AN ABSTRACT OF THE RESEARCH-PROJECT-IN-LIEU-OF-THESIS OF

Aidan Beery for the degree of Master of Science in Computer Science presented on December 10th, 2025.

Title: Towards Automatic Household Food Waste Measurement with the Smart Compost Bin

Abstract approved: _____

Patrick Donnelly

The rising amounts of food waste across the world constitute an environmental, social, and economic disaster. Over 60% of food waste is generated by household consumers, yet we currently lack reliable means to quantify this waste. A lack of publicly available image datasets of commingled food waste has prohibited researchers from leveraging advances in computer vision for the task of automatic food waste measurement. We present an AI-assisted compost bin that automatically measures kitchen compost waste by collecting waste images alongside environmental and weight data. We streamline the food waste annotation routine, combining weighing and imaging into a single step and enabling users to verbally label these waste images using onboard automatic speech recognition. We provide a companion mobile application as a platform for both instance-level food waste image annotation and an analytics dashboard to inform users of their waste habits. We will deploy 55 of these compost bins in a forthcoming field study to curate a large, novel dataset of commingled food waste images. Additionally, we explore the feasibility of using pretrained CLIP models for open-vocabulary food waste object detection and assess the viability of transferring a model trained on uneaten food images to compost image recognition tasks.

©Copyright by Aidan Beery
December 10th, 2025
All Rights Reserved

Towards Automatic Household Food Waste Measurement with the Smart Compost Bin

by

Aidan Beery

A RESEARCH-PROJECT-IN-LIEU-OF-THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented December 10th, 2025
Commencement June 2026

Master of Science research-project-in-lieu-of-thesis of Aidan Beery presented on December 10th, 2025.

APPROVED:

Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my research-project-in-lieu-of-thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my research-project-in-lieu-of-thesis to any reader upon request.

Aidan Beery, Author

ACKNOWLEDGEMENTS

The authors appreciate the efforts of Oregon State University undergraduate engineering students Daniel Eastman, Jake Enos, Vaylen Melendez, Kevin Ng, and William Richards, who contributed greatly in the development and fabrication of the smart compost bin; Daniel Lau, Kimberly Markley, Blake Pugh, and Micah Stalberg, who contributed to the development of the CompostSense app; and Vi Vuong, who assisted in the annotation of our compost image dataset.

This research was funded by the Foundation for Food and Agriculture Research and the Kroger Zero Hunger Zero Waste Foundation (FFAR Grant ID: 22−000420).

TABLE OF CONTENTS

Page

1 Introduction 1

2 Food Waste Impacts 3
2.1 Mitigating Food Waste 5
2.2 Food Waste Measurement 6
2.3 Waste Measurement Methods 6
2.4 Technological Approaches to Waste Measurement 8
2.5 Food Image Recognition 10
2.6 Datasets 10
2.7 Food Image Recognition Models 11

3 Smart Compost Bin 14
3.1 Hardware Design 14
3.2 Data Pipeline 17
3.3 CompostSense App 19

4 Contrastive Models for Food Object Detection 22
4.1 Model Architecture 23
4.1.1 CLIP Model Design 24
4.1.2 Open-vocabulary Object Detection with YOLO + CLIP 24
4.2 Results 25
4.2.1 Datasets 26
4.2.2 Food Image Classification 27
4.2.3 Food Object Detection 28
4.2.4 Compost Object Detection 31

5 Discussion 32
5.1 Limitations 34
5.2 Future Work 35

6 Conclusions 39

Bibliography 39

LIST OF FIGURES

Figure Page

3.1 The smart compost bin, our household food waste measurement device. . 14

3.2 3D renders of the main components of the smart compost bin. 16

3.3 System diagram detailing data collection, annotation, and storage workflow for the smart compost bin. Images are captured from the compost bin, annotated via the CompostSense mobile app, and stored in our central API server. 18

3.4 User interface for the CompostSense mobile app 20

3.5 Compost image annotation workflow in the CompostSense mobile app . . 21

4.1 Architecture diagram of our two-stage open-vocabulary food object detection model. Food images are processed by YOLOv11, then each detection is encoded by CLIP before being classified based on a similarity search over a vector database of precomputed food object embeddings. 25

LIST OF TABLES

Table		Page
2.1	Comparison of primary food waste measurement approaches.	7
2.2	List of common food image recognition datasets.	10
3.1	Description of sensors used in the smart compost bin.	17
4.1	Summary of food image datasets. "Combined" is the union of Food201, UECFoodPix, and FoodSeg103. We create the Food Waste Test dataset from manually annotated images taken from the smart compost bin.	26
4.2	Comparison of CLIP image classification models to current state-of-the-art on FoodX-251 validation set. Values marked with "−" for scores which were not presented in original publication.	28
4.3	Evaluation of CLIP models for food object classification. Localizations are provided by annotator labels, isolating the classification component of our food object detection task.	29
4.4	Performance of a fine-tuned YOLOv11 model on food object detection tasks.	29
4.5	Evaluation of YOLOv11 + `jina-clip-v2` for food object detection.	30
4.6	Evaluation of YOLOv11 + `jina-clip-v2` for open-vocabulary object detection on food waste images taken from the smart compost bin.	31

Chapter 1: Introduction

Food waste is a significant driver of environmental and socioeconomic harms. Estimates suggest that nearly 1.05 billon tons of edible food is lost each year, totaling $1 trillion USD in economic loses and contributing 8-10% of total greenhouse gas emissions[1]. 19% of all edible food available to consumers is wasted, with 60% of this waste occurring in households [36]. This gross inefficiency perpetuates global food insecurity, as 783 million people people suffer from chronic hunger.

A lack of descriptive, accurate measurements for food waste has limited the capability of targeted interventions to effectively reduce food waste. Policymakers and stakeholders lack information on the quantity and composition of consumer food waste streams, and this information deficit complicates the evaluation of waste reduction measures [104, 25]. The decentralized nature of household food waste streams poses a unique challenge for waste measurement studies. Existing methods for household food waste measurement typically rely on volunteers or researchers to directly weigh and categorize waste [97]. These manual observational methods are expensive and time-consuming, with low inter-study correlation due to a lack of standardization in waste measurement protocols [5].

To address the scarcity of annotated food waste data, we design and manufacture a smart compost bin to automatically measure household food waste. This consumer-scale compost bin is designed to accurately quantify food waste from individual households, providing detailed information about waste behaviors. Using a combination of imaging and environmental sensors, this device scans each disposal, providing stereoscopic image pairs alongside thermal images, ambient temperature and humidity, total volatile organic compound (tVOC) count, and disposal weight. We design a two-stage compost image annotation framework, leveraging an onboard automatic speech recognition model for easy-to-use unstructured image labeling and a companion mobile smartphone application for detailed instance segmentation annotation. Our annotation workflow enables the collection of a large set of coarsely labeled compost images with a subset of precise instance-level annotations, facilitating waste composition analysis without requiring

[1]Including both agricultural food loss and consumer food waste.

users to manually maintain a food waste journal. From these annotations, our goal is to produce a dataset of high-quality, annotated images of commingled food waste, enabling the development of computer vision models for automatic food waste recognition. Using this individualized food waste data, we provide a waste analytics dashboard to users, visualizing their waste habits and encouraging sustainable consumption behavior through reflection.

Towards the development of a computer vision model for food waste recognition, we evaluate the capability of contrastive image-language models such as CLIP to accurately classify food waste instances. CLIP models, with their natural language grounding and web-scale pretraining, facilitate continual learning over an open set of possible categorical labels from unstructured image-text labels, such as those provided by the transcribed speech annotations from our compost bin. We leverage the abundance of large, high-quality classification and instance segmentation datasets for images of uneaten food to train an open-vocabulary food image recognition model. Our model achieves state-of-the-art performance on food image classification tasks, and we evaluate the transfer of this model to the more challenging task of food object detection. We use a fine-tuned YOLOv11 convolutional neural network [44] to detect potential food objects within a scene, then pass these cropped detections CLIP for classification, with the goal of leveraging the zero-shot image classification capabilities of CLIP for our compost object detection task.

Chapter 2: Food Waste Impacts

The United Nations (UN) Sustainable Development Goals (SDG) outline critical environmental targets to coordinate an international response to the growing threats of global climate change and resource scarcity. UN SDG 12.3 declares the following: *"By 2030, halve per capita global food waste at the retail and consumer levels and reduce food losses along production and supply chains, including post-harvest losses"* [26]. Food waste and loss (FLW) is an immediate threat to global food security and climate, yielding significant economic losses while leading to significant direct and indirect greenhouse gas emissions. Residential households, the largest single food waste stream, comprise an estimated 60% of global food waste [36]. However, this waste source is also the most difficult to accurately measure, as a lack of standardized waste definitions and waste measurement systems results in inaccurate underestimates of total household food waste [25].

The United Nations Environmental Program (UNEP) is the custodian organization responsible for tracking food waste, defined as food disposed or spoiled at the retail, commercial, or residential level [36]. In 2024, the UNEP published their report on international progress towards SDG 12.3, finding that member nations have not made sufficient progress towards consistent, accurate food waste measurement, and therefore any analysis of progress towards food waste reduction is severely limited. Despite these deficits in measurement, the UNEP estimates global food waste volumes at 1.05 gigatonnes, with 631 million tonnes of this waste occurring in the household [36]. These estimates are drawn from six methods: direct measurement, waste composition analysis (WCA), volumetric assessment, mass balance, food item counting, and waste journaling (see Table 2.1 for a summary of these methods). Food waste estimates are aggregated for each nation by scaling estimates derived from national environmental agency measurements and academic studies. Only 20% of recent academic household food waste measurement studies involve direct measurement, many instead relying on indirect estimates of household food disposals [25]. The proportion and total volume of global household food waste, therefore, is poorly understood, and has been consistently under-

estimated on both a national and international level [5] [36].

Global FLW is responsible for 8-10% of global greenhouse emissions, contributing to adverse environmental impacts from climate change [34]. Consumer food waste which is directed to the landfill decomposes at a much faster rate than most landfilled waste, resulting in large methane emissions which are unable to be mitigated by typical methane capture systems [49]. Indirectly, food waste is responsible for greenhouse gas emissions from the energy and resources required to produce food which is ultimately wasted. 24% of freshwater used for agriculture and 30% of agricultural land is used to produce food which goes uneaten [34] [50].

This misappropriation of food resources also contributes to global hunger. An estimated 783 million people suffer from chronic hunger, with 2.4 billion experiencing moderate or severe food insecurity [27]. This has considerable downstream impacts on public health, economic advancement, and childhood development. 22.3% of children are impacted by growth stunting due to food insecurity and malnutrition. Even when considering only edible/avoidable food waste, the total volume of this food waste could provide up to 1.3 meals per person in hunger each day [36].

The cumulative effects of food loss and waste on national economies is drastic. It is estimated that FLW accounts for $1 trillion of global economic losses [27]. For US consumers, these losses are estimated $13.27 per-capita per-day, with $3.62 of these losses attributed to preventable, edible food waste. [15]. Reducing food waste by 50% as targeted by UN SDG 12.3 could have measurable benefits for household consumers, yielding a savings equivalent to 9% of a home consumer's annual food budget [87]. FLW reduction presents a $2.3 trillion annual global market opportunity, with the possibility of creating jobs and capital through public-private partnerships to reduce food waste and redirect wasted food to more energy-efficient channels [2].

To realize this economic potential and offset the insidious environmental and social harms of food waste, we must first design effective solutions for quantifying and analyzing food waste. These solutions must be accurate while providing detailed composition analysis of wasted food in order to benefit researchers and policymakers in their efforts to design sustainable interventions.

2.1 Mitigating Food Waste

Attaining accurate and detailed household food waste measurements is crucial to encourage sustainable consumer behavior. Papargyropoulou et. al. defines a hierarchy of food waste disposal behaviors [73]. Food waste is best mitigated by avoiding the waste entirely, either by preventing overallocation or redistributing food resources for re-use. Composting repurposes wasted food for agricultural use. Encouraging household consumers to compost their food waste does not prevent environmental harms from wasted food, but instead allows for the food to re-enter the agricultural lifecycle as nutrients for growing new food, mitigating greenhouse gas emissions caused by the landfilling of food waste [47]. Despite this, the U.S. EPA estimates only 3.7% of US residential food waste is composted [3]. Optimally, food waste measurement systems would differentiate avoidable and inevitable food waste, directing users towards sustainable habits to mitigate avoidable food waste while promoting the composting of inevitable food waste.

