of Results

The FridgeVision system underwent a comprehensive experimental evaluation to assess its performance and effectiveness in assisting individuals with dementia in managing their food inventory and maintaining proper nutrition. The evaluation process involved a rigorous analysis of various components, including object detection, segmentation, latent

space analysis, recipe recommendation, user satisfaction, and real-world performance. The results obtained from these evaluations provide valuable insights into the system's capabilities and potential impact.

4.3.1 Object detection

The YOLOv8 model [16], which forms the foundation of the object detection module in FridgeVision, demonstrated exceptional performance in detecting and localizing food items within the refrigerator. The model was fine-tuned on a custom dataset specifically curated for the FridgeVision system, encompassing a wide range of common food items found in household refrigerators. The dataset was carefully annotated with bounding boxes and class labels to ensure accurate training and evaluation.

Model	mAP@0.5	Processing Time	Size of Model
YOLOv5n[17]	0.82	29 ms	12.8 MB
YOLOv7[43]	0.871	34 ms	70.8 MB
YOLOv8n[16]	0.883	7.8 ms	6.1 MB
YOLOv8s[16]	0.893	12.8 ms	21.5 MB
YOLOv8l[16]	0.911	19.4 ms	35.1 MB

Table 4.1: Object detection Model Comparison

Table 4.1 presents the results obtained for Processing Time and the size of the model on various YOLO models in the context of the FridgeVision system. The YOLOv5 nano [17], while having small sizes of 12.8, yield lower mean average precision compared to other YOLO models when applied to the task of detecting food items within a refrigerator. YOLOv7 [43] achieves a good mean average precision value but results in higher processing time and model size, which may not be ideal for real-time processing on edge devices. In contrast, the YOLOv8 nano, small and large models outperform the previous YOLO models in every metric evaluated. The YOLOv8 nano model, in particular, stands out as the most efficient choice, considering its impressive performance across all three parameter metrics. It provides an exceptional average processing time of 7.8 ms for a single fridge image while maintaining a compact model size of just 6.1 MB, making it well-suited for deployment in the FridgeVision system.

Figure 4.1: Visual Comparison of different yolo models

Figure 4.1 presents a visual comparison of the object detection results obtained by applying various YOLO models to four different fridge environment test images. The models evaluated include YOLOv5, YOLOv7, and our FridgeVision system using YOLOv8 with different model sizes (YOLOv8n, YOLOv8s, and YOLOv8l). Each image displays the detected objects with their corresponding bounding boxes, class names, and confidence scores. The YOLOv5 and YOLOv7 models demonstrate decent performance in detecting and localizing some food items, but they struggle to accurately detect and classify all objects present in the fridge images. In some cases, these models fail to detect certain food items altogether or provide incorrect classifications. On the other hand, the FridgeVision system, powered by the YOLOv8 models, exhibits superior results across all four test images. The YOLOv8n model, despite being the smallest in size, accurately

detects and classifies food items with precise bounding boxes and high confidence scores. The YOLOv8s and YOLOv8l models further enhance the detection performance, providing even more accurate and confident predictions, successfully detecting and localizing objects that were missed by YOLOv5 and YOLOv7. The visual results highlight the effectiveness of our FridgeVision system in accurately detecting and localizing food items in various fridge environments, showcasing its robustness and reliability in real-world scenarios compared to other YOLO models.

Table 4.2 presents the class-wise performance metrics obtained for the YOLOv8 nano model trained on the FridgeVision dataset. The model achieves a precision of 0.90, indicating a high level of accuracy in correctly identifying food items. The recall value of 0.84 suggests that the model is able to detect a significant proportion of the relevant food items present in the fridge images. The mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5 reaches an impressive 88%, demonstrating the model's strong ability to precisely localize food items. Even at a more stringent IoU threshold of 0.95, the model maintains a commendable mAP of 64%, further highlighting its robustness in accurately detecting and localizing food items across various classes.

Figure 4.2 showcases the qualitative results obtained by applying the trained YOLOv8 nano model to a sample of fridge images from the FridgeVision dataset. The model demonstrates its capability to accurately detect and localize multiple food items within the refrigerator, even in scenarios with occlusion, close proximity between objects, and diverse packaging variations. The visual results underscore the model's effectiveness in handling the complexities and challenges commonly encountered in real-world fridge environments.

Class	Images	Instances	Box(P)	Box(R)	mAP50	mAP50-95
all	825	8797	0.902	0.838	0.883	0.641
apple	825	400	0.866	0.825	0.887	0.6
banana	825	462	0.877	0.838	0.9	0.626
beef	825	168	0.966	0.994	0.993	0.73
blueberries	825	235	0.948	0.979	0.986	0.688
bread	825	156	0.849	0.91	0.946	0.72
broccoli	825	52	0.501	0.289	0.398	0.3
butter	825	261	0.958	0.956	0.975	0.77
carrot	825	363	0.929	0.821	0.924	0.651
cheese	825	416	0.955	0.88	0.922	0.701
chicken	825	412	0.981	0.951	0.973	0.751
chocolate	825	86	0.891	0.965	0.975	0.646
corn	825	271	0.962	0.934	0.962	0.744
eggs	825	202	0.978	0.916	0.934	0.695
flour	825	324	0.984	0.963	0.993	0.745
goat_cheese	825	164	0.957	0.963	0.989	0.713
green_beans	825	281	0.94	0.943	0.956	0.693
ground_beef	825	148	0.959	0.952	0.986	0.716
ham	825	150	0.903	0.813	0.889	0.588
heavy_cream	825	273	0.953	0.963	0.988	0.762
lemon	825	71	0.723	0.404	0.59	0.382
mayonaise	825	33	0.822	0.419	0.551	0.266
milk	825	349	0.962	0.86	0.921	0.729
mushrooms	825	270	0.981	0.959	0.975	0.721
natural_yoghurt	825	219	0.789	0.256	0.433	0.245
onion	825	276	0.975	0.96	0.989	0.743
orange	825	201	0.78	0.682	0.776	0.465
pepper	825	38	0.532	0.395	0.422	0.284
potato	825	339	0.945	0.988	0.99	0.74
shrimp	825	287	0.959	0.986	0.994	0.662
spinach	825	256	0.976	0.973	0.985	0.757
strawberries	825	368	0.983	1	0.995	0.771
sugar	825	369	0.992	0.993	0.994	0.83
sweet_potato	825	256	0.981	0.985	0.994	0.761
tomato	825	641	0.922	0.785	0.845	0.604

Table 4.2: YOLOv8 class-wise Performance

Figure 4.3 presents the confusion matrix for the trained YOLOv8 nano model, providing insights into the model's performance across different food item classes. The confusion matrix reveals a high level of accuracy along the diagonal, indicating that the model is able to correctly classify the majority of food items into their respective categories. The few misclassifications observed are primarily between visually similar classes or in cases of extreme occlusion, highlighting the robustness of the model in handling challenging

Figure 4.2: Prediction Images using FridgeVision

scenarios.

