applied sciences

MDPI

Article

FLIC: A Real-World Dataset for Visual Estimation of Food Leftovers in Canteens

Flavio Piccoli [1], Damiano Callegaro [2], Davide Marelli [1], Marco Buzzelli [1,*], Cinzia Franchini [2], Lorenzo Stella [2], Simone Bianco [1], Gianluigi Ciocca [1], Raimondo Schettini [1] and Francesca Scazzina [2,3]

[1] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, 20126 Milan, Italy
[2] Human Nutrition Unit, Department of Food and Drug, University of Parma, Via Volturno 39, 43125 Parma, Italy
[3] Giocampus Scientific Committee, Parco Area delle Scienze 38, 43124 Parma, Italy
* Correspondence: marco.buzzelli@unimib.it

Abstract

We present FLIC, a real-world annotated dataset designed for the visual estimation of food leftovers in canteens and other collective catering environments using standard 2D RGB imagery. Collected over 22 days in an operational university canteen, the dataset includes 401 paired image acquisitions of full and leftover trays, each associated with pixel-precise semantic segmentation masks and physically measured food mass. The goal is to support research on the estimation of leftover food mass from tray images, a task that has received limited attention compared to pre-consumption food recognition, despite its relevance for sustainability and operational decision making in food services. Unlike existing food datasets, FLIC jointly provides paired before–after visual observations and reliable mass ground truth, enabling quantitative analysis of food leftovers under realistic conditions without relying on depth or multi-view information. To demonstrate the dataset's applicability, we rely on the concept of digital density, relating pixel area to food mass, and implement a lightweight, interpretable baseline mass estimation pipeline. This includes an automatic food/no-food segmentation stage, evaluated across multiple deep learning models (U-Net, DABNet, DINOv2+FeatUp, and SAM), followed by an assisted food recognition stage that leverages the fixed daily menu to map broad user input (e.g., "first course" vs. "second course") to a specific food class. Experimental results highlight both the potential and the intrinsic challenges of visual food leftover estimation.

Keywords: food leftovers; food waste; computer vision; semantic segmentation; mass estimation

Academic Editor: Cosimo Nardi

Received: 3 April 2026
Revised: 22 May 2026
Accepted: 27 May 2026
Published: 31 May 2026

1. Introduction

In recent years, the increasing focus on health, nutrition, and sustainability has led to the development of numerous systems and applications leveraging computer vision and artificial intelligence for food recognition and analysis [1,2]. These technologies are widely used for identifying food items and estimating key properties such as ingredients, portion size, volume, and mass, with applications ranging from dietary tracking to automated meal payment systems [3,4].

Most of these applications rely on machine learning models that require large annotated datasets for training. These datasets are typically created through manual labeling by human experts, a time-consuming and resource-intensive process [5,6].

Furthermore, existing food recognition systems are primarily designed for individual users, with image acquisition typically performed via mobile devices. In such settings, users capture images of their dishes before consumption, and the system processes these images to provide nutritional information or facilitate meal tracking [7,8]. However, this approach poses several challenges, including inconsistencies in image quality, variability in user behavior, and the frequent need for manual intervention. In contrast, collective catering environments, such as cafeterias and canteens, present a different set of requirements and opportunities. Fixed-camera systems, where cameras are installed in predefined locations relative to the food trays or even integrated with the checkout system, offer a standardized setup that facilitates automated recognition and analysis. Unlike mobile-based approaches, which suffer from significant variability in intrinsic and extrinsic parameters (e.g., camera calibration, distance, angle, and lighting), fixed setups maintain consistent acquisition geometry and illumination. This enables a much more accurate and repeatable estimation of food quantities. In this work, we deliberately focus on standard 2D RGB image acquisition, as it represents the most scalable and cost-effective sensing modality for large-scale deployment in collective catering environments.

Beyond food recognition, an important goal in collective catering is the estimation of plate waste, a task that must begin with the broader identification of food leftovers. Andrews et al. [9] define food waste (or plate waste) as any uneaten portion of food on a person's plate meal that was intended for consumption but ultimately went uneaten. Aloysius et al. [10] collected several contrasting definitions of what constitutes food leftovers, which range from "prepared but not plated food" to "takeaway leftover food". Here, we define food leftovers based on the diner's leftover interpretation, i.e., the total set of served but uneaten items. This, compared to plate waste, additionally encompasses non-edible fractions (such as bones, peels, and shells) and unrecoverable edible residues like scattered grains or residual sauces. In this sense, the identification of leftovers serves as a foundational step for the broader objective of measuring effective food waste. Initiatives aimed at identifying and quantifying leftovers can provide valuable insights into consumption patterns, helping diners to be more conscious consumers and food caterers to optimize meal offers. At a larger scale, tracking food leftovers at a collective catering level can contribute by highlighting trends, reducing overproduction, and minimizing environmental impact, with positive impacts on natural resources, economic growth, and society. This is particularly relevant when considering the current global food waste scenario. According to the most recent data [11], approximately 1.05 billion tons of food are wasted globally each year. At the European Union (EU) level, over 59 million tons of food are wasted annually, equal to 20% of the EU region's total food production. Of this, 11% is produced in food services such as restaurants and catering operations. Considering only the Italian scenario, food waste recorded for collective food services was 6% [12]. Given the recent increase in out-of-home meal consumption [13] and the expected growth of the European foodservice market by 2030 [14], the integration of digital tools in food service operations is becoming a sustainable practice for reducing food waste [15], and for reducing its environmental, social and economic impact [11,16] .