Household food waste drivers can be broadly classified into one of the following wasteful behaviors: household planning failures, commercial incentives to over-provision, food storage & preparation waste, unconsumed food, and disposal inefficiencies [75]. Consumer food waste should not be considered as a single behavioral failure on the part of the individual consumer. Instead, consumer food waste is a result of systemic inefficiencies in food distribution and utilization [1]. Consumers are often placed in positions where the economically optimal solution, in terms of both financial and labor costs, is to discard food, as the utility cost of reusing that food item would be too high [60]. To mitigate this waste, users can be encouraged to adopt sustainable consumption habits. Behavioral nudges can motivate users to modify their food utilization [80], however such suggestions must be tailored to the user's waste behavior and cultural norms, [10, 98], further motivating the need for individualized waste quantification rather than aggregate measures.

Directly measuring individual households' food waste both provides data to drive waste reduction interventions and serves as an intervention in of itself [79]. By informing consumers of the real quantity of their waste, it may motivate them to make more sustainable decisions with regards to food provisioning, preparation, and consumption. A lack of technological solutions for household consumers has been identified as a key limiting factor to household food waste reduction. In response, sustainability researchers

call for "affordable technology and tools (e.g., smart kitchen tools) to improve food management" [98]. Such technologies have seen some initial adoption in institutional kitchens and been demonstrated to be effective at reducing wasteful behaviors [106].

2.2 Food Waste Measurement

Data scarcity and a lack of standardization in waste measurement is a key limiting factor to reducing global food waste [35]. There are a growing number of food waste measurement studies aiming to address this lack of data. However, these approaches tend to rely on indirect measurements of food waste such as mass balancing, or self-reported waste measures. These methods tend to underreport actual waste quantities [97], however 80% of recent food waste measurement studies rely on such methods [104]. Such studies disproportionately survey high-income nations, with a particular lack of data on consumer waste behaviors in developing nations. Among the few studies which do rely on direct measures of food waste, they typically involve manual sorting and weighing of post-consumer waste, which is expensive, time-consuming, and potentially hazardous. Additionally, these studies are highly variable and incomparable, due to a lack of standardized definitions of food waste [25].

2.3 Waste Measurement Methods

Despite making up 60% of total food waste, household food waste is the most frequently underestimated and least understood waste stream [104]. There has been an increasing interest among sustainability researchers to attain accurate and representative data of household food waste streams [25]. Existing methods for measuring household food waste can be broken down to one of six categories, as shown in Table 2.1.

Due to the significant cost of precise waste measurement, food waste is often inferred indirectly. Mass balance estimates of food waste are calculated using preexisting data, taking the total volumes of food inventory in an area and measuring that against average food disposal and consumption volumes to approximate a food waste rate for some region [11, 40, 96]. Questionnaire methods similarly rely on indirect waste measures, asking participants to reflect on their waste habits and report what they believe to be their total waste volumes. Such approaches do not provide sufficient granularity [36] and

Study Type	Procedure	Merits	Weaknesses
Mass Balance	Estimate waste from proxy data	Scalable, low-cost	Highly inaccurate, insufficient granularity.
Questionnaires	Participants reflect and self-assess recent waste volumes	Low-cost, low participant commitment	Users may misrepresent and underreport waste
Waste Composition Analysis	Researchers manually collect, sort, and measure food waste	High precision, detailed food waste classification	Very high cost, no standard definitions for food categories
Food Waste Diary	Participants maintain a food waste log	Reduce reliance on participant memory	High participant burden, underreported waste
Direct Weighing	Participants weigh waste before disposal	Precise waste volume measurement	Moderate participant burden, lacks categorical data.
Waste Imaging	Participants take photos of waste	Accurate waste composition data, prevents misreporting	Requires complex CV models, lacks mass estimation.

Table 2.1: Comparison of primary food waste measurement approaches.

often do not correlate to direct waste measurements [92, 30].

Direct measurements quantify food waste at time of disposal. In household food waste studies, users are typically tasked with maintaining a written record of each discarded food item [81, 22]. These food waste diaries have been found to report much higher volumes of waste than self-assessment questionnaires [38], though they may still not accurately represent waste volumes [97], possibly due to significant annotator burden [91]. To reduce this burden, prior approaches have requested users to only record the total weight of their food waste, rather than itemizing each disposal [100]. However, reducing the granularity of food waste data inhibits the efficacy of targeted waste re-

duction interventions [6]. Self-reported waste measurements, including those including weight measurement, underestimate actual food waste by anywhere from 7% to 40% of real waste volume [76]. In these self-assessed measurement studies, social desirability bias and selection bias may result in the sample population misrepresenting the target data [76, 90].

To mitigate these biases, researchers have conducted waste measurement studies where researchers manually sort and quantify participant's disposals, requiring no behavioral change from the user. These waste composition analysis (WCA) studies enable researchers to attain accurate categorical data on a household's food waste [18], however the significant labor overhead in manual waste sorting prevents this approach from being applied at scale. Additionally, cultural norms and inconsistent definitions as to what defines avoidable waste or inedible food, as well as varying descriptors used to define certain foods and varying levels of granularity in food categorization (e.g. "vegetable" v.s. "green onion") leads to most WCA studies failing to be statistically representative of a target population's waste behavior [100, 51].

2.4 Technological Approaches to Waste Measurement

Existing methods to measure household food waste fall short of delivering accurate and scalable estimates. Precise WCA approaches are too labor-intensive to be implemented broadly, while low-cost questionnaires underrepresent waste volumes. Self-assessed direct waste measurements balance cost and accuracy, but misreporting due to annotator burden yields a misrepresentation of food waste data [76]. In response, researchers have attempted to leverage technological solutions to reduce this annotator burden and improve the accuracy of household-level food waste measurement systems.

Imaging-based approaches to food measurement have garnered significant research interest, particularly for nutritional tracking [69, 28, 66] and agricultural [74, 29] purposes. Initial attempts to extend the use of camera systems to food waste measurement have shown promise, with photographic measures correlating strongly to laboratory-measured waste [64, 17, 97]. Images of household food waste can be captured by the user's smartphone [83, 20, 97, 37], enabling waste data collection at scale. Users report these systems to be less burdensome and time-consuming than maintaining a food waste diary [84]. Direct weight measurement can be used in tandem with waste photography

methods to record accurate mass density and waste types [17].

Many existing household food waste imaging approaches share significant shortcomings. Many studies omit weight measurement entirely or require a two-step process where waste is weighed and imaged separately. There have been attempts at simplifying this process by estimating waste volumes from waste images [57, 41], however such methods do not attain sufficient accuracy to replace direct food mass measurement. Additionally, these proposed methods rely on waste image classification to be performed by researchers [37] or users [20]. This image labeling is an expensive and time-consuming endeavor, limiting the scale at which imaging-based waste measurement solutions can be deployed.

Dedicated food waste measurement appliances, designed for commercial food waste management, have enabled accurate waste measurement while minimizing annotation labor in busy kitchen environments. Initial approaches rely on mass measurement using smart kitchen scales [21, 93] to measure food preparation waste and customer plate waste. Iterations on these approaches include annotation tablets, with which staff can classify a record by menu item at measurement time [39]. For institutions with less rigid menus, overhead cameras have been added to smart kitchen scale systems to record waste images [65, 82]. The deployment of food waste measurement devices to institutional kitchens has led to waste reduction rates between 16% [82] and 64% [93], demonstrating the efficacy of waste measurement as a means to waste reduction [79]. This data can be used to generate data-driven food waste analytics, providing institutions with targeted recommendations to further reduce waste [71].

Similar technologies have been applied to the measurement of inorganic waste and common household refuse. Various prototypes for household-scale smart waste bins incorporate lid-mounted camera systems to capture images of waste at disposal time [94, 101]. Computer vision methods can be used to automatically sort and classify this waste, enabling the separation of organic and inorganic waste [94, 101] or recyclables from landfill waste [56, 7]. Smart waste bins have similarly been designed and deployed for large municipalities, using weight and environmental sensors to measure bin utilization and automate waste sorting [31]. Such bins can be combined into a robust Internet-of-Things (IoT) sensor network to aide municipal waste management authorities in waste generation prediction, waste retrieval route optimization, and precise evaluation of waste reduction interventions [31, 86, 43, 95].

2.5 Food Image Recognition

Food waste image classification systems are limited by the lack of publicly available, high-quality food waste image datasets. However, such datasets do exist for images of edible food, and there has been substantial work towards developing computer vision models for food identification [53]. As our approach seeks to leverage food image recognition models for food waste recognition, we briefly review major food image datasets, listed in Table 2.2, and survey existing food image recognition models.

Dataset	**Data Type**	**# Images**	**# Classes**
FoodX-251 [46]	Classification	118,000	251
Food-101 [9]	Classification	101,000	101
ISIA Food-500 [68]	Classification	399,726	500
Recipe-1M+ [63]	Image Caption	13,765,679	N/A
FoodSeg103 [102]	Segmentation	9,490	103
UECFoodPix [72]	Segmentation	10,000	102
Food-201 [70]	Segmentation	12,000	201

Table 2.2: List of common food image recognition datasets.

2.6 Datasets

Food image classification datasets contain text-image pairs, with each image categorized as a food item or dish. Often, these datasets are constructed by crawling image search engines [46] or food and cooking websites [63] as these platforms already host user-generated food images, searchable by food category. Food-101 [9], based on the now defunct `foodspotting.com`, collects the first 1000 images for the top 101 most popular food categories from this platform. FoodX-251 [46] extends Food-101's ontology, using WordNet [67] to identify similar food descriptor labels, then scraping images from Google Images which match these identifiers. 18,000 of these 118,000 images were manually validated by human annotators. A fine-tuned ResNet-101 model achieves 17% top-3 accuracy on this dataset, demonstrating both the difficulty of differentiating similar food classes and the need for more robust image classification for highly multi-class problems.

The Food-500 dataset identifies a list of 500 food categories from Wikipedia, and

collects nearly 400,000 matching food images from online image search platforms [68]. The authors propose SGLANet (Stacked Global-Local Attention Network), a two-stage transformer-based model which extracts both local and global spatial features, which achieves an 89% top-5 accuracy on the Food-500 dataset.

For many food recognition and waste analysis tasks, researchers care about not only the food item present in the image, but the location of this food item in a complex image. In the context of waste analysis, this can enable automatic sorting systems [48] or provide waste composition analysis from commingled waste bins [101]. In the case of food image recognition, the ability to identify and segment multiple food items in a single image enables better calorie estimation and helps identify multiple dishes present on a single plate, as is common in Western dining settings [63]. Such approaches rely on instance segmentation, a class of computer vision models which identify objects within an image and separate these objects by drawing pixel-level masks around the items of interest. Each segmented item can then be assigned a categorical label.

Food image segmentation datasets provide food images paired with these segmentation masks, identifying each food item present in an image. Food-201, an extension of Food-101, provides 12,000 segmented food images with 201 food item classes [70]. Bounding box annotations are provided by crowdsourced annotators, and segmentation annotations are generated using GrabCut [85], an algorithm for automatic foreground object segmentation. Each image contains one or more food item label. Similar datasets, such as UECFoodPix [72] and FoodSeg103 [102], each provide food image segmentation masks manually drawn by human annotators, rather than automated boundaries drawn by GrabCut.

2.7 Food Image Recognition Models

Various computer vision methods have been proposed for the task of food image classification, including support vector machines using hand-crafted features [14], convolutional neural networks (CNNs) [89, 28, 66], and vision transformers [103]. Food image recognition is often applied to calorie estimation tasks for nutrition monitoring [4]. These systems generally rely on user's smartphones for data entry and imaging, providing a convenient platform for automatic caloric intake measurement.