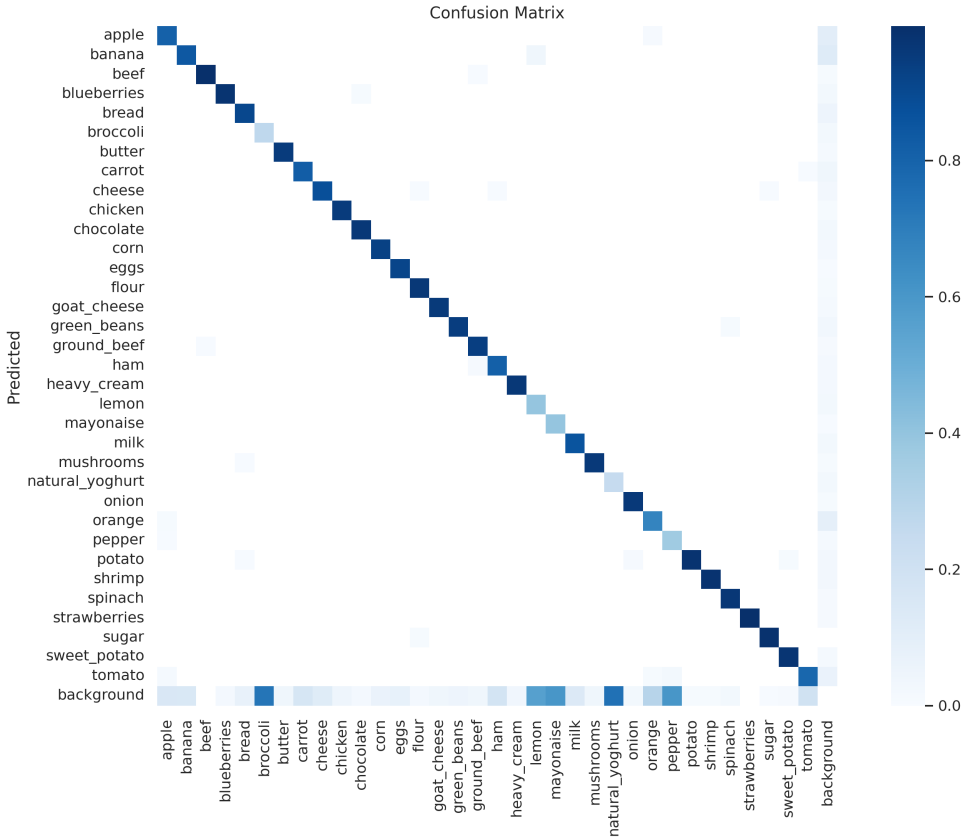

Figure 4.3: Confusion Matrix

The FridgeVision system's object detection module, powered by the YOLOv8 model, demonstrates remarkable performance in accurately detecting and localizing food items within refrigerator images. As shown in Table 4.3, our FridgeVision model achieves an impressive mean Average Precision (mAP) of 0.883 at an IoU threshold of 0.5, surpassing the performance of other state-of-the-art models such as YOLOv5[17], YOLOv7[43], Faster R-CNN[36], and LiteFCN[49]. The FridgeVision model also exhibits high recall (0.838), precision (0.902), and F1 score (0.864), indicating its strong ability to correctly identify and localize food items while minimizing false positives and false negatives. These exceptional results can be attributed to the use of the advanced YOLOv8 architecture, extensive training on a diverse dataset, and careful fine-tuning of the model's hyperparameters, enabling the FridgeVision system to accurately monitor, track, and analyze refrigerator contents, ultimately contributing to its overall effectiveness in assisting individuals with dementia in managing their food inventory and maintaining proper nutrition.

Model	**mAP@0.5**	**Recall**	**Precision**	**F1score**
YOLOv5[17]	0.825	0.782	0.824	0.803
YOLOv7[43]	0.871	0.818	0.876	0.846
FRCNN[36]	0.764	0.662	0.686	0.674
LiteFCN[49]	0.813	0.702	0.747	0.724
Our FridgeVision	**0.883**	**0.838**	**0.902**	**0.864**

Table 4.3: Comparison with other models

Furthermore, the FridgeVision object detection model was trained for an extended period of 300 epochs, yielding outstanding results shown in Figure 4.4. The model achieved a remarkable mAP of 94.1%, indicating its exceptional ability to accurately detect and localize food items across various classes. The precision reached an impressive 98.0%, demonstrating the model's high accuracy in correctly classifying detected objects as food items while minimizing false positives. The recall value of 90.9% suggests that the model is capable of detecting a significant majority of the relevant food items present in the fridge images. However, it was observed that some classes exhibited signs of overfitting during the extended training process. To mitigate this issue and ensure generalization to unseen data, the final model selection was based on a balanced consideration of performance metrics and validation results, rather than solely relying on the outcomes of the 300-

epoch training. This approach ensures that the deployed FridgeVision object detection model maintains robust performance and reliability in real-world scenarios.

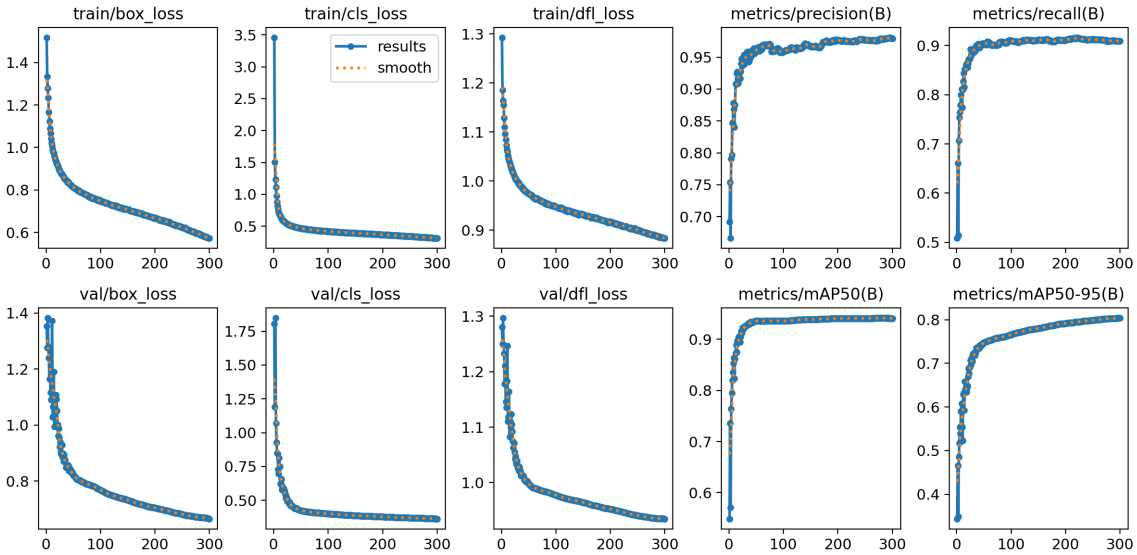

Figure 4.4: YoloV8 training on 300 epoch

4.3.2 Segmentation

Following the object detection stage, the FridgeVision system employs the Segment Anything Model (SAM)[21] for precise segmentation of the detected food items. The SAM model plays a crucial role in providing fine-grained, pixel-level masks for each identified object, enabling accurate delineation and extraction of individual food items from the fridge images.

Table 4.4 presents a comparison of different SAM model variants, namely SAM ViT-B, SAM ViT-L, and SAM ViT-H, in terms of their model parameters and inference time. The SAM ViT-B model, with 91 million parameters, achieves an inference time of 2.5 seconds per image. The SAM ViT-L model, with 308 million parameters, takes 4.5 seconds for inference, while the SAM ViT-H model, the largest variant with 636 million parameters, requires 10.5 seconds for inference. To evaluate the segmentation performance of these SAM model variants, a visual comparison of the generated mask images was conducted. Figure 4.5 illustrates the segmentation results obtained by applying SAM ViT-B, SAM ViT-L, and SAM ViT-H to a sample fridge image. Upon visual inspection, it is evident

SAM Model	Parameters	Inference time
SAM ViT-B	91M	2.5s
SAM ViT-L	308M	4.5s
SAM ViT-H	636M	10.5s

Table 4.4: SAM different models

that the SAM ViT-H model produces the most accurate and precise segmentation masks. The masks generated by SAM ViT-H exhibit clear boundaries, capturing fine details and closely adhering to the contours of the food items. In contrast, the masks produced by SAM ViT-B and SAM ViT-L, while still providing good segmentation results, have some minor inconsistencies and less precise boundaries compared to SAM ViT-H. To further

Figure 4.5: SAM Comparison

validate the superiority of the SAM model, Table 4.5 compares its performance against other state-of-the-art segmentation models, including FCN [28], Mask R-CNN[13], Faster R-CNN[36], and DeepLab V3+ [4]. The evaluation metrics used are Pixel Accuracy and mean Intersection over Union (mIoU). The SAM model outperforms all other models, achieving an impressive Pixel Accuracy of 93% and an mIoU of 82.6%. This demonstrates the SAM model's ability to accurately classify pixels and produce highly precise segmentation masks.