Despite growing research interest, there is still a lack of reliable, scalable tools for quantifying food mass in collective dining settings. Manual methodologies are labor-intensive and subject to observer variability, while many automated systems focus on pre-consumption food recognition, neglecting leftover analysis [17].

The main contributions of this work are summarized as follows:

- We introduce FLIC, a real-world dataset for the estimation of food leftovers in canteens and other collective catering environments, featuring paired full and leftover tray images, pixel-precise semantic annotations, and physically measured food mass

ground truth. This explicitly addresses the shortcomings of existing datasets, which only cover part of the aforementioned attributes.

- A reproducible acquisition and annotation protocol based on standard 2D RGB imaging and fixed-camera geometry is reported, designed to enable quantitative and scalable analysis of food leftovers under realistic operational conditions.
- We define a lightweight and interpretable baseline for leftover mass estimation based on the concept of digital density, together with systematic experiments that expose dataset characteristics, sources of variability, and intrinsic challenges of the problem.

In Section 2, we provide a detailed overview of existing public datasets related to the analysis of food leftovers and food estimation in order to better contextualize the value of our proposed FLIC dataset. In Section 3, we describe the context for acquisition of the dataset, including the canteen settings, our custom acquisition device, the physical mass acquisition protocol, and the annotation of semantic information. In Section 4, we define our methodology for mass estimation based on the concept of digital density. We also describe how assisted food recognition is used to perform automatic food/no-food segmentation and semi-automatic food classification. In Section 5, we extract an analysis of the collected data, and we present results on food/no-food segmentation and mass estimation.

2. Existing Datasets Related to Leftover and Mass Estimation

We organize existing datasets into two groups: (i) datasets explicitly targeting leftover food and food waste estimation and (ii) datasets providing ground-truth mass (and, in some cases, volume) annotations for food items or complete dishes. Although some level of cross-talk does exist, with certain leftover-oriented datasets also providing food mass data, this categorization highlights the existing structural gap in publicly available resources: datasets focusing on leftovers typically lack reliable physical quantity measurements, while datasets providing mass annotations rarely capture real post-consumption scenarios. A sample of images from the existing datasets is shown in Figure 1. Dataset information is summarized in Table 1, including information on our own FLIC, which is presented in detail in Section 3.

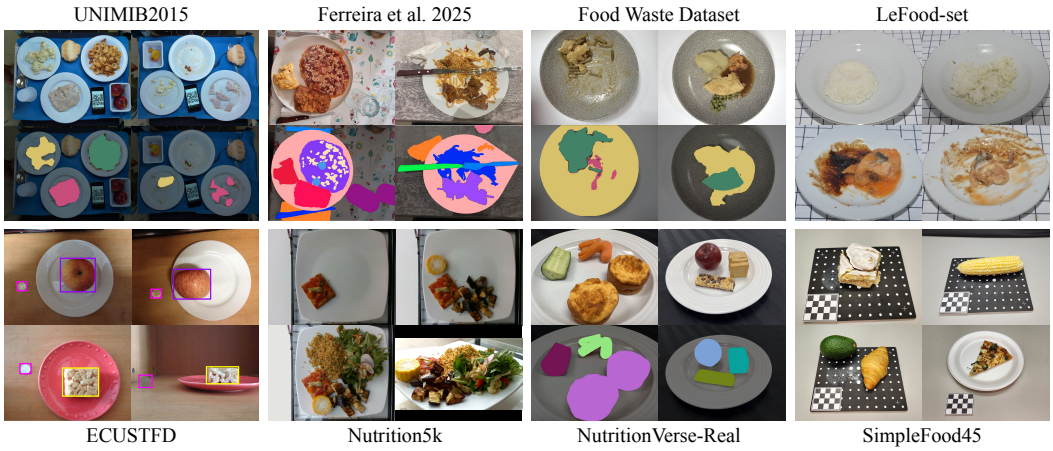

Figure 1. Sample images from existing datasets related to food leftover estimation (**top**) and mass estimation (**bottom**). Datasets focusing on leftovers typically lack reliable physical quantity measurements, while datasets providing mass annotations rarely capture real post-consumption scenarios. Spatial annotations are shown only where available, each dataset with their own color coding.