CNNs remain the most popular approach for food image recognition tasks, with

58% of recent food image recognition studies utilizing a CNN for food recognition [19][1]. Broadly, these systems are trained on publicly available food datasets, and fine-tuned on an annotated subset of food images [19, 89, 66], with many early models being pre-trained on the ImageNet [23] dataset of images [89, 66, 28], which, incidentally, contains a set of at least 1000 food categories [105].

Food recognition models can be broadly categorized into image classification [89, 105], object detection [45, 12, 14, 13], and instance segmentation [66] models. Image classification tasks simply apply a categorical label to an image, and are only accurate in images which contain a single food entity (e.g. a dish or single ingredient). Many food image recognition tasks involve the classification of multiple food entities on a single plate, as is the case in cafeteria tray images [13, 45] or ingredient recognition in plated food images [63]. For these tasks, object detection and instance segmentation models provide granular multiclass food recognition in complex scenes. Modern approaches to food object detection tasks often rely on YOLOv8 [44], a member of the You-Only-Look-Once (YOLO) family of single-shot object detection CNN models pre-trained on the COCO dataset [52]. Such models often fine-tune with only 1,000 to 20,000 annotated images, a relatively small training set given the complexity of the task, and attain a mAP@50[2] of 0.67 [45] to 0.70 [12], indicating this model's robust capability for adaptation to the downstream task of food instance localization in complex scenes. YOLOv8 has also been evaluated in other downstream food image recognition tasks, including plate waste in cafeteria environments [33] and damaged food detection for food processing quality control [24].

Convolutional neural networks, though very powerful, fail to encapsulate the complex relationships between food categories. As each image is classified into a separate category, the model retains no knowledge of similarity between classes. In food descriptors, this is

[1]The food recognition systems survey conducted by Dalakleidi et. al, 2022 [19], was presented at a time before the widespread adoption of novel computer vision methods such as vision transformers or contrastive image embedding models, and this proportion may have changed since the study's publication in 2022.

[2]Mean Average Precision (mAP) is a measure of an object detection model's ability to correctly localize an object of interest in an image. In object detection model evaluation, each predicted bounding box is classified as correct if some percentage of the predicted bounding box overlaps the with the ground truth bounding box. CV practitioners refer to this overlap as intersection-over-union (IoU). mAP@50 refers to the average precision, taken as a mean over each class, with an IoU threshold of 50, meaning at least 50% of a predicted bounding box must overlap with the ground truth to be classified as a correct prediction.

a key challenge, as a lack of standardized food ontology has limited the reproducibility of prior, manual food waste annotation studies [100]. Contrastive image-language models enable the multimodal fusion of language data and visual representations. Models such as CLIP train a Transformer text encoder model to represent language information, and a vision transformer or CNN for image encoding [77]. The encoders are fine-tuned to align text descriptors with matching image data, enabling the model to learn a unified embedding representation. Using these multimodal embeddings enables zero-shot classification, allowing a model to classify a sample belonging to a previously unseen category. In the context of food waste recognition, in which there exists potential hundreds of thousands of unique classes, and the same visual input may be given disparate labels by different users (e.g. "Green Onion" v.s. "Scallion"), CLIP-style models show potential for robust performance. When evaluated on Food-101, CLIP-ViT attains 95.4% classification accuracy [77].

Chapter 3: Smart Compost Bin

To address the limitations of existing methods for household food waste measurement, we design an AI-assisted household compost bin for the measurement of residential food waste. Our device implements an array of sensors to capture information about the visual, physical, and chemical qualities of each disposal. When combined with a robust compost image recognition model, this device will facilitate automatic quantification and classification of food waste items, with the goal to provide a similar level of granularity as waste composition analysis studies with minimal human input. To drive the development of such a model, we implement a easy-to-use dictation-based waste annotation framework, building a dataset of labeled food waste images to train and iteratively refine our compost image recognition model.



(a) Interior (Render) (b) Underneath (Render) (c) Front (Final) (d) Rear (Final)

Figure 3.1: The smart compost bin, our household food waste measurement device.

3.1 Hardware Design

The smart compost bin, shown in Figure 3.1, is a 3D printed device with a hinged lid and removable polycarbonate waste receptacle. Our expectation is that users place the

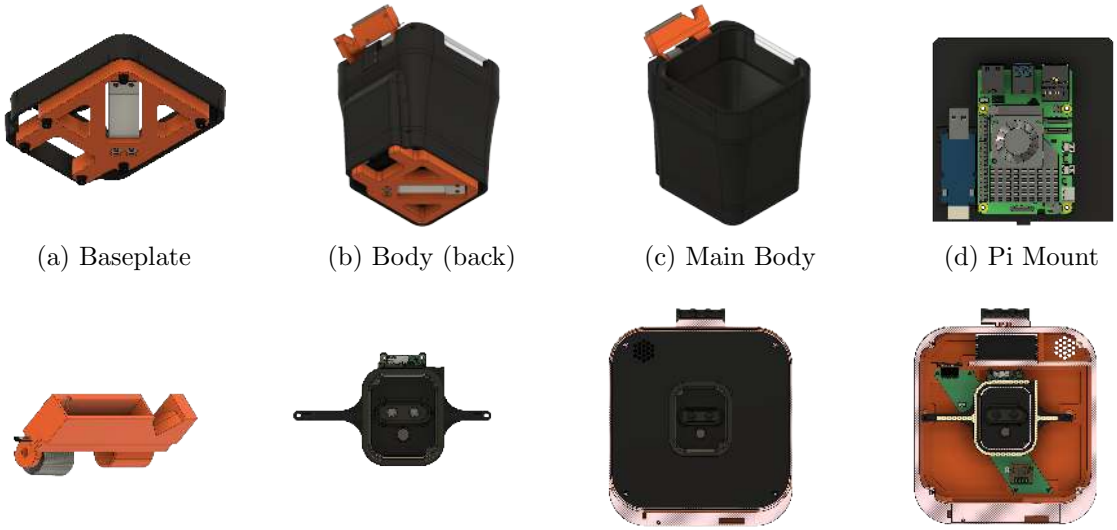
bin on their kitchen counter to temporarily hold food waste until its later disposal in an outdoor compost bin. As the bin will sit in a user's home, near their food preparation area, we design the bin with ease of use and sanitation in mind. We consider the following four core usability principles:

Sanitation: The device must enable the user to safely dispose of their waste and clean contamination. Waste is contained in a removable and dishwasher-safe inner liner, enabling easy disposal and cleaning.

Odor Mitigation: The device must mitigate offensive odors typically emitted from food waste. The lid of the bin features a passive air filtration system, directing emissions through a carbon pellet filter to capture volatile organic compounds. A rubber gasket maintains a seal when the lid is closed.

Convenience: Users should be able to quickly interact with the device while cooking or performing household tasks. Our bin features a springed hinge and hands-free latch, allowing users to deposit refuse while their hands are occupied.

Streamlined Data Collection: First-stage data annotation should be convenient and hands-free, reducing annotator burden. The bin records the user's speech and transcribes audio to a text label using an onboard speech-to-text model.



(a) Baseplate (b) Body (back) (c) Main Body (d) Pi Mount

(e) Hinge (f) Camera (g) Lid Assembly (h) Lid Cutaway

Figure 3.2: 3D renders of the main components of the smart compost bin.

The bin is made primarily of polylactic acid (PLA) and polyethylene terepthalate glycol (PETG) plastics. A C2-grade load cell is placed on the bottom of the bin to precisely measure the mass of each disposal (Figure 3.2a and Figure 3.2b). The main body (Figure 3.2c) holds the internal disposal bin, with the Raspberry Pi 5 microcontroller and audio adapter mounted in a separate protected enclosure in the rear (Figure 3.2d). The lid, which houses imaging and gas sensors, is connected to the main body using a spring-loaded hinge (Figure 3.2e). An Intel RealSense camera is selected for our imaging task, providing stereoscopic image pairs which can be used for food volume estimation [55]. An infrared dot-matrix camera and a BME688 gas sensor (Figure 3.2f) are also present in the lid. The color camera is protected from condensate and refuse by a polycarbonate cover, while the thermal camera is shielded by a infrared-transparent plastic. These imaging sensors are positioned in the lid such that they face towards the waste receptacle when the lid is closed (Figure 3.2g). A summary of the sensors included in the smart compost bin is provided in Table 3.1. An air duct allows emissions from the waste to first pass by our gas sensor, then pass through our passive carbon pellet filtration system before escaping the bin through ventilation holes. Each device has a total bill-of-materials cost of $468.

When the user opens the bin by pressing the latch, the lid gently pivots open using a tensioned spring. The user would then deposit their waste item and shut the lid, with the latch re-engaging to lock the lid in the closed position. The bin detects lid closure with a magnetic hall effect sensor, after which LEDs will illuminate the interior of the waste bin. A stereoscopic color image pair and a thermal image will be captured alongside ambient temperature and humidity, total volatile organic compounds (tVOC), and the mass of the deposited item.

We use a speech-based annotation workflow, which has been suggested to reduce annotator burden in food waste measurement studies [59]. During the data collection procedure, the bin will tell the user to describe the disposal, using the voice prompt "Please describe this item". We record a four-second window where the user's speech will be recorded, during which they should describe the wasted food. An audible chime indicates to the user when this recording has ended. This speech data is processed on-board the smart compost bin's Raspberry Pi, using an open-source C++ implementation[1] of the Whisper automatic speech recognition model [78].

Component Model	**Purpose**
Raspberry Pi 5 4GB	Connects to sensors to capture data.
Intel RealSense D401	Captures stereoscopic image pair.
Bosch BME688	Measures temperature, humidity, and tVOC.
NAU7802 24-bit ADC	Measures signal from load cell.
110° MLX90640 IR Camera	Generates thermal image and heatmap.
Waveshare USB Sound Card	Plays audio prompt, records user response.
US5881LUA Hall Effect Sensor	Detects when the lid is opened and closed.
10kg C3 Single-Point Load Cell	Measures mass of food items.

Table 3.1: Description of sensors used in the smart compost bin.

3.2 Data Pipeline

To facilitate the deployment of a distributed set of smart compost bins for a large-scale household food waste data collection study, we implement an integrated software system for data collection and annotation, summarized in Figure 3.3. Our system consists

[1] `https://github.com/ggml-org/whisper.cpp`

of initial data collection and preprocessing onboard the bin's Raspberry Pi, a central backend server for data ingress and storage, and a mobile smartphone application for data visualization and annotation (see section 3.3).

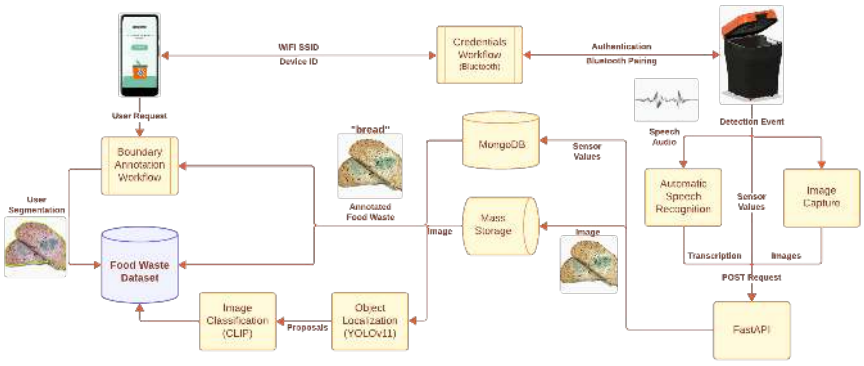


Figure 3.3: System diagram detailing data collection, annotation, and storage workflow for the smart compost bin. Images are captured from the compost bin, annotated via the CompostSense mobile app, and stored in our central API server.