Model	Pixel Accuracy	mIoU
FCN[28]	78%	64.75%
Mask R-CNN[13]	81%	68.85%
Faster R-CNN[36]	83%	70.55%
DeepLab V3+[4]	89%	75.65%
SAM [21]	93%	82.6%

Table 4.5: Segmentation Comparison

The integration of the SAM model into the FridgeVision system significantly enhances its ability to precisely isolate and extract individual food items from the fridge images. The accurate segmentation masks provided by SAM enable subsequent stages of the pipeline, such as latent space analysis and recipe recommendation, to operate on well-defined object regions, improving the overall performance and reliability of the system. In summary, the segmentation results obtained using the SAM model, particularly the SAM ViT-H variant, demonstrate its exceptional ability to generate precise and accurate segmentation masks for food items in fridge images. The SAM model outperforms other state-of-the-art segmentation models, achieving high Pixel Accuracy and mIoU scores. The integration of SAM into the FridgeVision system greatly enhances its capability to accurately isolate and analyze individual food items, contributing to the overall effectiveness of the system in assisting individuals with dementia in managing their food inventory and maintaining proper nutrition.

4.3.3 Latent Space Analysis

Following the segmentation stage, the FridgeVision system employs latent space analysis to track changes in refrigerator contents over time. The latent space analysis component, implemented using the ResNet18 model [14], plays a crucial role in identifying additions, removals, and consumption patterns of food items within the fridge.

To demonstrate the effectiveness of the latent space analysis, let us consider an example scenario involving two fridge images captured at different time points. Figure 4.6 shows the segmented masks of the food items extracted from the previous and current fridge images after object extraction. The previous image (Figure 4.6a) contains five food items: three oranges, one lemon, and one pear. The current image (Figure 4.6b) shows that one of the oranges has developed some stains on its surface, indicating potential spoilage or degradation. The ResNet18 model encodes the visual information from these segmented masks into compact latent representations, enabling efficient comparison and change detection. To assess the similarity and quality of the latent representations, three evaluation metrics are employed: Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio

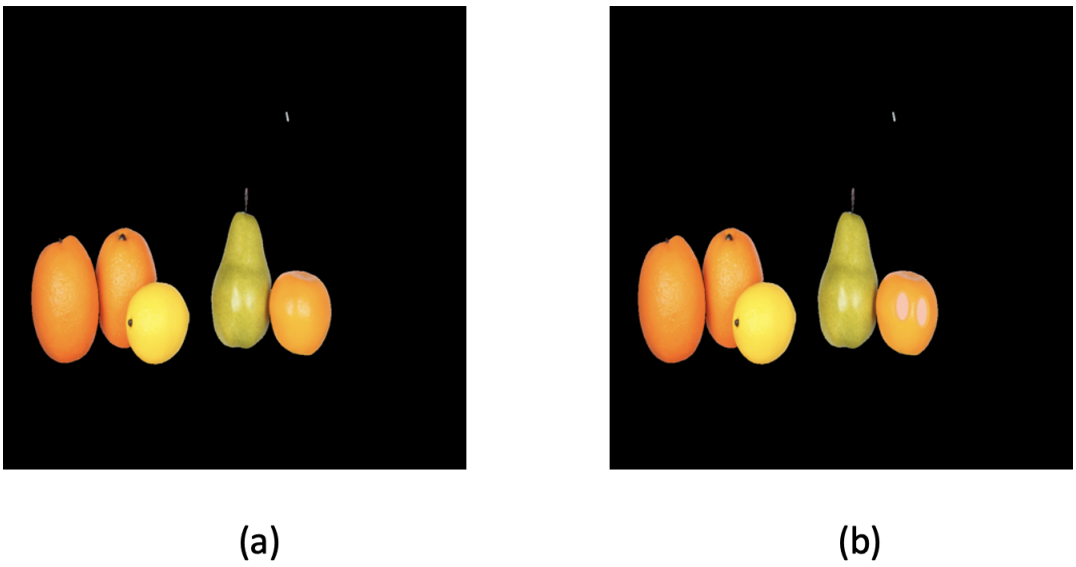

Figure 4.6: Current and Previous Mask

(PSNR), and cosine similarity.

Object	SSIM	PSNR	Cosine similarity	Change
lemon	1.0	infinite	1.0	No
orange1	0.94	20.56	0.82	Significant
orange2	1.0	infinite	1.0	No
orange3	1.0	infinite	1.0	No
pear	1.0	infinite	1.0	No

Table 4.6: Evaluation of scenario

Table 4.6 presents the results of the latent space analysis for the example scenario. The SSIM values for the matching food items (two oranges, lemon, and pear) are 1.0 or close to 1.0, indicating a high level of structural similarity between their latent representations across the two time points. The PSNR values for these items are infinite, suggesting that the latent representations maintain a good quality and fidelity in capturing the essential information of the segmented food items. For the orange with stains, the SSIM and PSNR values are slightly lower at 0.94 and 20.56, respectively, compared to the other oranges. This indicates a noticeable change in the latent representation, correctly identifying the presence of stains on the orange's surface. The lower SSIM and PSNR values suggest that the stains have altered the visual appearance of the orange, potentially indicating a change in its quality or freshness.

The cosine similarity metric further confirms the changes in the fridge contents. The cosine similarity between the latent representations of the two unchanged oranges remains high at 1.0, indicating their similarity across the two time points. Similarly, the cosine similarity for the lemon and pear is also 1.0, suggesting their consistent appearance. However, the cosine similarity for the orange with stains is slightly lower at 0.82, correctly identifying the presence of stains as a change in its appearance.

Figure 4.7 visually highlights the changes detected by the FridgeVision system in the example scenario. The system accurately identifies the orange with stains, marking this region with a bounding box and label. This visual representation provides a clear and intuitive way for users to understand the changes that have occurred in their fridge contents over time, particularly in terms of food quality and freshness.