Table 1. Schematic comparison of datasets related to leftover estimation (first group) and quantitative food analysis (second group), including our proposed FLIC dataset (last row). 'C' identifies the number of semantic classes, if present.

Dataset (Year)	Setting/Location	Scene/Layout	Presentation	Full/Leftover	# Images/Pairs	Spatial Annotations	Quantity Annotations
UNIMIB2015 [5]	University canteen (Italy)	Staged presentation; low realism for leftovers	Trays; multi-item	Both (paired)	1080 images (700 full, 380 leftover)	Polygonal semantic segmentation masks (C = 15)	None
Ferreira et al. 2025 [18]	School canteen (Portugal)	Natural canteen views; varied meal progression	Plates; multi-item	Both (unpaired)	175 images	Instance segmentation masks (C = 3)	None
Food Waste Dataset [19]	Institutional food service (Germany)	Natural cafeteria presentation	Plates; multi-item	Leftover only	375 images	Automatic ingredient segmentation masks (C = 41)	Plate mass before/after (g); energy (kcal); macronutrients
LeFood-set [20]	Hospital cafeteria (Indonesia)	Controlled checkerboard background	Plates; single-item	Both (paired)	524 pairs	None (C = 34)	Plate mass before/after (g); leftover score
ECUSTFD [21]	Controlled indoor setup (China)	Structured views (top + side)	Plates; single-item	Full only	2978 images	Bounding boxes (C = 19)	Item mass (g); item volume (mL); item energy (kcal)
Nutrition5k [22]	Google cafeterias (USA)	Real cafeteria meals; overhead RGB-D rig	Plates; multi-item	Full only	5000 images; 20 k videos; 3500 RGB-D images	None	Ingredient masses (g); macronutrients; dish energy (kcal)
NutritionVerse-Real [23]	Indoor natural scenes (Canada)	Varied smartphone viewpoints; black mat	Plates; multi-item	Full only	889 images	Pixel-precise semantic segmentation masks (C = 45)	Ingredient masses (g); dish nutrition
SimpleFood45 [24]	Controlled environment	Checkerboard reference pattern	Single item	Full only	513 images	None	Item volume (mL); item mass (g); item energy (kcal)
FLIC (this work)	Public canteen (children's summer camp, Italy)	Natural messy canteen presentation; top-down capture	Trays; multi-item	Both (paired)	401 pairs (802 images)	Pixel-precise semantic segmentation masks (C = 36)	Plate mass before/after (g)

2.1. Datasets for Food Waste and Leftover Estimation

2.1.1. UNIMIB2015 (Ciocca et al. 2015 [5])

UNIMIB2015 is one of the earliest datasets explicitly designed for leftover estimation in a canteen setting. It contains images acquired in a university canteen, with different dishes placed on a tray and photographed before and after consumption. The dataset comprises 1080 total images (700 images of full trays and a subset of 380 corresponding leftover trays), with 15 food classes and polygonal annotations of each food region. The arrangement of food on the trays is staged and quite tidy, which reduces realism, especially in leftover images. In addition, some images were obtained by rearranging the same plates seen in other images, changing only their orientation and layout on the tray. A later extension, UNIMIB2016 [6], increased the number of food categories and relaxed the acquisition conditions but only includes full trays and therefore cannot be used for leftover estimation.

2.1.2. Ferreira et al. 2025 [18]

A more recent dataset focuses on food waste detection in a school canteen environment. It contains 175 real images and an average of 7.2 annotated objects per image, including both food and non-food items. Annotations are provided as pixel-precise semantic masks. Unlike UNIMIB2015, the images are not organized as explicit full-leftover pairs: the collection contains a mixture of plates at different stages of a meal (before, during, and after consumption), typically as close-ups of one or two plates rather than complete trays. This makes the dataset useful for semantic understanding of food waste and for the training of segmentation-based methods, but it is less suited to paired before/after analyses of the same meal.

2.1.3. Food Waste Dataset (AI Service Center, 2024 [19])

The Food Waste Dataset released by the AI Service Center (Hasso Plattner Institute) provides images of plates with leftover food, together with rich tabular metadata. The dataset currently includes 375 images (215 train, 160 test) of meals in institutional food-service settings, along with detailed ingredient-level nutritional information and per-plate measurements recorded both before and after consumption (mass, calories, macronutrients, and derived waste metrics such as returned quantity and return percentage). An enhanced third-party version [25] also includes segmentation masks automatically generated via

YOLO-E [26] and feature embeddings for computer-vision experiments. Some shortcomings of the automated annotation are shown in figure: in the first example, the green beans and chicken are assigned one unique label, whereas some residuals of the chicken's condiment are marked with a separate label. In the second example, only the chicken is assigned its own label, whereas four other food items are merged into a unique label. This dataset does not provide images of the full plate prior to consumption for each leftover image; instead, the before/after information is purely numeric, as previously described. Consequently, it is well suited for modeling the relation between the visual appearance of leftovers and quantitative waste but less suited for paired visual analysis of the same tray before and after the meal.