Each bin is connected to a user's home Wi-Fi network using our companion smartphone app (see section 3.3). After each disposal, the compost bin takes the aggregate image and sensor data, alongside the text transcription from the user's spoken label, and prepares a payload to be uploaded to our central database server. Data processing and upload happens asynchronously, allowing the user to initiate another food waste disposal routine immediately after the recording window for the previous scan completes. These data payloads are written to the Raspberry Pi's storage and queued for upload to the API server. In cases where internet connectivity has been lost, the device will continue to operate, maintaining the payload queue on disk until a connection to the API server has been re-established. Once data has been successfully uploaded, the LEDs in the bin will flash green to indicate to the user that the data collection has completed.

Annotated compost records are stored on a centralized database server, with each compost bin client interfacing with our backend system via a REST API to protect our database from malformed or malicious operations. The bins send a POST request to the API to upload a compost scan, and the API will validate the format and integrity of this data. Images are written to disk, while sensor data and logging information is written

to MongoDB, a non-relational (NoSQL) database. When a user opens the CompostSense mobile app, that app will request from the API the most recent images associated with that user's compost bin. The user can provide image segmentation annotations as demonstrated in Figure 3.5b. If provided, the mobile client will upload these segmentations and categorical labels to the API server, where the detailed annotation is appended to the associated compost scan's database record.

3.3 CompostSense App

We design CompostSense, a companion application to the smart compost bin, as a cross-platform mobile app. The app provides a conveinent interface for users to interact with their compost bin, improving our food waste measurement protocol by providing the following functionalities:

Configuration & Setup The app allows users to set up their compost bin, connect it to their home Wi-Fi network, and reset the device in case of errors.

Detailed Image Annotations Using the smartphone's touch input, users can provide instance-level compost image annotations, which we use to refine our food waste object localization model (see section 4.1).

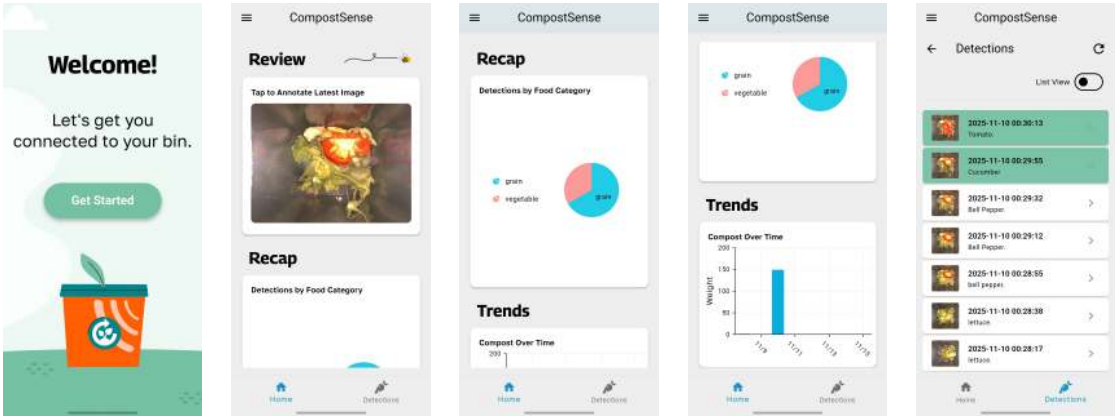
Waste Monitoring Analytics CompostSense informs users of waste habits and total waste volumes over time through a convenient analytics dashboard.

On first launch, the CompostSense app will guide the user through initial setup of their bin (see Figure 3.5a). The app will initiate Bluetooth pairing with the user's compost bin, and prompt the user to provide a Wi-Fi SSID and credentials. Once the bin is connected to the user's home network, the app will exit the first-time setup and load the compost analytics dashboard.

This dashboard (see Figure 3.5e) informs users on their general food waste trends. Waste items are categorized into food groups (e.g., fruits, nuts, meats) and the user is presented with a circle chart detailing the proportions of their waste by food group. Waste trends over time are given by total waste mass, as measured by the weight sensor built into the smart compost bin. This dashboard displays the most recent image taken from the bin, nudging the user towards an annotation task.

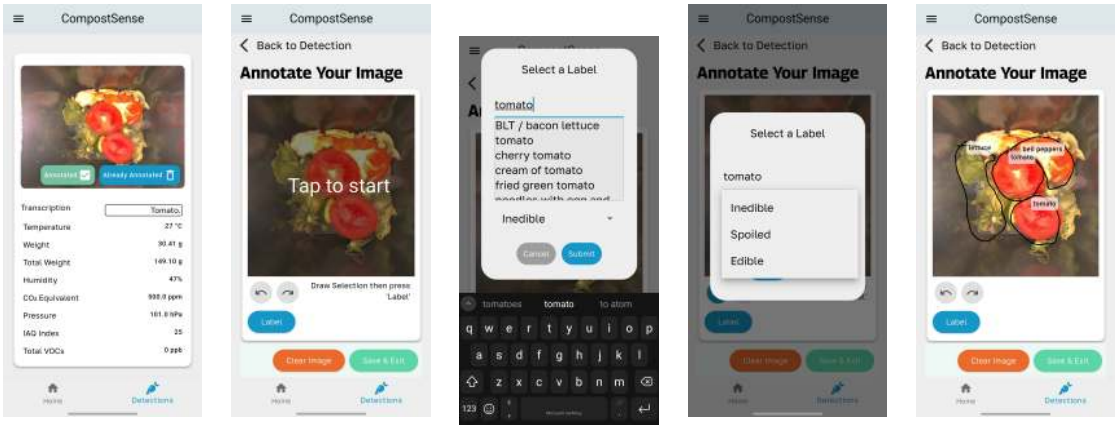
Under the Detections tab (see Figure 3.5d), users can select a compost scan and view details about the sensor data from this scan (see Figure 3.5c). This panel will display a text transcription of the user's spoken label. We allow users to edit this label to correct mistakes from accidental misinputs or erroneous transcriptions by our automatic speech recognition model. From here, the user can tap the compost image to enter a detailed image annotation interface, from which we collect fine-grained instance segmentation labels.

We augment our coarsely labeled image dataset with fine-grained, localized annotations to iteratively refine our computer vision model's ability to separate food waste instances from commingled waste images. In the CompostSense annotation interface (see Figure 3.5), users are asked to circle each food item present in the image and assign a categorical label to this object (see Figure 3.5b). This interface also allows users to mark each food item as spoiled, inedible, or edible, allowing the quantification of avoidable food waste relative to the overall waste output.



(a) Splash Screen (b) Home Screen (c) Waste Groups (d) Waste Mass (e) Detections

Figure 3.4: User interface for the CompostSense mobile app



(a) Scan Example (b) Labeling Screen (c) Food Label (d) Waste State Tag (e) Labeled Image

Figure 3.5: Compost image annotation workflow in the CompostSense mobile app

Chapter 4: Contrastive Models for Food Object Detection

Our proposed device in chapter 3 iterates on existing household food waste measurement systems by simultaneously weighing and imaging waste, enabling users to quickly label wasted food items as part of their standard kitchen routine. This approach is an improvement on existing food waste measurement protocols, but still requires some amount of human input in the form of dictated labels. An entirely automatic system for food waste measurement would improve user continuance intention [97] and enable robust waste behavior analytics without the need for human annotation. Various imaging-based approaches to household food waste measurement have been evaluated (see section 2.4), however none provide a method for automatic waste quantification and classification from images taken in a commingled waste setting (e.g., a waste bin). Our goal is to assess the feasibility of contrastive image-language pretraining (CLIP) models for open-vocabulary compost recognition towards an automated system for food waste measurement.

Images captured from the smart compost bin depict many possible food waste items in a complex, commingled scene. This computer vision task is challenging, as there is a need both for precise localization of food waste objects within an image and for open-vocabulary, fine-grained object classification to capture the many possible types of food a user may dispose of. Though modern models can be trained to recognize hundreds of unique categorical labels, these approaches are still limited by the use of a fixed label set. The descriptive labels used to identify food items can vary significantly, with potentially hundreds of thousands of terms used to describe food and different terms for describing the same food item. Traditional image recognition models struggle to attain robust classification accuracy as the number of potential categories grows [107], motivating the need for an open vocabulary food image recognition model.

Contrastive image-language models leverage broad web-scale pretraining to enable zero-shot, open-vocabulary image classification [77]. Using CLIP image embeddings, our model can make inferences about previously unseen food waste classes, allowing the label space of food waste categories to grow adaptively over time. However, CLIP is designed for image classification tasks in which a caption is paired with a single image. This is

incompatible with our compost object detection task, in which many relevant objects are present in an image. We experiment with a two-stage object detection model, first using a fine-tuned convolutional neural network to identify regions-of-interest in each image, then passing these cropped detections to CLIP for classification.

To date, there exists no publicly available dataset of annotated food waste images. In response, we rely on datasets of uneaten food images (see subsection 4.2.1) to train a computer vision model for food image recognition with the goal of evaluating the transferability of food image recognition models to compost image recognition tasks. We evaluate our two-stage object detection model first on these datasets of uneaten food, then attempt to transfer this model to our compost object detection task. We determine that training on uneaten food images alone is insufficient for an accurate compost object detection model, highlighting the need for a large-scale annotated compost image dataset.

4.1 Model Architecture

CLIP models use jointly trained text and image encoders, learning a multimodal feature space by aligning the output embeddings from these encoders. These models are often pretrained using an image captioning task, in which the goal is to maximize the cosine similarity of embedding vectors from matching text-image pairs. By grounding the image feature space in natural language context, CLIP models demonstrate strong zero-shot image classification performance, enabling these models to classify images belonging to classes not seen in the training data. The use of a separate text encoder allows for an open set of labels to be provided for an image, rather than being restricted to a fixed set of categorical labels. The ability to learn from unstructured image captions is uniquely important for our application, as our users provide compost image labels in the form of spoken-word annotations (see section 3.2).

We first compare various CLIP implementations detailed in subsection 4.1.1 for food image classification, then evaluate the use of these CLIP embeddings for food object classification. Then, we combine a fine-tuned instance of YOLOv11 with CLIP as described in subsection 4.1.2 in an attempt to leverage the robust object localization capabilities of this large pre-trained CNN to enable the use of CLIP for classification of detected objects in a scene.

4.1.1 CLIP Model Design

We consider two CLIP models for our food image recognition tasks, including `CLIP-ViT-B-32` and an open-source alternative, `jina-clip-v2`. `CLIP-ViT-B-32` uses a 32-layer vision transformer as its image encoder and an uninitialized text encoder model, producing an 512-dimension embedding vector [77]. `Jina-clip-v2` iterates on CLIP by using a pre-trained instance of RoBERTa [54] as its text encoder rather than training a text encoder from scratch. The additional natural language grounding provided by a pre-trained text encoder enables `jina-clip-v2` to achieve greater zero-shot image classification performance than `CLIP-ViT-B-32`. Additionally, the `EVA-02` image encoder used in `Jina-clip-v2` provides a modest improvement in downstream image classification tasks over the standard ViT implementation used by `CLIP-ViT-B-32` [32].

`Jina-clip-v2` uses Matryoshka representation learning to enable the truncation of output embeddings to a meaningful lower-dimension embedding. `CLIP-ViT-B-32`, meanwhile, outputs a fixed 512-dimension output embedding. We consider both the 1024-dimension and 512-dimension embedding vectors from `Jina-clip-v2` for food image classification to assess if the additional representation capacity of a higher-dimension embedding is necessary for accurate food image classification.

As CLIP is an embedding model, it does not directly output categorical labels. To classify food images (and, later, food objects), we use Qdrant[1], an open-source vector database, to retrieve the image embeddings most similar to a given detection. This vector database contains CLIP embeddings generated from food instances and allows for the efficient retrieval of relevant vectors based on cosine similarity. Each food instance vector is stored with a corresponding categorical label in this vector database. To classify a sample, the vector database retrieves the five most similar food instance vectors, assigning the majority class from these retrieved vectors as the categorical label for this detection.