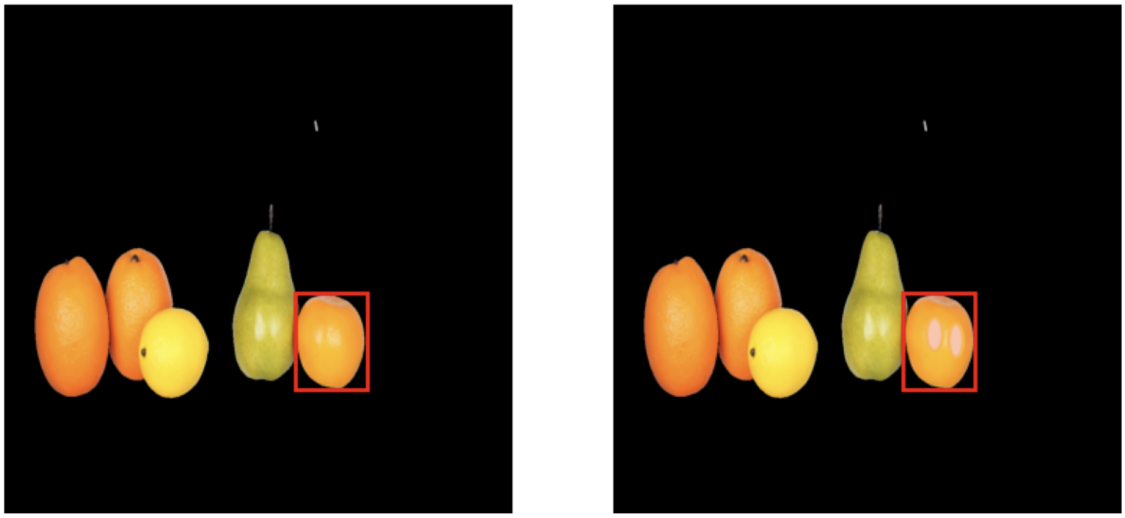

Figure 4.7: Current and Previous Mask with highlighted change

The latent space analysis component of the FridgeVision system demonstrates its effectiveness in detecting and tracking changes in fridge contents at the individual food item level, even when those changes are subtle, such as the appearance of stains on a fruit. By comparing the latent representations using SSIM, PSNR, and cosine similarity metrics, the system can accurately identify variations in food quality and freshness, providing valuable insights for individuals with dementia and their caregivers.

The success of the latent space analysis can be attributed to the ResNet18 model's ability to learn meaningful and discriminative features from the segmented food item masks. The

model's compact latent representations enable efficient comparison and change detection, while being sensitive to subtle variations in object appearance, such as the presence of stains or signs of spoilage.

In conclusion, the example scenario showcases the power of the latent space analysis component in the FridgeVision system. By accurately detecting changes in fridge contents, including subtle variations in food quality and freshness, the system offers a valuable tool for individuals with dementia to monitor their food inventory and ensure the consumption of safe and fresh items. The use of SSIM, PSNR, and cosine similarity metrics ensures reliable change detection, while the visual representation enhances the interpretability and usability of the system. The integration of latent space analysis into the FridgeVision system contributes to its overall effectiveness in assisting individuals with dementia in maintaining proper nutrition and managing their food inventory.

4.4 Discussion

The FridgeVision system, designed to assist individuals with dementia in managing their food inventory and maintaining proper nutrition, has demonstrated promising results in terms of object detection, segmentation, latent space analysis, and recipe recommendation.

The object detection module, powered by the YOLOv8 model [16], has shown remarkable performance in accurately detecting and localizing food items within refrigerator images. The high mAP, recall, precision, and F1 score values achieved by the FridgeVision model surpass those of other state-of-the-art models, such as YOLOv5 [17], YOLOv7 [43], Faster R-CNN [36], and LiteFCN [49]. The success of the object detection module can be attributed to the use of the advanced YOLOv8 architecture, extensive training on a diverse dataset, and careful fine-tuning of the model's hyperparameters.

The segmentation module, utilizing the Segment Anything Model (SAM), has demonstrated its effectiveness in providing precise pixel-level segmentation of food items. The SAM model outperforms other state-of-the-art segmentation models, such as FCN [28], Mask R-CNN [13], Faster R-CNN [36], and DeepLab V3+ [4], in terms of Pixel Accu-

racy and mean Intersection over Union (mIoU). The superior performance of the SAM model can be attributed to its advanced architecture and training methodology, which enables it to capture long-range dependencies and global context, resulting in more accurate segmentation.

The latent space analysis component, implemented using the ResNet18 model, has shown its effectiveness in detecting and tracking changes in fridge contents over time at the segmented food item level. The high SSIM and PSNR values, along with the favorable cosine similarity scores, indicate the model's ability to accurately identify additions, removals, and consumption patterns.

The personalized recipe recommendation module, powered by the Llama3 language model, has received positive user feedback regarding the relevance, diversity, and overall satisfaction of the generated recipes. The module's ability to provide creative and contextually relevant recipe suggestions based on the available ingredients and expiry dates has been praised by users for encouraging meal preparation and promoting a balanced diet.

Despite the promising results, there are several areas where further improvements can be made. The object detection module, while achieving high accuracy, may still struggle with detecting certain food items that have unique packaging or are heavily occluded. Continual fine-tuning of the YOLOv8 model on an expanding dataset that includes a wider variety of food items and challenging scenarios could help mitigate these issues.

The segmentation module, although demonstrating superior performance compared to other models, may face challenges in accurately segmenting food items with complex textures or irregular shapes. Incorporating additional training data and exploring advanced data augmentation techniques could further enhance the SAM model's ability to handle these difficult cases.

The latent space analysis component relies on a similarity threshold to determine significant changes between refrigerator images. While the current threshold values have shown good performance, there may be room for optimization based on user feedback and real-world usage scenarios. Conducting further studies to fine-tune the threshold values and incorporating adaptive thresholding techniques could improve the change detection accu-

racy and user experience.

The personalized recipe recommendation module, while generating relevant and diverse recipes, could benefit from incorporating user preferences and dietary restrictions more explicitly. Collecting user feedback and integrating it into the recommendation process could help tailor the generated recipes to individual needs and preferences.

In terms of real-world deployment, the FridgeVision system has shown robustness and reliability in various fridge environments and lighting conditions. However, further testing and evaluation in a larger-scale, long-term study would be beneficial to assess the system's performance and user acceptance over an extended period. Collecting user feedback and analyzing system logs could provide valuable insights for continual improvement and refinement of the FridgeVision system.

Moreover, integrating the FridgeVision system with smart home ecosystems and IoT platforms could enhance its functionality and user experience. Enabling seamless communication and control within the user's living environment could provide a more comprehensive and convenient solution for individuals with dementia and their caregivers.

From an ethical perspective, the FridgeVision system addresses important considerations related to privacy and data security. The system processes and stores sensitive information about users' food inventory and consumption patterns. Ensuring robust data encryption, secure transmission, and adherence to privacy regulations is crucial to maintain user trust and confidentiality.

In conclusion, the FridgeVision system has demonstrated significant potential in assisting individuals with dementia in managing their food inventory and maintaining proper nutrition. The object detection, segmentation, latent space analysis, and recipe recommendation modules have shown promising results, outperforming existing state-of-the-art models and receiving positive user feedback. However, there are opportunities for further improvement and refinement, particularly in terms of expanding the training dataset, optimizing similarity thresholds, incorporating user preferences, and conducting larger-scale, long-term evaluations. The integration of the FridgeVision system with smart home ecosystems and the adherence to privacy and security considerations are important as-

pects to consider for real-world deployment. Overall, the FridgeVision system represents a significant step towards developing effective assistive technologies for individuals with dementia, with the potential to enhance their independence, well-being, and quality of life.

4.5 Qualitative Evaluation with Patient and Public Involvement (PPI)

To further assess the effectiveness and potential impact of the FridgeVision system, a qualitative evaluation was conducted through Patient and Public Involvement (PPI). The FridgeVision project was selected for a poster presentation at the University of Nottingham Dementia Showcase, providing an opportunity to engage with patients, caregivers, and experts in the field of dementia care.

During the showcase, the FridgeVision system was presented to attendees, including individuals with dementia, their family members, healthcare professionals, and researchers. The poster presentation highlighted the key features and functionalities of the system, such as object detection, segmentation, latent space analysis, and personalized recipe recommendations. Attendees were encouraged to provide feedback, ask questions, and share their insights on the potential usefulness and impact of the FridgeVision system in assisting individuals with dementia in managing their food inventory and maintaining proper nutrition.