2.1.4. LeFood-Set (Sari et al. 2025 [20])

LeFood-set contains 524 image pairs, each depicting an individual Indonesian dish before and after consumption. The dataset spans 34 food categories and was collected at a hospital in Malang, Indonesia, using a Nikon D3200 DSLR camera . Each sample focuses on a single food item, which may include both solid and liquid components: solid dishes are served on plates, while liquids are provided in bowls. All items are presented on white dishware placed over a wooden placemat featuring a black-and-white checkerboard pattern that serves as a geometric reference for scale. For every image pair, the dataset provides physical mass measurements obtained by weighing the plate or bowl with a digital scale before and after the meal.

Several other works describe datasets specifically acquired for food waste or leftover analysis, but to the best of our knowledge, the corresponding data is not publicly downloadable. These include datasets reported by Sari et al. 2021 [27] (about 340 images and 5 classes), Hsieh et al. 2022 [28] (803 images and 16 classes), Rokhva and Teimourpour 2025 [29] (1413 images and 5 classes), and Li et al. 2025 [30] (40 images and no explicit class information).

2.2. Datasets with Ground-Truth Mass Annotations

In addition to leftover-centric datasets, there is a growing family of datasets that provide ground-truth quantity information for single items or complete dishes—typically based on mass and sometimes with volume and full nutritional breakdown. These datasets are highly relevant to quantitative food intake estimation, even though they usually do not depict leftovers.

2.2.1. ECUSTFD (Liang and Li, 2017 [21])

The ECUST Food Dataset (ECUSTFD) consists of images of single food items placed on a plate. For each portion, several pairs of images are captured using smartphones, with each pair containing a top view and a side view of the same food. The dataset includes 19 food classes and a total of 2978 images. Annotations include bounding boxes around the food, the mass (measured with an electronic scale), the volume (measured via volume displacement using the drainage method), and the corresponding energy (computed from nutritional tables). ECUSTFD provides a clean, controlled setting with high-quality mass and volume labels but focuses exclusively on isolated items rather than realistic multi-item meals or leftovers.

2.2.2. Nutrition5k (Thames et al., 2021 [22])

Nutrition5k is a large-scale dataset of approximately 5000 real-world dishes collected in Google cafeterias using a custom scanning rig. For each dish, the dataset provides multiple side-view videos, overhead RGB-D images (when available), a fine-grained list of ingredients, per-ingredient mass, and a complete nutritional breakdown (calories, fat,

protein, and carbohydrates). The dishes are assembled incrementally: ingredients are added one at a time, and images and metadata are recorded at each step. This design supports the training of models for dish-level mass and macronutrient estimation, but no segmentation masks or bounding boxes are provided, and the dataset does not include leftovers. The incremental assembly theoretically allows for the construction of synthetic "full vs. partial" states, but these represent artificial constructions rather than real consumption.

2.2.3. NutritionVerse-Real (Tai et al., 2023 [23])

NutritionVerse-Real is a manually collected dataset targeting dietary intake estimation. It contains 889 images of 251 distinct dishes and 45 unique food types acquired using an iPhone 13 Pro Max from multiple random viewpoints. For each dish, the mass of every ingredient was measured with a food scale, and nutritional values were obtained either from packaging or from the Canada Nutrient File [31]. These per-ingredient values were aggregated to obtain dish-level macronutrients and micronutrients, which is the only information publicly exposed. In addition, the dataset provides manually labeled segmentation masks for a subset of images, combining detailed mass/nutrition labels with dense spatial annotations. As with Nutrition5k, the images depict full dishes rather than leftovers.

2.2.4. SimpleFood45 (Vinod et al., 2024 [24])

SimpleFood45 is a portion-estimation benchmark of 2D images of real foods captured with a smartphone camera. The dataset contains images of 45 individual food items (grouped into a smaller number of food types) placed on a planar checkerboard pattern that serves as a physical scale reference. For each image, the dataset provides ground-truth volume (mL), mass (g), and energy (kCal), enabling evaluation of single-image volume and energy estimation approaches. SimpleFood45 explicitly targets realistic camera poses and the use of 3D models for portion estimation, but, again, it considers single food items and does not cover leftovers or full multi-item trays.