4.1.2 Open-vocabulary Object Detection with YOLO + CLIP

Object detection models identify regions-of-interest within an image and classify each detection, mapping multiple categorical labels to a single image. Unlike image classification

[1] `https://github.com/qdrant/qdrant`

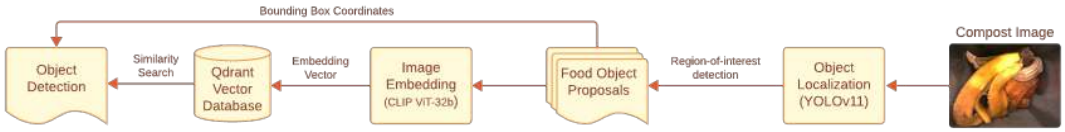


Figure 4.1: Architecture diagram of our two-stage open-vocabulary food object detection model. Food images are processed by YOLOv11, then each detection is encoded by CLIP before being classified based on a similarity search over a vector database of precomputed food object embeddings.

models, which have been thoroughly studied for food image recognition [103, 46, 9, 68], object detection models are able to localize multiple food instances in an image, enabling more accurate quantification of foods from image data. However, common object detection architectures such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) make predictions from a closed set of categorical labels specified at training time, making these models alone insufficient for open-vocabulary compost object detection.

We design a two-stage object localization and classification approach to enable open-vocabulary detection of food objects in commingled scenes. Proposals—bounding boxes indicating the location of food items present in the image—are extracted from a source image using YOLOv11, an open-source CNN model for object detection tasks [44]. We choose YOLOv11 for this task due to its efficient single-shot architecture and deep pre-training on object detection tasks. The fine-tuning of YOLO models for food image recognition tasks is a common strategy well-supported in the literature [45, 12, 33, 24].

4.2 Results

We first evaluate the capability of CLIP-based models for food image classification in subsection 4.2.2. Through these evaluations, we assess if CLIP models produce embedding vectors which encode semantic information relevant for food image recognition tasks. We transfer this model to the more difficult task of food object classification in subsection 4.2.3, assessing CLIP classification performance on annotator-localized food objects, then using YOLOv11 to automatically localize food objects to be later classified by CLIP. Finally, in subsection 4.2.4 we assess the performance of this two-stage YOLOv11

+ CLIP food object detection model on a dataset of annotated compost images captured from the smart compost bin to measure the capability of an open-vocabulary food object detection model to measure food waste.

4.2.1 Datasets

		Training		Evaluation	
Dataset	**# Classes**	# Samples	# Instances	# Samples	# Instances
Food Image Classification					
FoodX-251 [46]	251	118,000	118,000	12,000	12,000
Food Object Detection					
Food201 [66]	201	9,653	15,979	2,439	4,149
UECFoodPix [72]	102	9,000	13,032	1,000	1,575
FoodSeg103 [102]	103	4,983	18,318	2,135	7,697
Combined	389	23,637	47,329	5,574	13,421
Compost Object Detection					
Food Waste Test	77	–	–	467	1600

Table 4.1: Summary of food image datasets. "Combined" is the union of Food201, UECFoodPix, and FoodSeg103. We create the Food Waste Test dataset from manually annotated images taken from the smart compost bin.

We assess the classification accuracy of CLIP on the FoodX-251 dataset, as its large, fine-grained label space and visually similar classes serves as a good proxy for our downstream food waste recognition task [46]. The dataset consists of 251 food categories, using WordNet to expand upon a set of 101 food categories originally given in Food101 [9].

To train our food object detection models, we use a combination of Food201, UECFoodPix-Complete, and FoodSeg103 (see Table 4.1). These datasets contain labeled images of food, with a categorical label and pixel mask segmenting each food instance in the scene. As CLIP is trained for image inputs rather than segmentations, we reduce image segmentation labels to bounding boxes using the minimum and maximum coordinate in the pixel mask. After combining 18 duplicate classes across these datasets, our joint dataset

contains 389 labels. Training and test splits are detailed in Table 4.1.

In order to evaluate the transferability of our food object detection model to compost object detection tasks, we create a dataset of 467 images collected from six smart compost bins. These bins were deployed with beta testers for one to four months. From these 467 images, we annotate 1600 instances across 77 classes. Image labels were transcribed from user speech at time of disposal and manually reviewed for consistency. Class names were normalized for capitalization (Apple v.s. apple) and pluralization ("grape" v.s. "grapes"), but state descriptors ("apple slice" v.s. "apple core") were preserved.

4.2.2 Food Image Classification

We begin by comparing `CLIP-ViT-B-32` and `jina-clip-v2` for food image classification on the FoodX-251 dataset, with the goal of selecting a CLIP configuration for our food object detection model. We compare both the standard 1024-dimension embedding vectors and the truncated 512-dimension embedding vectors output from `jina-clip-v2`, pre-computing a separate vector database from the FoodX-251 training set with each model and evaluating vector similarity search classification from these embeddings.

With 512-dimension embeddings, `CLIP-ViT-B-32` and `jina-clip-v2` achieve near-identical classification performance on the FoodX-251 test set. This indicates that truncated Matryoshka embeddings have similar visual encoding capabilities as those generated by `CLIP-ViT-B-32`. Increasing the embedding vector dimension of `jina-clip-v2` from 512 to 1024 improves food image classification accuracy by 17% (from 0.675 to 0.790). This suggests that larger embeddings may encode spatial information with greater specificity, improving performance in fine-grained food image classification tasks.

When using 1024-dimension embeddings from `jina-clip-v2`, we find that our model attains near state-of-the-art performance on the FoodX-251 dataset. As shown in Table 4.2, our model sightly improves upon the baseline ResNet-101 model, reducing top-3 error from 0.16 to 0.15. We observe only a 2.1% decrease in top-1 classification accuracy on the FoodX-251 validation set compared to the current best publicly available model, a fine-tuned Swin-DR vision transformer [103]. This is despite `jina-clip-v2` not requiring fine-tuning, providing a more compute efficient alternative to the current state-of-the-art. These results imply that the embedding vectors generated from pre-trained CLIP models have the potential to encode semantic information relevant to food

image recognition tasks.

Model	Output Dim.	Top-1 Acc.	Top-5 Acc	Top-3 Error
CLIP-ViT-B-32	512	0.676	0.838	0.21
jina-clip-v2	512	0.675	0.838	0.21
jina-clip-v2	1024	0.790	0.900	0.15
Swin-DR ViT [103]	–	0.811	–	–
ResNet-101 [46]	–	–	–	0.16

Table 4.2: Comparison of CLIP image classification models to current state-of-the-art on FoodX-251 validation set. Values marked with "–" for scores which were not presented in original publication.

4.2.3 Food Object Detection

Next, we assess the use of `CLIP-ViT-B-32` and `jina-clip-v2` for the more difficult task of food object classification. Using the annotated bounding boxes from each food object detection dataset listed in Table 4.1, we generate CLIP embeddings for each instance in the training set and store these in a vector database. As shown in Table 4.3, `jina-clip-v2` (1024-dim.) on average achieves 13% greater food object classification accuracy than `CLIP-ViT-B-32` (512 dim.) (0.61 v.s. 0.54), reinforcing the usefulness of 1024-dimension embedding vectors for fine-grained food image recognition tasks.

Overall, we see a 23% decrease in classification accuracy performance between food object classification and food image classification tasks (0.610 v.s. 0.790). These test sets are not identical, making direct comparison between image classification and object detection results challenging. However, we generally find that that CLIP-style models, pretrained on image classification tasks, demonstrate decreased classification accuracy on food image recognition tasks when given cropped objects without background context.

As both a baseline food object recognition model and a component of our proposed CLIP-based approach, we fine-tune an instance of YOLOv11 on the combined food object detection dataset described in Table 4.1. Our model is trained for 85 epochs on the "Combined" food object detection dataset, using early stopping based on a weighted fitness criteria of $(0.1(\text{mAP50}) + 0.9(\text{mAP50-95}))$. We evaluate this YOLOv11 model's performance against the hold-out subsets of each of our three food object detection

Model	Dataset	Top-1 Acc.	Top-5 Acc.
`CLIP-ViT-B-32` 512-dim. embedding	Food201	0.507	0.651
	UECFoodPix	0.686	0.864
	FoodSeg103	0.500	0.664
	Combined	0.466	0.632
`jina-clip-v2` 1024-dim. embedding	Food201	0.584	0.724
	UECFoodPix	0.741	0.864
	FoodSeg103	0.581	0.738
	Combined	0.521	0.683

Table 4.3: Evaluation of CLIP models for food object classification. Localizations are provided by annotator labels, isolating the classification component of our food object detection task.

datasets, as well as against a combined evaluation set sampled from all three datasets. On this combined dataset, we find that a fine-tuned YOLOv11 model attains a modest mAP50 score of 0.293 as shown in Table 4.4.

Dataset	Precision	Recall	mAP50	mAP50-95
Food201	0.524	0.467	0.340	0.284
UECFoodPix	0.692	0.687	0.200	0.186
FoodSeg103	0.589	0.554	0.342	0.296
Combined	0.584	0.545	0.293	0.251

Table 4.4: Performance of a fine-tuned YOLOv11 model on food object detection tasks.

Having identified `jina-clip-v2` with 1024-dimension embeddings as the optimal CLIP-style model for our food image recognition tasks, we evaluate our two stage YOLO + CLIP model on the food object detection datasets listed in Table 4.1. We compare our model on in-domain food object detection against a fine-tuned YOLOv11 model (see Table 4.4).

In addition to the pretrained `jina-clip-v2`, we fine-tune two instances of `jina-clip-v2`, one on FoodX-251, and one on the "Combined" food object detection dataset detailed in Table 4.1. For each model, we define a image captioning task for fine-tuning, where food images are paired with captions corresponding to their classification label (e.g., an

image of an apple would be paired with the caption "An image of an Apple"). Each pre-training dataset is comprised of positive and negative image-caption pairs, with a positive-to-negative ratio of one. We fine-tune each model for a maximum of five epochs, using matching accuracy as an early-stopping criteria evaluated on a held-out subset of each model's respective finetuning dataset[2]. `jina-clip-v2-foodx251` trains for the full five epochs, while `jina-clip-v2-combined` trains for only one.

Model	Dataset	Precision	Recall	mAP50	mAP50-95
`jina-clip-v2`	Food201	0.446	0.397	0.297	0.261
	UECFoodPix	0.652	0.648	0.188	0.178
	FoodSeg103	0.534	0.503	0.304	0.263
	Combined	0.456	0.425	0.226	0.197
`jina-clip-v2-foodx251`	Food201	0.466	0.415	0.305	0.267
	UECFoodPix	0.541	0.537	0.141	0.156
	FoodSeg103	0.510	0.480	0.291	0.250
	Combined	0.418	0.390	0.219	0.185
`jina-clip-v2-combined`	Food201	0.398	0.355	0.264	0.230
	UECFoodPix	0.554	0.550	0.160	0.145
	FoodSeg103	0.533	0.502	0.295	0.259
	Combined	0.424	0.396	0.207	0.176

Table 4.5: Evaluation of YOLOv11 + `jina-clip-v2` for food object detection.

We use the aforementioned fine-tuned instance of YOLOv11 to extract object proposals from each food image, then generate embeddings from these proposals with `jina-clip-v2`. For each dataset-specific experiment, we build a vector database with image embeddings from that specific dataset's training subset, whereas the "Combined" experiment includes embeddings from all three datasets. The results in Table 4.5 show that, of the three variants of `jina-clip-v2` we evaluate, the base `jina-clip-v2` model attains the best object classification performance, with an average mAP50 score across four benchmarks of 0.254. However, this result is still a 13.61% decrease in mAP50 score compared to our baseline YOLOv11 model in Table 4.4 (0.254 v.s. 0.294).

[2]For `jina-clip-v2-combined`, this is a random 1000-image sample taken from Food201, UECFoodPix, and FoodSeg103.

4.2.4 Compost Object Detection

In return for a trade-off in classification performance, our model provides zero-shot image classification capabilities which our baseline model can not achieve. We leverage these capabilities in an compost object detection experiment, using compost images taken from the smart compost bin as described in subsection 4.2.1. Our model uses YOLOv11 to detect regions of interest in these compost images, then performs vector similarity classification using embeddings generated by `jina-clip-v2`. We add an additional set of vectors to our vector database generated in the "Combined" experiments in Table 4.5, creating a text embedding for each classification label in the set of food waste class name given in our evaluation dataset.