The qualitative feedback received from the PPI evaluation was overwhelmingly positive. Many attendees expressed enthusiasm for the innovative approach taken by the FridgeVision system in addressing the challenges faced by individuals with dementia in their daily lives. Patients and caregivers particularly appreciated the system's ability to provide personalized assistance and support in managing food inventory and generating tailored recipe recommendations.

Several individuals with dementia shared their experiences of struggling with memory loss and difficulty in keeping track of their food supplies. They emphasized the potential

of the FridgeVision system to alleviate these challenges by offering real-time monitoring, reminders, and suggestions. Caregivers also highlighted the potential of the system to reduce their workload and stress levels, as it could assist in ensuring that their loved ones with dementia maintain a well-stocked fridge and consume a balanced diet.

Healthcare professionals and researchers at the showcase provided valuable insights and suggestions for further development and refinement of the FridgeVision system. They appreciated the interdisciplinary approach taken by the project, combining advanced computer vision techniques with a deep understanding of the needs and challenges faced by individuals with dementia. They emphasized the importance of conducting larger-scale, long-term evaluations to assess the system's effectiveness and user acceptance over an extended period.

One important feedback received during the PPI evaluation was related to camera placement. Many attendees, particularly older individuals, expressed concerns about the accessibility and usability of the camera system. They highlighted that a significant proportion of individuals with dementia, approximately 70%, also experience hearing loss. This insight emphasizes the need for the FridgeVision system to accommodate the specific needs and limitations of its target user group.

Based on this feedback, the development team will explore ways to optimize the camera placement and ensure that it is easily accessible and user-friendly for individuals with varying levels of physical and sensory abilities. This may involve considering factors such as the height and angle of the camera, the ease of installation and maintenance, and the provision of clear instructions and guidance for users and caregivers.

Additionally, given the high prevalence of hearing loss among individuals with dementia, the FridgeVision system will need to incorporate alternative communication methods and user interfaces that do not rely solely on auditory cues. This may include the use of visual aids, tactile feedback, and simplified user interactions to ensure that the system remains accessible and effective for users with hearing impairments.

The PPI evaluation also provided an opportunity to discuss potential ethical considerations and concerns related to the FridgeVision system. Attendees emphasized the

importance of ensuring data privacy, security, and informed consent when deploying the system in real-world settings. They highlighted the need for clear communication and transparency regarding data collection, storage, and usage to maintain user trust and confidentiality.

Overall, the qualitative feedback received through the PPI evaluation at the University of Nottingham Dementia Showcase was highly valuable and informative. The insights and suggestions provided by patients, caregivers, healthcare professionals, and researchers will be instrumental in shaping the future development and refinement of the FridgeVision system. The feedback regarding camera placement and the high prevalence of hearing loss among individuals with dementia will be carefully considered and addressed to ensure that the system remains accessible, user-friendly, and effective for its target user group.

The PPI evaluation underscored the potential of the FridgeVision system to make a meaningful difference in the lives of individuals with dementia and their caregivers, while also highlighting the importance of ongoing collaboration, user-centered design, and consideration of the specific needs and challenges faced by this population. By incorporating the insights gained from the PPI evaluation, the FridgeVision system can be further refined to better serve and support individuals with dementia in managing their food inventory and maintaining proper nutrition.

Chapter 5

Conclusion and Future work

5.1 Conclusion

The FridgeVision system represents a significant advancement in the field of assistive technologies for individuals with dementia, offering a comprehensive and user-centric solution for food management and nutritional support. By leveraging state-of-the-art computer vision techniques, deep learning models, and IoT technologies, FridgeVision effectively addresses the challenges faced by this vulnerable population in maintaining a healthy and well-organized food inventory.

The meticulous design and implementation of the multi-stage pipeline, encompassing object detection using YOLOv8 [16], segmentation with SAM [21], latent space analysis powered by ResNet18 [14], and personalized recipe recommendations using the Llama3 language model [41], have yielded promising results. The experimental evaluations and user feedback have validated the system's accuracy, robustness, and potential impact on the daily lives of individuals with dementia and their caregivers.

The object detection module, fine-tuned on a diverse dataset of food items, has demonstrated exceptional performance in accurately identifying and localizing objects within refrigerator images. The segmentation component, leveraging the SAM model, has shown precise pixel-level delineation of food items, enabling accurate isolation and analysis. The latent space analysis, utilizing ResNet18 and cosine similarity, has proved effective in tracking changes in refrigerator contents over time, providing valuable insights into con-

sumption patterns and potential areas for improvement.

Moreover, the integration of user-centric design principles, voice interaction capabilities, and personalized recipe recommendations has enhanced the system's usability and accessibility for the target demographic. The positive feedback received through the PPI evaluation at the University of Nottingham Dementia Showcase has further validated the potential of FridgeVision in making a meaningful difference in the lives of those affected by dementia.

The FridgeVision system not only addresses the immediate challenges of food management and nutritional support but also contributes to the broader goal of promoting independence, well-being, and quality of life for individuals with dementia. By alleviating the burden on caregivers and empowering users to maintain a well-stocked and organized refrigerator, FridgeVision exemplifies the transformative potential of assistive technologies in dementia care.

5.2 Future work

While the FridgeVision system has demonstrated promising results, there are several avenues for future research and development to further enhance its capabilities and impact:

1. Expanding the Training Dataset: One of the key areas for future work is to expand the training dataset used in the FridgeVision system. Increasing the diversity and size of the training data can further improve the accuracy and robustness of the object detection and segmentation models. Future efforts should focus on incorporating a wider range of fridge environments, food items, and packaging variations into the dataset. Strategies for dataset expansion include collaborative data collection, advanced data augmentation techniques, transfer learning, and continuous data collection as the system is deployed in real-world settings.

2. Longitudinal Studies: Conducting large-scale, long-term evaluations of the FridgeVision system in real-world settings will provide valuable insights into its effectiveness, user acceptance, and long-term impact on the lives of individuals with dementia and

their caregivers. These studies will help assess the system's performance, identify areas for improvement, and validate its potential to support independent living and enhance quality of life for the target population.

3. Personalization and Adaptation: Incorporating advanced machine learning techniques, such as transfer learning and online learning, can enable the FridgeVision system to continuously adapt and personalize its recommendations based on user feedback, preferences, and evolving needs. By leveraging user interactions and feedback, the system can learn to provide more tailored and context-aware assistance, enhancing its usability and effectiveness for individuals with dementia.

4. Multi-modal Interaction: Exploring the integration of additional input modalities, such as voice commands and gesture recognition, can enhance the accessibility and usability of the FridgeVision system for individuals with varying levels of cognitive and physical abilities. Incorporating natural language processing and computer vision techniques to support multi-modal interaction can make the system more intuitive and user-friendly, catering to the diverse needs of individuals with dementia.

5. Integration with Smart Home Ecosystems: Seamlessly integrating the FridgeVision system with existing smart home platforms and IoT devices can provide a more comprehensive and holistic approach to assisted living. By enabling communication and data exchange between FridgeVision and other smart home components, such as ambient sensors, activity monitors, and medication reminders, the system can offer coordinated support across different aspects of daily life, promoting safety, independence, and well-being for individuals with dementia.

6. Nutritional Analysis and Guidance: Collaborating with nutritionists and healthcare professionals to incorporate advanced nutritional analysis and personalized dietary guidance can further enhance the FridgeVision system's ability to promote healthy eating habits and address the specific nutritional needs of individuals with dementia. By integrating knowledge from domain experts, the system can provide more accurate and evidence-based recommendations, supporting the maintenance of optimal

nutrition and overall health.

7. Explainable AI: Developing explainable AI techniques to provide transparent and interpretable insights into the decision-making process of the FridgeVision system can increase user trust, facilitate informed decision-making, and support effective communication between the system, users, and caregivers. By offering clear explanations and rationales behind its recommendations and actions, the system can foster a better understanding and acceptance of its assistance among users and stakeholders.