Konstantakopoulos et al. [32] introduced MedGRFood, a Mediterranean (mainly Greek) food image dataset used for both food recognition and mass estimation studies. To the best of our knowledge, the mass-annotated dataset is not currently available for public download.

Beyond the datasets described above, several works have proposed datasets designed primarily for volume estimation rather than mass. However, these are not publicly released, and we therefore do not consider them in detail here.

The identified limitations of existing datasets motivate the need for a dataset that jointly combines paired full-leftover observations, pixel-level semantic information, and physically measured mass ground truth acquired in a real operational setting, which is the focus of the next section.

3. FLIC Dataset Acquisition

Data collection took place in a canteen of the University of Parma, Italy, located on the scientific disciplines campus. Every year during the summer season, one part of the dining hall is devoted to the provision of lunch to participants of the summer camp of the broader Giocampus initiative [33], which promotes healthy and sustainable lifestyle among children aged 6–12 in Parma.

Meals were offered according to a biweekly rotating menu: each day's lunch consisted of a first course (typically carbohydrate-based dishes such as pasta or rice), a second course (protein-based preparations), and a vegetable side dish. First and second courses were predetermined in advance for the entire rotation, while side dishes changed daily to reflect

seasonality. Although every child received the same fixed menu, they had the option to request a reduced portion or to decline any dish altogether.

For dataset acquisition, we collected information on the meal before and after consumption in the form of its physical mass (Section 3.1), and pictures captured using an ad hoc device (Section 3.2). Figure 2 shows the interaction with the user during acquisition: First of all, the user autonomously composes his/her tray. A QR code specific for that user on that day is then added onto the tray for acquisition through our device, and the individual plates are weighed. After meal consumption, the same acquisition procedure with the same QR code is repeated. QR codes allow for association of the before- and after-consumption pictures while preserving user anonymity: because the study involved only anonymized photographs and plate-level mass measurements without collecting personal or identifiable data, formal ethical approval was not required.

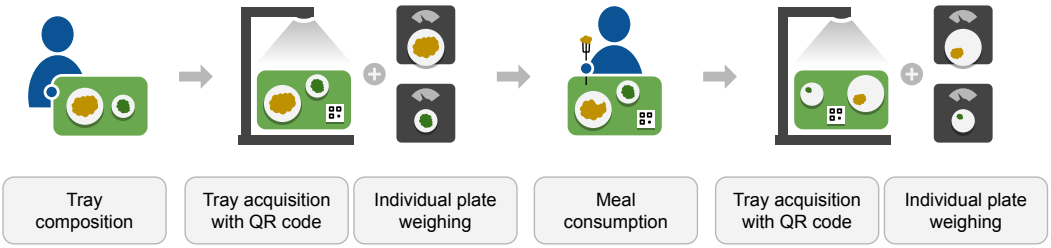

Figure 2. Interaction with the system during dataset acquisition. The user autonomously composes his/her tray. Then, before- and after-consumption information is captured in the form of pictures, and individual plate weighing.

It is worth noting that pictures were shot to frame the whole tray at once, whereas physical mass values were captured per plate: our goal was to design and implement a semi-automatic mass estimation procedure that operates on the whole tray, and the physical mass constitutes our ground-truth reference.

Although data were collected in a specific institutional setting, the acquisition protocol and tray-based organization reflect common operational practices in collective catering, making the dataset representative of a broad class of real-world scenarios.

3.1. Physical Mass Acquisition Protocol

All the food mass measurements were carried out by trained researchers using a digital scale calibrated to ±1 g accuracy. Immediately after plating, each compostable plate (tare 10 g or 15 g) carrying the first course, second course, and vegetable side dish was weighed to record the gross served mass. Bread was not weighed before consumption because it was prepackaged in a fixed portion of 50 g. Once diners finished their meals, the same plates were re-weighed under identical conditions.

Plate tares were systematically subtracted from each mass annotation by referring to the corresponding plate type.

3.2. Acquisition Device

We developed a custom device to ensure consistent acquisition conditions across all trays and to speed up the overall process.

As shown in Figure 3, the device comprises a bottom section where operators place the trays and an L-shaped vertical arm that supports the acquisition camera and lighting system, both of which are oriented downwards. We integrated a camera based on a Sony IMX378 sensor with a resolution of 4056 × 3040 pixels. Images were later subsampled for processing and sharing, as detailed in Section 3.4. We positioned the camera 65 cm above the tray to ensure that each tray fully fit within the field of view. We arranged high-power

natural white LEDs around the camera to provide uniform illumination over the tray surface and to reduce shadows caused by ambient light sources, which effectively reduces unknown variability due to uncontrolled illumination conditions. A 7-inch capacitive touchscreen placed in front of the operator provides an intuitive user interface that shows a real-time image preview and offers on-screen controls that allows for the triggering of data acquisition. The lower section of the device houses the electronics required for the camera, LEDs, and display. A Raspberry Pi 5 single-board computer with 8 GB of RAM controls all hardware components, including the LED driver circuitry, while an embedded power supply delivers a regulated 5 V to the system. We built the structure using sealed plastic components to facilitate cleaning and to reduce the risk of damage in the event of food or liquid spills. The proposed acquisition setup relies exclusively on low-cost, off-the-shelf components and is intended as a reference design rather than a specialized sensing solution.