Model	**Precision**	**Recall**	**mAP50**	**mAP50-95**
`jina-clip-v2`	0.291	0.116	0.063	0.076
`jina-clip-v2-foodx251`	0.266	0.106	0.059	0.052
`jina-clip-v2-combined`	0.180	0.072	0.048	0.032

Table 4.6: Evaluation of YOLOv11 + `jina-clip-v2` for open-vocabulary object detection on food waste images taken from the smart compost bin.

We find that our approach, using CLIP to classify detections identified by YOLOv11, does manage to identify some of the compost items in our small evaluation dataset. We find these capabilities noteworthy, as this model was not trained on compost images and the majority of class labels in the compost image dataset were not present in the uneaten food datasets. However, we note that our CLIP models which are fine-tuned for food instance classification, `jina-clip-v2-foodx251` and `jina-clip-v2-combined`, yield lower mAP50 scores than`jina-clip-v2` by 6.3% and 23.8%, respectively. This suggests that a CLIP instance fine-tuned for food image recognition does not necessarily transfer to the task of compost object detection. The lack of transferability between these tasks further motivates the need for a large dataset of annotated food waste images to train a computer vision model for automatic food waste quantification.

Chapter 5: Discussion

We present three key contributions towards the development of an automatic household food waste measurement system. First, we design and manufacture a food waste measurement device with the needs of residential consumers in mind. This device facilitates a standardized food waste measurement study protocol, unifying the waste classification and weighing steps and reducing annotator burden. Our approach improves upon historically expensive and time-consuming waste composition analysis studies, seamlessly imaging each disposal and using state-of-the-art automatic speech recognition technology to allow for the dictation of waste annotations.

To support our food waste measurement device, we provide an integrated software solution to household food waste data collection and analysis. We develop a central database server and API backend to connect our fleet of IoT smart compost bins for data collation and processing. Our companion smartphone application, CompostSense, provides a platform for users to annotate food waste images with instance-level detail using the touch input of their mobile device. This app presents users with detailed information about their waste habits, including food waste volume over time and waste distribution by food group. A personalized waste behavior analytics dashboard enables users to monitor their food waste output, which may lead to personal reflections on one's waste habits and encourage the adoption of more sustainable consumption behaviors [79].

We evaluate the viability of CLIP embedding models for food object recognition tasks towards the development of a compost image recognition system. We build a fast, efficient food image classification model using `jina-clip-v2` to generate embedding vectors for food images from the FoodX-251 dataset. We find that this pre-trained model is able to achieve near state-of-the-art performance compared to a fine-tuned Swin-DR vision transformer [103], with only 2.07% loss in classification accuracy. By using vector similarity search for food image classification, our approach remains competitive with large fine-tuned task-specific models while both being being faster to train and enabling additional functionality, such as zero-shot classification of previously unseen food classes and the addition of new classes to the label space without necessitating retraining.

Using this CLIP-based model, we consider the more difficult task of food object classification. Using human-annotated bounding boxes from Food201, UECFoodPix, and FoodSeg103, we extract object proposals from each image. These cropped detections are out-of-distribution for CLIP models, which are typically trained on image-caption pairs where each image retains its background context. Regions-of-interest cropped from a larger scene lack this background context, instead highlighting only the visual information relevant to the object. We find that `jina-clip-v2` demonstrates an average 23% decrease in classification accuracy when evaluated on our food object classification tasks relative to performance achieved on the FoodX-251 dataset. We hypothesize that the out-of-distribution nature of objects cropped from a larger scene is incompatible with the expected inputs for `jina-clip-v2` and similar CLIP models.

We combine the zero-shot instance classification capabilities of `jina-clip-v2` with the robust object localization performance of YOLOv11 to build an open-vocabulary food object detection model. When compared to the use of a fine-tuned YOLOv11 instance alone, we observe a 13.61% decrease in average mAP50 score across our four benchmark tasks. This indicates that simply substituting the classification predictions from YOLOv11 with those generated by a pretrained CLIP model does not alone yield a stronger classifier for food object detection tasks. In an attempt to mitigate this performance degradation, we fine-tune `jina-clip-v2` on food image captioning (using FoodX-251) and food object captioning (using a combination of Food201, UECFoodPix, and FoodSeg103). However, we find that neither of these approaches improve object detection performance when compared against the baseline pre-trained `jina-clip-v2`. This could indicate that simple image captioning (or object captioning) pretraining is insufficient for improving the quality of CLIP models for downstream food object recognition tasks.

Despite the limitations of combining YOLOv11 and `jina-clip-v2` for food object detection tasks, we do still observe some zero-shot food waste recognition capabilities. Our best model achieves a mAP50 score of 0.063 with a precision of 0.291 on our small dataset of 467 annotated food waste images taken from the smart compost bin, highlighting the feasibility of CLIP-based models to classify out-of-domain and out-of-distribution food waste instances in highly complex and commingled scenes. Additionally, we find that fine-tuning `jina-clip-v2` for food image captioning or food object captioning does not yield an improvement on our compost image recognition task, indicating that train-

ing on images of uneaten food alone is not sufficient to design a robust model for compost image recognition. This further motivates the need for the production of a large, high-quality dataset of annotated compost images to enable the development of an automatic food waste measurement model.

5.1 Limitations

Our ability to train and evaluate food waste object detection models is constrained by the size of our annotated food waste image dataset. This food waste image dataset represents data collection from six smart compost bins deployed for beta testing with volunteer users. These beta deployments were limited in scope, focusing primarily on device functionality rather than the collection of a large set of food waste images. This lack of sufficient data prevents us from fine-tuning a model on compost image data, instead requiring the use of uneaten food image datasets. This prohibits a direct evaluation of the transferability of food image recognition models to compost image recognition tasks, as we lack a fine-tuned compost image recognition model to compare against.

We find that the use of YOLOv11 for object localization restricts the zero-shot inference capabilities of CLIP. Of the 1600 ground-truth instances in our food waste dataset, YOLOv11 only yields 289 proposals, leading to a theoretical maximum recall of 0.18. Our YOLOv11 instance is fine-tuned on a set of 389 food classes, however the fine-tuning datasets consist entirely of images of uneaten food. Many of the post-consumer food waste objects in our evaluation set are not visually similar to their uneaten forms (e.g., an apple core may not resemble an apple). Additionally, many of the food classes identified in our food waste data may fall outside the label space of our fine-tuned YOLOv11 model. This results in fewer food waste objects present in a scene being identified as regions-of-interest.

The use of either food image classification datasets or food object detection datasets for fine-tuning `jina-clip-v2` is not sufficient to improve performance on either in-domain food recognition tasks or out-of-domain compost recognition tasks. In Table 4.5, we find that `jina-clip-v2-combined` achieves worse performance than `jina-clip-v2` without fine-tuning, despite being fine-tuned on training data directly from the target datasets. We hypothesize that CLIP models, which are typically pre-trained on image-text pairs, do not respond well to fine-tuning on food object detection data. The

detections used for fine-tuning `jina-clip-v2-combined` are cropped from a larger scene, and do not include the background context which would typically be present in many image captioning datasets.

Food201, UECFoodPix, and FoodSeg103 are all originally food instance segmentation datasets. We reduce these annotations to bounding boxes with the goal of assessing the feasibility of pretrained CLIP models for food object recognition and compost object recognition. Standard implementations of CLIP are not trained on instance segmentations, and generating embeddings from segmentations with with CLIP or `jina-clip-v2` is not practically feasible. However, we find that in many cases, food instances in commingled scenes are not able to be contained within a bounding box without introducing significant noise into the detection. Our model design was guided by the assumption that any decrease in classification accuracy as a result of the use of imprecise bounding box labels would be compensated for by the broad, robust classification capabilities of contrastive image encoding models, which are pretrained on a broad set of image captioning tasks. However, as we apply this model to compost image recognition tasks, we find that bounding boxes are not sufficient to accurately localize a food waste instance. This is especially true of images taken from the smart compost bin, where food waste items are highly commingled in a relatively small scene. Annotators often found that fully encompassing a food waste item with a bounding box led to the majority of the bounding box area containing visual information irrelevant to the object of interest.

5.2 Future Work

To develop a robust computer vision model for automatic food waste measurement, we must first collect a sufficiently large dataset of annotated food waste images. To date, we have manufactured 55 smart compost bins, with the goal of distributing these units to volunteer households across the state of Oregon. Through the data annotation protocol detailed in chapter 3, we collect a large set of coarse-grained dictated waste labels as well as a small subset of instance-level annotations provided by these volunteers. Through a combination of coarse-grained image-text pairs given by user dictation, and instance segmentation annotations provided through the CompostSense app, we will train future models on this dataset to provide a task-specific open-vocabulary food waste image recognition model without the need to rely on datasets of uneaten food images.

The combination of a pretrained CNN object localization model and a contrastive image-language model for object detection does not improve performance on in-distribution food object detection tasks. The pretrained nature of the localization CNN may fail to detect objects of interest if the object is not visually similar to a class in the CNN's label space, inhibiting downstream task performance by failing to identify regions-of-interest for CLIP to classify. Future implementations of food object detection or compost object detection models would benefit from an end-to-end object detection architecture, rather than separating the object localization and object classification stages. Recent advances in contrastive image-language models have demonstrated the capability for open-vocabulary instance segmentation without relying on a separate localization model. Such approaches have been implemented by extending CLIP's loss function to enable superpixel-level segmentation reconstruction [58], using a decoder head to predict a segmentation map from the CLIP embedding [61], or using knowledge distillation to merge foundation semantic segmentation models such as Segment Anything (SAM) with CLIP [99]. We will evaluate these various techniques for embedding spatial knowledge directly into our CLIP model for compost instance segmentation. Using an segmentation-aware variant of CLIP will both resolve the limitations of using a separate pretrained CNN for food object localization as well as provide greater spatial specificity through the use of instance segmentation labels rather than potentially noisy bounding boxes.

We identify several improvements to the fine-tuning tasks of CLIP models for food image recognition. Rather than only training on image-class label pairs, we propose an additional image-to-image fine-tuning task. As our model's classification relies on the comparison of similarity between image embeddings, fine-tuning the model on food image matching may lead to the generation of image embedding vectors which better encode the visual and spatial information relevant to the differentiation of food classes. Additionally, to optimize our CLIP instance for fine-grained image classification tasks, we will tune our model's fine-tuning loss to actively identify "hard" image-image pairs, for images which belong to highly visually similar but distinct classes [42]. Such a loss function may improve our model's ability to accurately classify visually similar instances, improving our performance on fine-grained food image recognition tasks. Finally, we will investigate the use of large language models to improve the quality of captions paired with each food image by synthetically generating descriptive labels from food class labels [88].

Foundation vision-language models (VLMs) are becoming increasingly common for image understanding tasks. Though many current VLMs struggle with fine-grained image classification [108], they still demonstrate promising open-vocabulary image recognition capabilities thanks to rich natural text grounding and web-scale pretraining [16, 62]. Such models have become capable of instance localization tasks, learning to generate bounding box coordinates or tokenized segmentation masks through task-specific fine-tuning [8]. We will assess the capabilities of such models for food waste segmentation tasks, leveraging our forthcoming compost image dataset to design an instruction tuning task for the adaptation of VLMs for food waste measurement.

Our approach relies on a vector database to store the image embeddings generated by `jina-clip-v2`, classifying detections based on their similarity to other image embeddings stored in this database. We use a cosine similarity search to retrieve five candidate image embeddings for classification. While this approach yields noteworthy classification accuracy performance for food image classification, recent developments in retrieval-augmented generation (RAG) models demonstrate novel methods for the retrieval of semantically relevant embeddings from vector databases. These reranker models accept a query vector and a subset of candidate vectors from the vector database, rescoring this subset of embeddings to improve retrieval relevance[1]. Many reranker models are pre-trained on question-answering tasks, in which a query will take the form of a question about a topic and the reranker model will produce the document which best answers that question. Our image similarity search approach requires a different definition of document relevance than that of a question-answering model, and this domain gap was evident in our preliminary experiments with reranker-augmented food image classification. In future experiments, we intend to pretrain our own reranker model for image similarity retrieval with the goal of improving retrieval-based image classification.