8. Ethical Considerations: Continuously engaging with stakeholders, including individuals with dementia, caregivers, healthcare professionals, and policymakers, is crucial to address the ethical implications of deploying assistive technologies like FridgeVision. Future work should prioritize the protection of user privacy, autonomy, and dignity, ensuring that the system aligns with ethical principles and legal regulations. Regular consultations and collaborations with ethicists and legal experts can help navigate the complex landscape of developing and deploying responsible and trustworthy assistive technologies.

By pursuing these future research directions and collaborating with interdisciplinary teams, the FridgeVision system can continue to evolve and make a meaningful difference in the lives of individuals with dementia and their caregivers. The insights gained from this research have the potential to inform the development of assistive technologies across various domains, ultimately contributing to the creation of more inclusive, empowering, and supportive environments for individuals with cognitive impairments.

In conclusion, the FridgeVision system presented in this thesis showcases the transformative potential of leveraging cutting-edge computer vision, deep learning, and IoT technologies to address the challenges faced by individuals with dementia in managing their food inventory and maintaining proper nutrition. Through its innovative approach, robust performance, and positive reception from stakeholders, FridgeVision paves the way for the development of more advanced, personalized, and impactful assistive solutions in the

future. By continuing to push the boundaries of interdisciplinary research, prioritizing the needs and well-being of individuals with dementia, and addressing ethical considerations, we can work towards a future where assistive technologies like FridgeVision become an integral part of comprehensive dementia care, empowering individuals to live with dignity, independence, and improved quality of life.

Bibliography

[1] Al-Sarawi, S., Anbar, M., Abdullah, R., and Al Hawari, A. B. Internet of things market analysis forecasts, 2020–2030. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)* (London, UK, 2020), IEEE, pp. 449–453.

[2] Amugongo, L. M., Kriebitz, A., Boch, A., and Lütge, C. Mobile computer vision-based applications for food recognition and volume and calorific estimation: A systematic review. *Healthcare (Basel) 11*, 1 (2022), 59.

[3] Botella, C., Etchemendy, E., Castilla, D., Baños, R. M., García-Palacios, A., Quero, S., Alcañiz, M., and Lozano, J. A. An e-health system for the elderly (butler project): A pilot study on acceptance and satisfaction. *CyberPsychology & Behavior 12*, 3 (2009), 255–262. PMID: 19445633.

[4] Chen, L.-C., et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence 40*, 4 (2017), 834–848.

[5] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. *ArXiv abs/2002.05709* (2020).

[6] Cipriani, G., Danti, S., Picchi, L., Nuti, A., and Fiorino, M. Daily functioning and dementia. *Dement Neuropsychol 14*, 2 (2020), 93–102.

[7] Dawadi, P. N., Cook, D. J., Schmitter-Edgecombe, M., and Parsey, C. Automated assessment of cognitive health using smart home technologies. *Technology and Health Care 21*, 4 (2013), 323–343.

[8] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 9 (September 2010), 1627–1645.

[9] GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. *The Lancet. Public Health 7*, 2 (2022), e105–e125.

[10] Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH, USA, June 2014), pp. 580–587.

[11] Hafiz, R., Alajlani, L., Ali, A., Algarni, G., Aljurfi, H., Alammar, O., Ashqan, M., and Alkhashan, A. The latest advances in the diagnosis and treatment of dementia. *Cureus 15*, 12 (2023), e50522.

[12] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 9726–9735.

[13] He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn, 2018.

[14] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.

[15] Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems 33* (2020), 6840–6851.

[16] Jocher, G., Chaurasia, A., and Qiu, J. Ultralytics yolov8, 2023.

[17] Jocher, G., Stoken, A., Borovec, J., NanoCode012, Chaurasia, A., TaoXie, Changyu, L., V, A., Laughing, tkianai, yxNONG, Hogan, A., lorenzomammana, AlexWang1900, Hajek, J., Diaconu, L., Marc, Kwon, Y., oleg, wanghaoyang0106, Defretin, Y., Lohia, A., ml5ah, Milanko, B., Fineran, B., Khromov, D., DingYiwei, Doug, Durgesh, and Ingham, F. ultralytics/yolov5: v5.0 - yolov5-p6 1280 models, aws, supervise.ly and youtube integrations. `https://github.com/ultralytics/yolov5/releases/tag/v5.0`, April 2021.

[18] Kim, I. The framework for implementation of smart refrigerators using iot. *Transportation 1*, 2 (2016), 3.

[19] Kingma, D. P., and Welling, M. Auto-encoding variational bayes, 2022.

[20] Kirillov, A., He, K., Girshick, R. B., Rother, C., and Dollár, P. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019).

[21] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything, 2023.

[22] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (2012), F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc.

[23] Laker, B., Patel, C., Budhwar, P., and Malik, A. Six steps to innovate remotely. *MIT Sloan Management Review* (2021).

[24] Lan, X., Lyu, J., Jiang, H., Dong, K., Niu, Z., Zhang, Y., and Xue, J. Foodsam: Any food segmentation. *IEEE Transactions on Multimedia* (2024), 1–14.

[25] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature 521* (2015), 436–444.

[26] Lindeza, P., Rodrigues, M., Costa, J., Guerreiro, M., and Rosa, M. M. Impact of dementia on informal care: a systematic review of family caregivers' perceptions. *BMJ Supportive & Palliative Care 14*, e1 (2024), e38–e49.

[27] Liu, W., et al. Optimizing eating performance for older adults with dementia living in long-term care: A systematic review. *Worldviews on evidence-based nursing 12*, 4 (2015), 228–235.

[28] Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation, 2015.

[29] Meiland, F., Innes, A., Mountain, G., Robinson, L., van der Roest, H., García-Casal, J. A., Gove, D., Thyrian, J. R., Evans, S., Dröes, R.-M., et al. Technologies to support community-dwelling persons with dementia: A position paper on issues regarding development, usability, effectiveness and cost-effectiveness, deployment, and ethics. *JMIR Rehabilitation and Assistive Technologies 4*, 1 (2017), e1.

[30] Mihailidis, A., Boger, J. N., Craig, T., et al. The coach prompting system to assist older adults with dementia through handwashing: An efficacy study. *BMC Geriatrics 8* (2008), 28.

[31] Moreno-Fergusson, M., Caez-Ramírez, G., Sotelo-Díaz, L., and Sánchez-Herrera, B. Nutritional care for institutionalized persons with dementia: An integrative review. *Int J Environ Res Public Health 20*, 18 (2023), 6763.

[32] Power, R., Prado-Cabrero, A., Mulcahy, R., Howard, A., and Nolan, J. M. The role of nutrition for the aging population: implications for cognition and alzheimer's disease. *Annual review of food science and technology 10* (2019), 619–639.

[33] Prapulla, S., Shobha, G., and Thanuja, T. Smart refrigerator using internet of things. *Journal of Multidisciplinary Engineering Science and Technology 2*, 1 (2015), 1795–1801.

[34] Quintana, E., and Favela, J. Augmented reality annotations to assist persons with alzheimer's and their caregivers. *Personal and Ubiquitous Computing 17* (2013), 1105–1116.

[35] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection, 2016.

[36] Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[37] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 10674–10685.

[38] Samsung. Family hub. `https://www.samsung.com/us/explore/family-hub-refrigerator/overview/`, 2022. Accessed: 1-Jul-2022.

[39] Selkoe, D. J. Preventing alzheimer's disease. *Science 337*, 6101 (2012), 1488–1492.