Figure 3. The acquisition device as used within the data collection campaign for the FLIC dataset.

The acquisition software was written in Python 3.10.12 and executed directly on the device. The system stores data locally and simultaneously transmits it to a remote database (via Wi-Fi or Ethernet). This architecture ensures data redundancy and supports disaster recovery in case of device failure while still enabling continued operation when the network connection is unavailable or unstable.

3.3. Semantic Annotation

To support our experiments on mass estimation, we defined a ground truth for the semantic annotation of food items on each tray (both full and leftover). This ground truth consists of per-pixel labels that specify the food category associated with each region of the image.

Annotations were created using the Computer Vision Annotation Tool (CVAT) [34], specifically interacting through the Segment Anything Model (SAM) [35]: with minimal user input (often as little as a single click), we obtained a good delineation of a food region, which was then manually refined if necessary using a brush-like tool.

Beyond binary segmentation, we assigned a semantic label to each food region based on one of the following predefined classes: “first course”, “second course”, “side dish”, “bread”, and “food” (a generic class used when the item was not listed in the daily menu). A complementary “background” class was used to label all non-food regions in the image. Samples of the resulting annotations are shown in Figure 4.

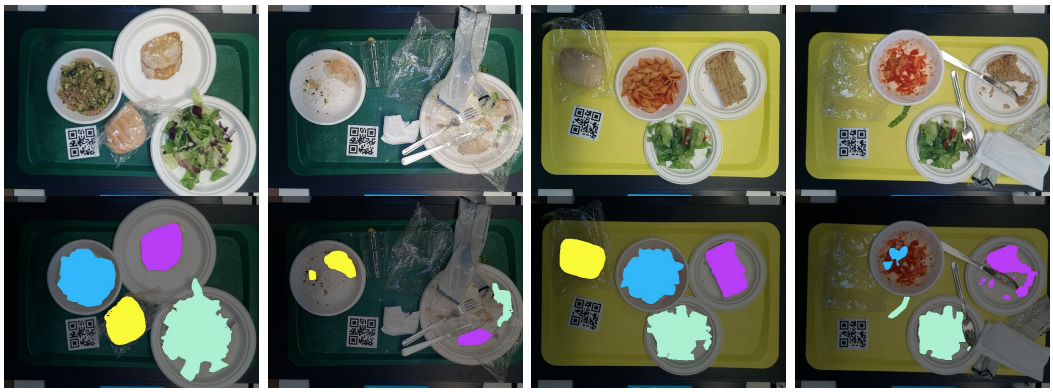

Figure 4. Sample images from the proposed FLIC dataset with corresponding pixel-precise semantic annotations. For the two tray samples (green tray and yellow tray), we show both full and leftover states. Blue segments indicate first course, magenta second course, teal side dish, yellow bread.

We then exploited the known menu structure of the canteen, where each day features, at most, one first course and one second course: this constraint allowed us to further map these general class labels to specific dishes from the corresponding daily menu.

3.4. Resulting FLIC Dataset

The final FLIC dataset was collected over 22 days between June and July 2024. It contains a total of 401 full-tray acquisitions plus the corresponding 401 leftover tray acquisitions. Images in the final dataset have a spatial resolution of 640 × 480 pixels as a result of downsampling the original full-resolution output of the acquisition device.

For each tray, we provide a segmentation mask that identifies food regions with an associated class (first course, second course, side dish, bread, and generic food). As explained in Section 3.3, the first-course and second-course classes can be further detailed by referencing the daily menu. This accounts for 17 first courses and 17 second courses plus bread and a generic side dish, for a total of 36 classes.

Each plate is also associated with a physically measured mass, following the protocol described in Section 3.1. At the tray level, the average tray mass for full trays is 388.58 g, and the average tray mass per leftover tray is 61.69 g. Concerning individual food items, the average measured mass on full trays is 121.56 g (for a total of 1166 distinct food items), whereas the average mass on leftover trays is 18.54 g per item. Detailed statistics are provided in Section 5.3.

The final week of acquisitions (5 days, totaling 111 tray pairs) is characterized by trays with the second course and side dish served on the same plate to enable future experimentation on mixed food-dish recognition.