We can use multimodal embeddings from `jina-clip-v2` to encode documents beyond just food images and descriptors, such as recipes. Recipe1M+ provides a robust set of recipes, itemized by ingredients [63]. These recipes could be retrieved by querying a vector database for documents mentioning certain ingredients. The individualized food waste measurements provided by our smart compost bin could enable targeted food waste interventions by providing recipe recommendations based on a user's frequently wasted ingredients. Targeted recipe recommendations could be integrated into the waste

[1] `https://huggingface.co/jinaai/jina-reranker-m0`

analytics dashboard in our smartphone app, offering a convenient interface for food waste reduction through behavioral nudging.

Our long-term goal with the development of the smart compost bin is to provide a food waste measurement device which can quantify individual household food waste while minimizing user burden. The development of an open-vocabulary compost image recognition model will enable wide-scale household food waste measurement and provide personalized waste analytics without the need for labor-intensive manual annotation. To achieve this, we pursue the development of a robust open-vocabulary compost instance segmentation model. The smart compost bin is designed with this future automatic measurement routine in mind, and in the future, we aim to simply provide a software update to disable data annotation and convert the smart compost bin from a convenient waste image annotation framework to an automatic food waste measurement appliance.

Chapter 6: Conclusions

We present a 3D-printed hardware device for the automatic measurement of household food waste, addressing a critical gap in existing food waste and loss (FLW) measurement literature. Our food waste measurement protocol enables the collection of detailed compost image annotations, contributing towards the development of a computer vision model for automatic food waste classification. We evaluate the feasibility of pretrained contrastive image-language models for fine-grained open-vocabulary food and food waste image recognition tasks. We find that the use of contrastive image embeddings and an efficient vector database for image similarity search provides a robust food image classification system which contends with current state-of-the-art food image recognition models. The transfer of this food image recognition model to food object classification tasks results in a substantial performance loss, indicating the need for modified CLIP pretraining objectives for spatial reasoning tasks to effectively measure food or food waste from visual data.

We use a combination of YOLOv11 and `jina-clip-v2` for a two-stage food object detection model. Though this method yields reduced performance on in-domain tasks, it enables open-vocabulary object detection on out-of-domain food waste images. We do not find that the use of YOLOv11 and `jina-clip-v2` together yields a strong food or food waste object detection model, demonstrating a need for deeper evaluation of spatially aware vision-language models and alternative contrastive image-language pretraining objectives for compost image recognition tasks. We conclude that CLIP embeddings encode semantic information useful for fine-grained food image recognition tasks, and such models show promise for open-vocabulary compost instance classification tasks. However, the inability of food object recognition models to transfer to our food waste image recognition task highlights the need for a large dataset of annotated compost images to train a food waste object detection model. Using the data collection protocol enabled by our smart compost bin, we aim to provide such a dataset in our ongoing food waste image annotation study.

Bibliography

[1] *A National Strategy to Reduce Food Waste at the Consumer Level*. National Academies Press, Washington, D.C., October 2020.

[2] Justin Adams, Samson Kanai, Nedaa Abdulaziz Al-Mubarak, Tove Stuhr Sjoblom, Martin Stuchtey, Jessica Long, Philipp Buddemeier, Celine Herweijer, Louise Scott, Daniel Dowling, Jonathan Grant, Jim Stephenson, Laura Plant, Raimund Bleischwitz, Geoff Clarke, Pavel Kabat, John Elkington, Peter Head, Ecosequestration Trust, Guido Schmidt-Traub, Murray Birt, and Gert Wrigge. Valuing the sdg prize in food and agriculture. October 2016.

[3] United States Environmental Protection Agency. 2019 wasted food report. April 2023.

[4] Dario Allegra, Sebastiano Battiato, Alessandro Ortis, Salvatore Urso, and Riccardo Polosa. A review on food recognition technology for health applications. *Health Psychology Research*, 8(3), December 2020.

[5] Vera Amicarelli and Christian Bux. Food waste measurement toward a fair, healthy and environmental-friendly food system: a critical review. *British Food Journal*, 123(8):2907–2935, 2021.

[6] Jayanath Ananda, Gamithri Gayana Karunasena, and David Pearson. Identifying interventions to reduce household food waste based on food categories. *Food Policy*, 111:102324, 2022.

[7] Rahmi Arda Aral, Şeref Recep Keskin, Mahmut Kaya, and Murat Hacıömeroğlu. Classification of trashnet dataset based on deep learning models. In *2018 IEEE International Conference on Big Data (Big Data)*, page 2058–2062, December 2018.

[8] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024.

[9] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, page 446–461, Cham, 2014. Springer International Publishing.

[10] Mark Boulet, Annika Stott, and Sarah Kneebone. Which behaviours matter? prioritising food waste reduction behaviours for targeted policy and program approaches. *Journal of Environmental Management*, 345:118668, 2023.

[11] Jean C Buzby, Hodan Farah-Wells, and Jeffrey Hyman. The estimated amount, value, and calories of postharvest food losses at the retail and consumer levels in the United States. *USDA-ERS Economic Information Bulletin*, 121:1–39, 2014.

[12] Gustav Chrintz-Gath, Meena Daivadanam, Laran Matta, and Steve McKeever. Image-Based Dietary Assessment Using the Swedish Plate Model: Evaluation of Deep Learning-Based You Only Look Once (YOLO) Models. 9:e70124.

[13] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. Food Recognition: A New Dataset, Experiments, and Results. 21(3):588–598.

[14] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. Food recognition and leftover estimation for daily diet monitoring. In Vittorio Murino, Enrico Puppo, Diego Sona, Marco Cristani, and Carlo Sansone, editors, *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, pages 334–341, Cham, Switzerland, 2015. Springer International Publishing.

[15] Zach Conrad. Daily cost of consumer food wasted, inedible, and consumed in the united states, 2001–2016. *Nutrition Journal*, 19(1):35, April 2020.

[16] Avi Cooper, Keizo Kato, Chia-Hsien Shih, Hiroaki Yamane, Kasper Vinken, Kentaro Takemoto, Taro Sunagawa, Hao-Wei Yeh, Jin Yamanaka, Ian Mason, and Xavier Boix. Rethinking vlms and llms for image classification, 2024.

[17] M. Cropley, M. Sprajcer, and D. Dawson. Wastogram: Validation of a new tool to measure household food waste. *Journal of Environmental Psychology*, 84:101896, December 2022.

[18] Lisa Dahlén and Anders Lagerkvist. Methods for household waste composition studies. *Waste Management*, 28(7):1100–1112, 2008. Citation Key: DAHLEN20081100.

[19] Kalliopi V Dalakleidi, Marina Papadelli, Ioannis Kapolos, and Konstantinos Papadimitriou. Applying Image-Based Food-Recognition Systems on Dietary Assessment: A Systematic Review. 13(6):2590–2619.

[20] Kaitlin Danible, Chloe Panizza, Carol Boushey, Deborah Kerr, Fengqing Zhu, and Jinan C Banna. A novel to method to measure food waste: The mobile food record. *The Journal of Extension*, 59(3):1, 2021.

[21] Thais de Oliveira Pontes, Aldara da Silva César, Marco Antonio Conejero, Lucas Rodrigues Deliberador, and Mário Otávio Batalha. Food waste measurement in a chain of industrial restaurants in brazil. *Journal of Cleaner Production*, 369:133351, October 2022.

[22] Jan den Boer, Agata Skiba, Emilia den Boer, Gudrun Obersteiner, and Arkadiusz Dyjakon. Prevention and under-reporting effects of food waste diaries. *Sustainability*, 16(2411009), 2024. Citation Key: su162411009.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[24] Ayushi Dhelia, Srishti Chordia, and Kanisha B. Yolo-based food damage detection: An automated approach for quality control in food industry. In *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pages 1444–1449, 2024.

[25] Zhengxia Dou and John D. Toth. Global primary data on consumer food waste: Rate and characteristics – A review. *Resources, Conservation and Recycling*, 168:105332, 2021.

[26] United Nations General Assembly Economic and Social Council. Progress towards the sustainable development goals. May 2024.

[27] United Nations General Assembly Economic and Social Council. *Progress towards the Sustainable Development Goals.* May 2024.

[28] Takumi Ege and Keiji Yanai. Estimating food calories for multiple-dish food photos. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, page 646–651. IEEE, 2017.

[29] Ersin Elbasi, Nour Mostafa, Zakwan AlArnaout, Aymen I. Zreikat, Elda Cina, Greeshma Varghese, Ahmed Shdefat, Ahmet E. Topcu, Wiem Abdelbaki, Shinu Mathew, and Chamseddine Zaki. Artificial intelligence technology in the agricultural sector: A systematic literature review. *IEEE Access*, 11:171–202, 2023.

[30] Efrat Elimelech, Eyal Ert, and Ofira Ayalon. Bridging the gap between self-assessments and measured household food waste: A hybrid valuation approach. *Waste Management*, 95:259–270, July 2019.

[31] Bingbing Fang, Jiacheng Yu, Zhonghao Chen, Ahmed I. Osman, Mohamed Farghali, Ikko Ihara, Essam H. Hamza, David W. Rooney, and Pow-Seng Yap. Artificial intelligence for waste management in smart cities: A review. *Environmental Chemistry Letters*, 21(4):1959–1989, may 2023.

[32] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, September 2024.

[33] João Ferreira, Paulino Cerqueira, and Jorge Ribeiro. Food waste detection in canteen plates using yolov11. *Applied Sciences*, 15(13), 2025.

[34] Food and Agriculture Organization (FAO). Food wastage footprint: Impacts on natural resources. *Summary Report*, pages 1–63, 2013.

[35] Food and Agriculture Organization (FAO) of the United Nations. Tracking progress on food and agriculture-related sdg indicators 2021: A report on the indicators under fao custodianship, 2021.

[36] Hamish Forbes, Eloise Peacock, Nettie Abbot, and Michael Jones. Think eat save: Tracking progress to halve global food waste. *Food Waste Index Report*, 2024.

[37] Eva Ganglbauer, Geraldine Fitzpatrick, and Florian Güldenpfennig. Why and what did we throw out? probing on reflection through the food waste diary. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 1105–1114, New York, NY, USA, 2015. Association for Computing Machinery.

[38] Claudia Giordano, Fabrizio Alboni, and Luca Falasconi. Quantities, determinants, and awareness of households' food waste in italy: A comparison between diary and questionnaires quantities. *Sustainability*, 11(1212):3381, January 2019.

[39] Yanne Goossens, Dominik Leverenz, and Manuela Kuntscher. Waste-tracking tools: A business case for more sustainable and resource efficient food services. *Resources, Conservation Recycling Advances*, 15:200112, November 2022.

[40] J. Gustafsson, C. Cederberg, U. Sonesson, and A. Emanuelsson. The methodology of the fao study: Global food losses and food waste - extent, causes and prevention"- fao, 2011. 2013.

[41] Hamid Hassannejad, Guido Matrella, Paolo Ciampolini, Ilaria De Munari, Monica Mordonini, and Stefano Cagnoni. A new approach to image-based estimation of food volume. *Algorithms*, 10(2):66, 2017.

[42] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification, 2017.

[43] Jun-Ho Huh, Jae-Hyeon Choi, and Kyungryong Seo. Smart trash bin model design and future for smart city. *Applied Sciences*, 11(1111):4810, January 2021.

[44] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023.

[45] Aknur Karabay, Huseyin Atakan Varol, and Mei Yen Chan. Improved food image recognition by leveraging deep learning and data-driven methods with an application to Central Asian Food Scene. 15(1):14043.

[46] Parneet Kaur, Karan Sikka, Weijun Wang, Serge Belongie, and Ajay Divakaran. Foodx-251: A dataset for fine-grained food classification. (arXiv:1907.06167), July 2019. arXiv:1907.06167 [cs].