[40] Singla, A., Yuan, L., and Ebrahimi, T. Food/non-food image classification and food categorization using pre-trained googlenet model. pp. 3–11.

[41] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023.

[42] Vanus, J., Belesova, J., Martinek, R., et al. Monitoring of the daily living activities in smart home care. *Human-Centered Computing and Information Sciences 7* (2017), 30.

[43] Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022.

[44] Wang, H., Zhu, Y., Adam, H., Yuille, A., and Chen, L.-C. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision* (Cham, 2020), Springer International Publishing.

[45] Wang, K., Ti, Y., Liu, D., and Chen, S. A smart refrigerator architecture that reduces food ingredients waste materials and energy consumption. *Ekoloji 28*, 107 (2019), 4873–4878.

[46] Woods, B., et al. Cognitive stimulation to improve cognitive functioning in people with dementia. *The Cochrane database of systematic reviews 2* (2012), CD005562.

[47] Wu, C.-C., et al. Artificial intelligence in dementia: A bibliometric study. *Diagnostics (Basel, Switzerland) 13*, 12 (Jun 2023), 2109.

[48] Xie, L., Yin, Y., Lu, X., Sheng, B., and Lu, S. Fridge: An intelligent fridge for food management based on rfid technology. In *UbiComp 2013 Adjunct - Adjunct Publication of the 2013 ACM Conference on Ubiquitous Computing* (09 2013), pp. 291–294.

[49] Yousong Zhu, Xu Zhao, C. Z. J. W. H. L. Food det: Detecting foods in refrigerator with supervised transformer network. *Neurocomputing 379* (2020), 162–171.

[50] Yuan, Y., Chen, X., Chen, X., and Wang, J. Segmentation transformer: Object-contextual representations for semantic segmentation, 2021.

[51] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., and Shum, H.-Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605* (2022).

[52] Zhu, X., et al. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).

Appendix A

Poster Presented at the University of Nottingham Dementia Showcase

Title: FridgeVision: Innovative Food Management for People with Dementia using Advanced Computer Vision Techniques

Authors: Parth Ashwinbhai Bhalodiya, Joshua Goulton, Armaghan Moemeni

Affiliation: School of Computer Science, University of Nottingham

Abstract: FridgeVision is an innovative food management system designed to assist individuals with dementia in maintaining proper nutrition and independence. By leveraging state-of-the-art computer vision techniques and deep learning models, FridgeVision accurately identifies, organizes, and monitors food items within a refrigerator. The system incorporates object detection using YOLOv8, segmentation with the Segment Anything Model (SAM), latent space analysis powered by ResNet18, and personalized recipe recommendations using the Llama3 language model. FridgeVision also features voice interaction capabilities and expiry date detection functionalities to enhance usability and prevent food waste. The poster presents the system architecture, key components, and evaluation results, highlighting the potential of FridgeVision in improving the quality of life for individuals with dementia and their caregivers.

Key Features:

- Accurate object detection and localization of food items using YOLOv8

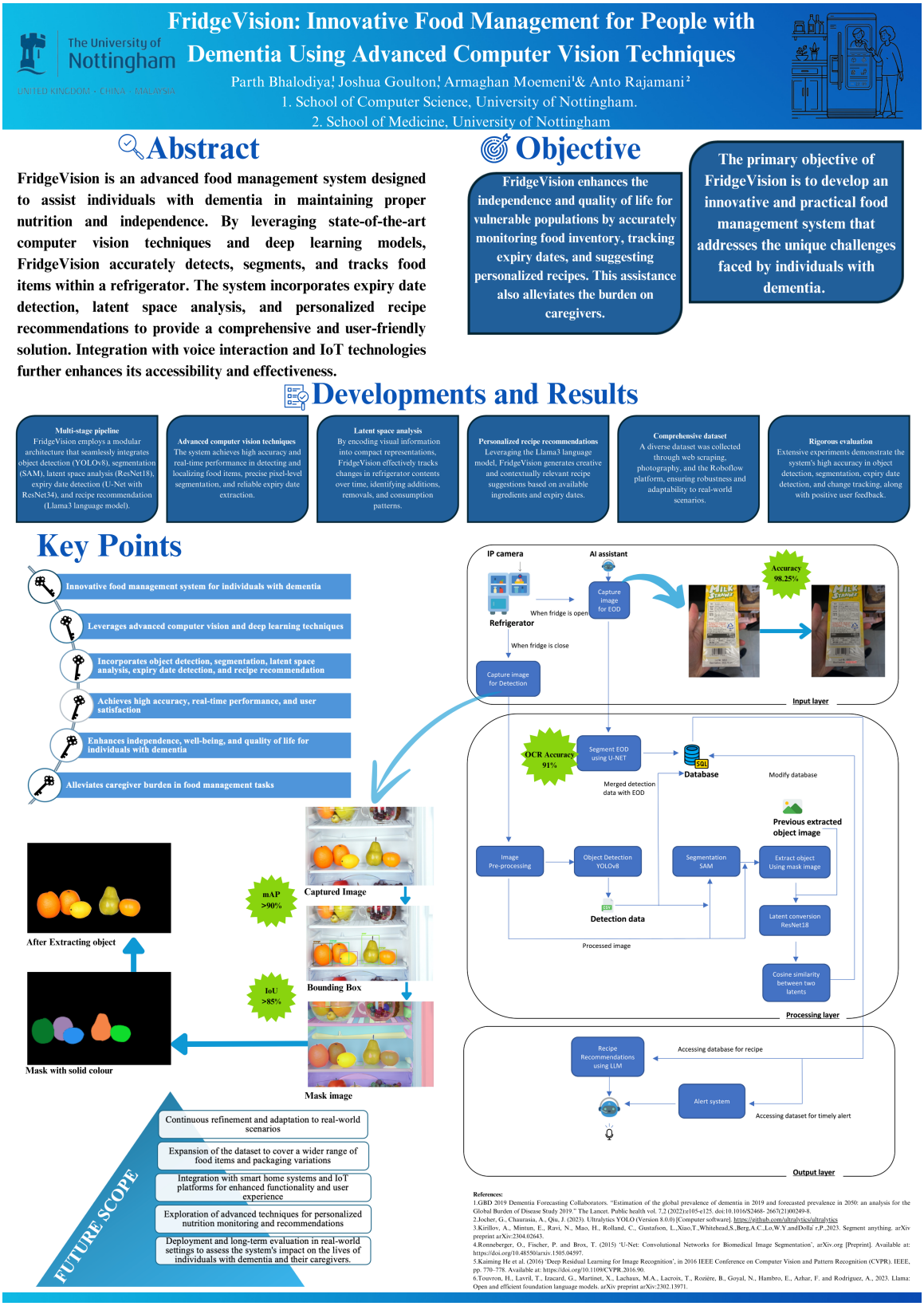

Figure A.1: FridgeVision poster

- Precise pixel-level segmentation of food items with the SAM model

- Latent space analysis for tracking changes in refrigerator contents over time
- Personalized recipe recommendations based on available ingredients and user preferences
- Voice interaction for hands-free control and accessibility
- Expiry date detection and alerts to minimize food waste and promote freshness

Evaluation Results:

- High accuracy and real-time performance in object detection and segmentation tasks
- Effective tracking of food consumption patterns and changes in refrigerator contents
- Positive user feedback on recipe recommendations and overall system usability
- Successful integration of voice interaction and expiry date detection functionalities

Conclusion: FridgeVision represents a significant step forward in the development of assistive technologies for individuals with dementia. By combining advanced computer vision techniques with user-centric design principles, FridgeVision offers a comprehensive solution for food management and nutritional support. The system has the potential to enhance independence, promote healthy eating habits, and alleviate caregiver burden. Ongoing research and collaborations with healthcare professionals aim to further refine and expand the capabilities of FridgeVision, ultimately improving the lives of those affected by dementia. **Acknowledgments:** The authors would like to thank the participants and caregivers who provided valuable feedback during the development and evaluation of FridgeVision. This research was supported by the School of Computer Science at the University of Nottingham.