This combination of paired visual data, dense semantic annotation, and physically measured mass provides an experimental setup that makes it possible to study food leftover estimation under realistic conditions. The annotated dataset is made available at https://drive.google.com/drive/folders/1IGCjr7y-_zsZfHTsL4CyxdkspTp8Uhnt (accessed on 26 May 2026).

3.5. FLIC in the Context of Multimodal Foundation Models

The newly introduced FLIC dataset can also be interpreted in the context of the current shift toward foundation models and Multimodal Large Language Models (MLLMs) [36]. The field of food recognition is increasingly moving away from task-specific supervised models toward these general-purpose architectures, which leverage extensive visual–linguistic pre-training to perform complex semantic reasoning. In the food domain, this

allows models to interpret culinary context, identify ingredients through natural language interaction, or evaluate the adequacy of dietary records through domain-specific knowledge [37]. In this evolving landscape, the FLIC dataset may serve as a high-fidelity benchmark for the "physical grounding" of advanced AI. While modern deep learning often prioritizes massive Web-crawled collections for pre-training, recent research has demonstrated that self-supervised representations, such as those from DINOv2 [38], can serve as a robust, unified backbone for diverse food-related tasks, including classification and segmentation in few-shot regimes [39]. Furthermore, as MLLMs are increasingly explored for portion and mass estimation through the integration of volumetric reasoning [40], FLIC addresses a critical gap by providing a physically measured mass ground truth paired with visual observations.

4. Methodology for Mass Estimation

This section introduces a simple and interpretable methodology whose primary purpose is to validate and analyze the proposed dataset rather than to define a state-of-the-art mass estimation model.

4.1. Digital Density

In order to provide a baseline for food mass estimation in our dataset, in the following, we define the concept of "digital density". In physical measurements, "density" (ρ) is a material property that relates the volume (V) of a sample with its mass (M):

$$\underset{[\mathrm{g}]}{M} = \underset{[\mathrm{m}^3]}{V} \cdot \underset{\left[\frac{\mathrm{g}}{\mathrm{m}^3}\right]}{\rho} . \tag{1}$$

In contrast, "digital density" (δ) is defined here as relating the apparent area (A) of a food sample, expressed in pixels, with its mass:

$$\underset{[\mathrm{g}]}{M} = \underset{[\mathrm{px}]}{A} \cdot \underset{\left[\frac{\mathrm{g}}{\mathrm{px}}\right]}{\delta} , \tag{2}$$

where the area (A) is measured under fixed imaging conditions (camera distance, focal length, and angle). The relationship assumed by the digital density formulation is not intended as a physically exhaustive model but as a controlled approximation that enables systematic analysis of the dataset.

To allow for generality, we cannot assume that two instances of food will have the exact same ratio (δ) of mass over area (even in the case of the same type of food). Therefore, we compute δ by solving an overdetermined system of linear equations defined by a *set* of apparent areas ($\vec{A}$) and corresponding annotated mass values ($\vec{W}$). The solution is obtained by minimizing the least-square error using the Moore–Penrose pseudo-inverse (PInv) [41]:

$$\delta = \mathrm{PInv}(\vec{A}) \cdot \vec{W}. \tag{3}$$

Specializing δ for different food classes is expected to reduce mass estimation errors: to assess the impact of such specialization, we conduct dedicated experiments that require the identification of the individual food classes on each tray. Thanks to the specific setup of our canteen environment, where only one first course and one second course are served each day, this identification can be achieved with minimal operator input. The full procedure is described in Section 4.2, which details the assisted food recognition stage.

The procedure to learn the digital densities is schematized in Figure 5: from a set of several trays, we generate semantically segmented visual maps using assisted food

recognition. From these semantic maps, a pixel count of their areas is easily extracted for each food class. Concurrently, we collect physical mass values following the protocol described in Section 3.1. This information is combined, per class, following Equation (3) to compile a table of digital densities.

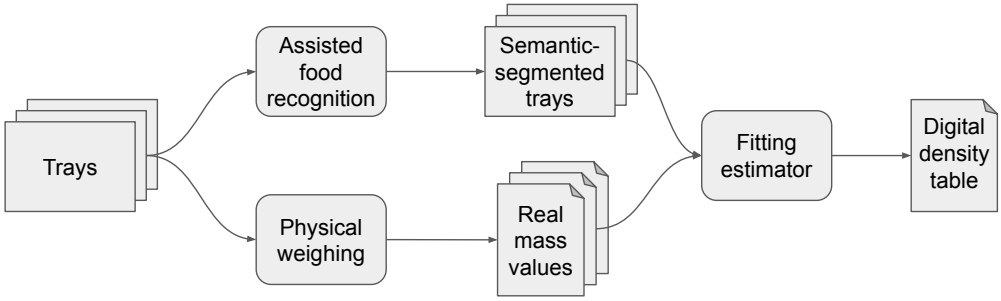

Figure 5. Procedure defined to learn digital densities, starting from physical weighing and assisted food recognition on full trays.