[47] Shannon Kenny, Jenny Stephenson, Alexandra Stern, Juliana Beecher, Ben Morelli, Andrew Henderson, Emelyn Chiang, Andrew Beck, Sarah Cashman, Emily Wexler, Kyle McGaughy, and Abby Martell. From field to bin: The environmental impacts of u.s. food waste management pathways. October 2023.

[48] Maria Koskinopoulou, Fredy Raptopoulos, George Papadopoulos, Nikitas Mavrakis, and Michail Maniadakis. Robotic waste sorting technology: Toward a vision-based categorization system for the industrial robotic separation of recyclable waste. *IEEE Robotics & Automation Magazine*, 28(2):50–60, 2021.

[49] Max Krause, Shannon Kenny, Jenny Stephenson, and Amanda Singleton. Quantifying methane emissions from landfilled food waste. October 2023.

[50] M.S. Kummu, H. de Moel, M. Porkka, S. Siebert, O. Varis, and P.J. Ward. Lost food, wasted resources: global food supply chain losses and their impacts on freshwater, cropland, and fertiliser use. *Science of the Total Environment*, 438:477–489, 2012.

[51] S. Lebersorger and F. Schneider. Discussion on the methodology for determining food waste in household waste composition studies. *Waste Management (New York, N.Y.)*, 31(9–10):1924–1933, 2011.

[52] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[53] Detianjun Liu, Enguang Zuo, Dingding Wang, Liang He, Liujing Dong, and Xinyao Lu. Deep learning in food image recognition: A comprehensive review. *Applied Sciences*, 15(14):7626, 2025.

[54] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[55] Frank Po Wen Lo, Yingnan Sun, Jianing Qiu, and Benny Lo. Image-Based Food Classification and Volume Estimation for Dietary Assessment: A Review. 24(7):1926–1939.

[56] Edoardo Longo, Fatih Alperen Sahin, Alessandro E. C. Redondi, Patrizia Bolzan, Massimo Bianchini, and Stefano Maffei. Take the trash out... to the edge. creating a smart waste bin based on 5g multi-access edge computing. In *Proceedings of the Conference on Information Technology for Social Good*, GoodIT '21, page 55–60, New York, NY, USA, September 2021. Association for Computing Machinery.

[57] Jelena Lubura and Lato Pezo. Food recognition and food waste estimation using convolutional neural network. *Electronics*, 11(22):3746, 2022.

[58] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation, 2023.

[59] Yuhan Luo, Young-Ho Kim, Bongshin Lee, Naeemul Hassan, and Eun Kyoung Choe. Foodscrap: Promoting rich data capture and reflective food journaling through speech input. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, DIS '21, page 606–618, New York, NY, USA, 2021. Association for Computing Machinery.

[60] Jayson L. Lusk and Brenna Ellison. Economics of household food waste. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, 68(4):379–386, 2020.

[61] Timo Lüddecke and Alexander S. Ecker. Image segmentation using text and image prompts, 2022.

[62] Anish Madan, Neehar Peri, Shu Kong, and Deva Ramanan. Revisiting few-shot object detection with vision-language models. *Advances in Neural Information Processing Systems*, 37:19547–19560, 2024.

[63] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):187–203, January 2021.

[64] Corby K. Martin, Hongmei Han, Sandra M. Coulon, H. Raymond Allen, Catherine M. Champagne, and Stephen D. Anton. A novel method to remotely measure food intake of free-living people in real-time. *The British journal of nutrition*, 101(3):446–456, March 2009.

[65] Hrair Mekhsian, Kevin Duffy, and Marc Zornes. System and method for monitoring food waste, December 2023.

[66] Austin Meyers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin P Murphy. Im2calories: towards an automated mobile vision food diary. In *Proceedings of the IEEE international conference on computer vision*, pages 1233–1241, 2015.

[67] George A. Miller. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[68] Weiqing Min, Linhu Liu, Zhiling Wang, Zhengdong Luo, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. (arXiv:2008.05655), August 2020. arXiv:2008.05655 [cs].

[69] Zhao-Yan Ming, Jingjing Chen, Yu Cao, Ciarán Forde, Chong-Wah Ngo, and Tat Seng Chua. Food photo recognition for dietary tracking: System and experiment. In *International Conference on Multimedia Modeling*, pages 129–141. Springer, 2018.

[70] Austin Myers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin Murphy. Im2calories: Towards an automated mobile vision food diary. In *2015 IEEE International Conference on Computer Vision (ICCV)*, page 1233–1241, Santiago, Chile, December 2015. IEEE.

[71] P Nagasharmila, G K Aadhini, V Mohanasubi, T Pramika, and R Babitha Lincy. A Holistic Approach to Food Waste Management: Leveraging Full-Stack Technologies for Efficient Resource Utilization. In *2023 Third International Conference on Smart Technologies, Communication and Robotics (STCR)*, volume 1, pages 1–5.

[72] Kaimu Okamoto and Keiji Yanai. Uec-foodpix complete: A large-scale food image segmentation dataset. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part V*, pages 647–659. Springer, 2021.

[73] Effie Papargyropoulou, Rodrigo Lozano, Julia K. Steinberger, Nigel Wright, and Zaini bin Ujang. The food waste hierarchy as a framework for the management of food surplus and food waste. *Journal of cleaner production*, 76:106–115, 2014.

[74] Diego Inácio Patrício and Rafael Rieder. Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. *Computers and electronics in agriculture*, 153:69–81, 2018.

[75] Ludovica Principato, Giovanni Mattia, Alessio Di Leo, and Carlo Alberto Pratesi. The household wasteful behaviour framework: A systematic review of consumer food waste. *Industrial Marketing Management*, 93:641–649, 2021.

[76] Tom E Quested, Griff Palmer, Laura C Moreno, Christa McDermott, and Kelsea Schumacher. Comparing diaries and waste compositional analysis for measuring food waste in the home. *Journal of Cleaner Production*, 262:121263, 2020. Publisher: Elsevier.

[77] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[78] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 23–29 Jul 2023.

[79] Gabriel Jäger Ramos, João Augusto Rossi Borges, Carla Heloisa de Faria Domingues, and Erica van Herpen. Reducing food waste by simply measuring it: insights from interventions to reduce household food waste. *British Food Journal*, 126(2):812–833, November 2023.

[80] Lucia A. Reisch, Cass R. Sunstein, Mark A. Andor, Friederike C. Doebbe, Johanna Meier, and Neal R. Haddaway. Mitigating climate change via food consumption and food waste: A systematic map of behavioral interventions. *Journal of Cleaner Production*, 279:123717, January 2021.

[81] Beate Richter and Wolfgang Bokelmann. Explorative study about the analysis of storing, purchasing and wasting food by using household diaries. *Resources, Conservation and Recycling*, 125:181–187, 2017.

[82] Steve Rock. Food waste reduction in military kitchens, a tracking technology demonstration at fort jackson. September 2019.

[83] Brian E. Roe, John W. Apolzan, Danyi Qi, H. Raymond Allen, and Corby K. Martin. Plate waste of adults in the united states measured in free-living conditions. *PLOS ONE*, 13(2):e0191813, February 2018.

[84] Brian E Roe, Danyi Qi, Robbie A Beyl, Karissa E Neubig, Corby K Martin, and John W Apolzan. The validity, time burden, and user satisfaction of the foodimagetm smartphone app for food waste measurement versus diaries: a randomized crossover trial. *Resources, Conservation and Recycling*, 160:104858, 2020.

[85] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut — interactive foreground extraction using iterated graph cuts.

[86] Alberto Rovetta, Fan Xiumin, Federico Vicentini, Zhu Minghua, Alessandro Giusti, and He Qichang. Early detection and evaluation of waste through sensorized containers for a collection monitoring application. *Waste Management*, 29(12):2939–2949, December 2009.

[87] Martine Rutten, Peter Nowicki, Marc-Jeroen Bogaardt, and Lusine Aramyan. *Reducing food waste by households and in retail in the EU.*

[88] Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17542–17552, June 2024.

[89] Doyen Sahoo, Wang Hao, Shu Ke, Wu Xiongwei, Hung Le, Palakorn Achananuparp, Ee-Peng Lim, and Steven C. H. Hoi. FoodAI: Food Image Recognition via Deep Learning for Smart Food Logging. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2260–2268.

[90] Christie N Scollon, Chu Kim-Prieto, and Ed Diener. Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness studies*, 4(1):5–34, 2003.

[91] Laura Sheble and Barabara Wildemuth. Research diaries. *Applications of social research methods to questions in information and library science*, pages 211–221, 2009.

[92] Yiheng Shu, Brian E Roe, and Kathryn Bender. Adapting, refining and expanding a validated questionnaire to measure food waste in us households. *MethodsX*, 8:101377, 2021.

[93] Kirsi Silvennoinen, Sampsa Nisonen, and Oona Pietilainen. Food waste case study and monitoring developing in finnish food services. *Waste Management*, 97:97–104, 2019. Publisher: Elsevier.

[94] Florian Taurer, Florian Wolling, Josh Moore, and Florian Michahelles. Smart trash can: Easy collection of photos of organic waste in the home. In *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia*, page 571–573, Vienna Austria, December 2023. ACM.

[95] Anastasia Tsita, Stylianos Eleftheriadis, Thanasis Vafeiadis, Dimosthenis Ioannidis, and Dimitrios Tzovaras. Ai-assisted gamification platform for smart city food waste management. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25 Companion, page 12–16, New York, NY, USA, March 2025. Association for Computing Machinery.

[96] Monika van Den Bos Verma, Linda de Vreede, Thom Achterbosch, and Martine M Rutten. Consumers discard a lot more food than widely believed: Estimates of global food waste using an energy gap approach and affluence elasticity of food waste. *PloS one*, 15(2), 2020.

[97] Erica van Herpen, Ivo A van der Lans, Nancy Holthuysen, Mariska Nijenhuis-de Vries, and Tom E Quested. Comparing wasted apples and oranges: An assessment of methods to measure household food waste. *Waste Management*, 88:71–84, 2019. Publisher: Elsevier.

[98] Matteo Vittuari, Laura Garcia Herrero, Matteo Masotti, Elisa Iori, Carla Caldeira, Zhuang Qian, Hendrik Bruns, Erica van Herpen, Gudrun Obersteiner, Gulbanu Kaptan, et al. How to reduce consumer food waste at household level: A literature review on drivers and levers for behavioural change. *Sustainable Production and Consumption*, 2023.

[99] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding, 2024.

[100] Sohani Vihanga Withanage, Goretty Maria Dias, and Komal Habib. Review of household food waste quantification methods: Focus on composition analysis. *Journal of Cleaner Production*, 279:123722, 2021.

[101] Florian Wolling, Gabor Pal, Lukas Larcher, and Florian Michahelles. Smart trash can: Image classification and segmentation for waste quality management in the home. In *Proceedings of the 14th International Conference on the Internet of Things*, IoT '24, page 65–71, New York, NY, USA, March 2025. Association for Computing Machinery.

[102] Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven CH Hoi, and Qianru Sun. A large-scale benchmark for food image segmentation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 506–515, 2021.

[103] Zhiyong Xiao, Guang Diao, and Zhaohong Deng. Fine grained food image recognition based on swin transformer. 380:112134.

[104] Li Xue, Gang Liu, Julian Parfitt, Xiaojie Liu, Erica Van Herpen, Åsa Stenmarck, Clementine O'Connor, Karin Östergren, and Shengkui Cheng. Missing food, missing data? A critical review of global food losses and food waste data. *Environmental Science & Technology*, 51(12):6618–6633, June 2017. Publisher: American Chemical Society.

[105] Keiji Yanai and Yoshiyuki Kawano. Food image recognition using deep convolutional network with pre-training and fine-tuning. In *2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–6, 2015.

[106] Yue Yu, Sophia Yi, Xi Nan, Leo Yu-Ho Lo, Kento Shigyo, Liwenhan Xie, Jeffry Wicaksana, Kwang-Ting Cheng, and Huamin Qu. Foodwise: Food waste reduction and behavior change on campus with data visualization and gamification. In *Proceedings of the 6th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, COMPASS '23, page 76–83, New York, NY, USA, August 2023. Association for Computing Machinery.

[107] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023.

[108] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification?, 2024.