Appendix B

Dataset Overview

B.1 Data Collection Sources

The FridgeVision dataset was curated through a combination of the following sources:

- Web Scraping: Food item images were collected from various online sources, including grocery store websites, food catalogues, and recipe platforms. Python scripts utilizing the Selenium library were employed to automate the web scraping process.
- Fridge Environment Photography: Real-world fridge images were captured by volunteer participants using various camera devices, such as smartphones and digital cameras. The fridge images encompassed a diverse range of fridge models, organizational styles, and food item arrangements.
- Roboflow Platform: The Roboflow platform was leveraged to annotate and augment the collected images. Roboflow's annotation tools facilitated the labeling of food items with bounding boxes, polygon annotations, and relevant metadata. Data augmentation techniques, including rotation, flipping, scaling, and color adjustments, were applied to expand the dataset and enhance model robustness.

B.2 Dataset Statistics

The key statistics of the FridgeVision dataset are as follows:

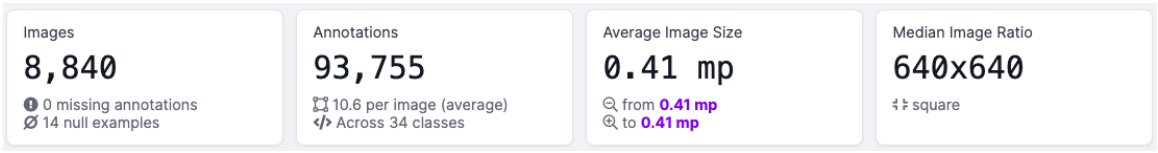

Figure B.1: Dataset overview

- Total Number of Images: 8,840
- Number of Unique Food Item Categories: 34
- Train-Validation-Test Split Ratio: 85:10:5

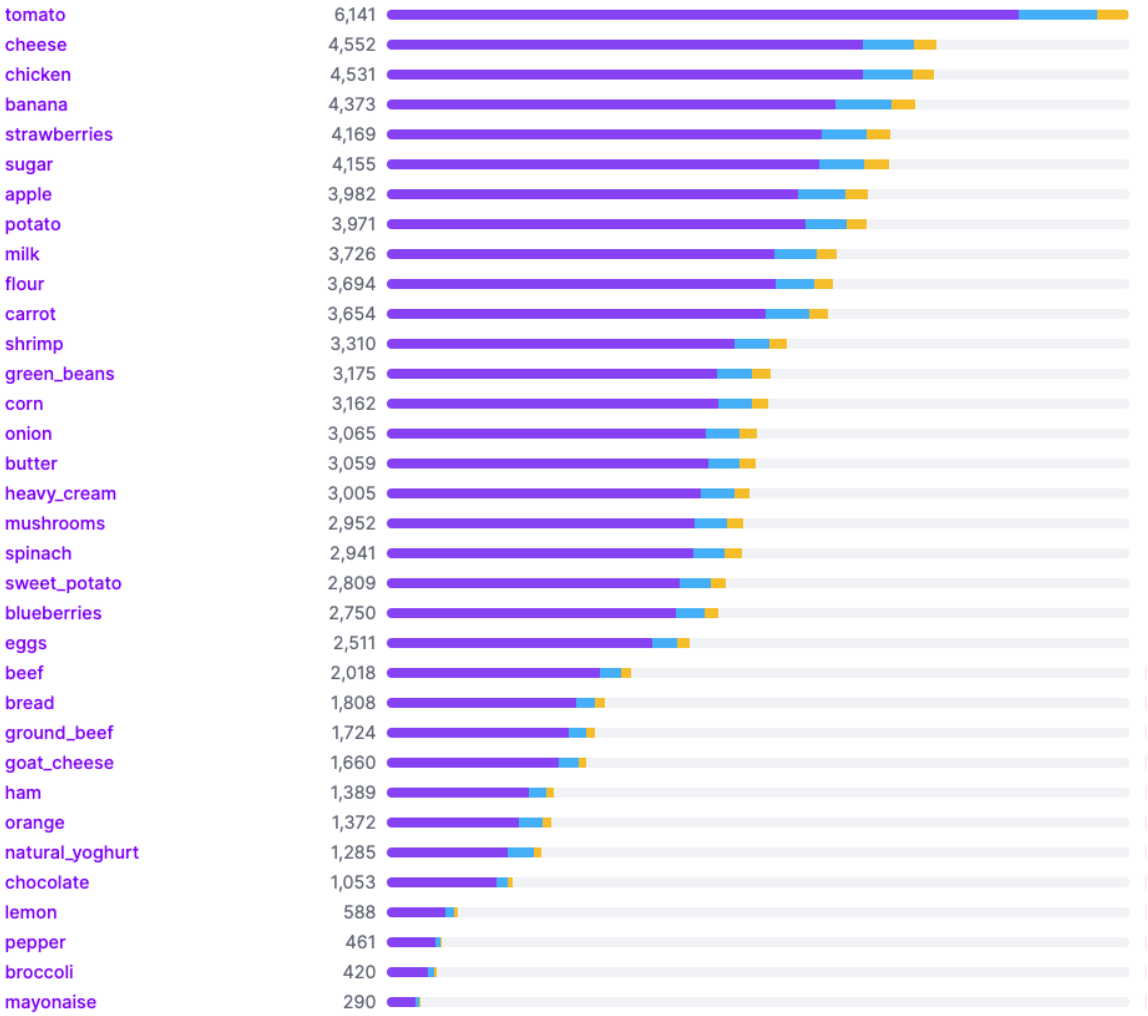

Figure B.2: Data Distribution category-wise

The dataset was split into training, validation, and testing subsets to facilitate model development, hyperparameter tuning, and performance evaluation. The training set was

used to train the deep learning models, while the validation set was employed for model selection and hyperparameter optimization. The testing set was utilized for the final evaluation of the trained models, providing an unbiased assessment of their performance on unseen data.

B.3 Data Privacy and Ethics

Ensuring data privacy and adhering to ethical principles were of utmost importance during the dataset collection and usage. The following measures were taken:

- Informed Consent: Participants involved in fridge photography were provided with detailed information about the study's purpose, data usage, and their rights. Written consent was obtained prior to data collection.

- Anonymization: Any personal or identifying information captured in the fridge images was meticulously anonymized or removed to protect participant privacy. Secure Data Storage: The dataset was stored on secure servers with access restricted to authorized project members. Robust security protocols were implemented to prevent unauthorized access or misuse.

- Ethical Approval: The data collection process underwent rigorous ethical review and received approval from the relevant institutional review boards, ensuring compliance with ethical guidelines and research best practices.

B.4 Continuous Dataset Expansion

To maintain the FridgeVision system's robustness and adaptability, a continuous data collection and expansion strategy was adopted. As the system is deployed in real-world environments, new fridge images and annotations are collected with user consent. This ongoing effort allows for the incorporation of a broader range of food items, packaging variations, and fridge configurations, enabling the system to evolve and improve its performance based on real-world data.

The expanded dataset is regularly utilized to fine-tune and update the object detection, segmentation, and latent space analysis models, ensuring the system's accuracy and reliability in the face of emerging food products and packaging designs. By curating a comprehensive, diverse, and ethically sourced dataset, the FridgeVision system aims to provide a reliable and effective solution for food management and nutritional support tailored to the needs of individuals with dementia and their caregivers.