Super-Linear Models

Acknowledging the simplification introduced by the linear model of digital density, we take into consideration two additional models of increasing complexity that map the apparent area in pixels to the mass value: a second-degree polynomial and a multilayer perceptron (MLP). The second-degree polynomial is determined via least-squares optimization under the same constraints as the linear digital density model for mapping 0 pixels to 0 g. Regarding the multilayer perceptron, we opt for a configuration of 2 processing layers with 5 intermediate neurons each, as determined after preliminary experimentation. The perceptron weights are optimized via gradient backpropagation.

As we show with appropriate ablation studies in Section 5.4, super-linear models improve estimation performance on specific dishes but result in overall equivalent or worse performance when compared with the digital density approach.

4.2. Assisted Food Recognition

In this section, we delineate the procedure for assisted food recognition employed within this study.

As shown in Figure 6, assisted food recognition is achieved in two stages:

1. Food/no-food segmentation:
   Given an image of a tray, we automatically return a segmentation mask that identifies, per pixel, whether this represents food or other.
2. Food classification:
   The class of each food portion identified in the previous stage is defined at a broad level by manual input (first course, second course, side dish, or bread), and further refined automatically by relying on the daily menu.

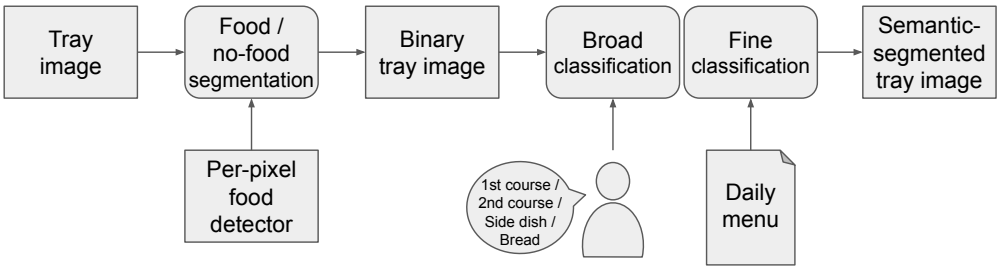

Figure 6. Detailed steps for assisted food recognition based on automatic food/no-food binary segmentation, manual broad classification of the food classes, and automatic fine-grained classification based on the daily menu.

Details for each stage are provided in the following. The result is a segmentation map of the food-tray image. This procedure is applied to both full trays and leftover trays.

4.2.1. Food/No-Food Segmentation

At the first stage of assisted food recognition, we are interested in providing an automatic binary segmentation of food vs no-food regions. To this extent, we implement, train, and evaluate different solutions based on deep learning:

- U-Net-like architecture [42]:
  As a baseline, we consider a traditional U-Net-like neural architecture where the encoder and decoder are composed of four blocks with skip connections. Each block is structured as two consecutive sequences of convolution–batch-normalization–ReLU layers and is connected to other blocks through a two-dimensional max-pooling layer.
- DABNet [43].
  Based on the Depth-wise Asymmetric Bottleneck module (DAB), DABNet generates a coarse segmentation map directly from a low-dimensional feature map without explicitly relying on a decoder module. In [44], it was shown to exhibit excellent accuracy in food/no-food segmentation while maintaining a moderate computational footprint. This led to the selection of DABNet as a representative for advanced supervised semantic segmentation within this work.
- DINOv2 features [38] combined with FeatUp upsampling [45] (DINOv2+FU):
  DINO (self-DIstillation with NO labels) is a self-supervised foundation vision transformer whose features have demonstrated remarkable capability in distinguishing semantic content in open-world scenarios. One significant limitation of DINO, i.e., the low spatial resolution of its features, has been addressed in past [46] with FeatUp, a model-agnostic framework capable of learning an upsampling function specific to a given embedding space. In this work, the 384-dimensional DINOv2 features, upsampled with FeatUp, are eventually mapped to our pixel-wise binary output through a single convolutional head with a kernel size of 1×1.
- SAM [35] prompted through Grounding DINO [47] (SAM+GDINO):
  The Segment Anything Model (SAM) is a general-purpose foundation model for segmentation designed to handle different types of input prompts, including points and bounding boxes. To fully automate its usage and avoid the need for user prompts, we use it in its "Language SAM" configuration [48]. In this setup, the input image is processed with Grounding DINO (DETR with Improved deNoising anchor boxes), an open-world object detector that returns bounding boxes based on an input textual prompt. In this case, the prompt is "food".

Of the four approaches, the first three require training to some extent. Namely, U-Net and DABNet were trained from scratch, adhering to a configuration that was consolidated on previous studies for food localization on canteen trays [44] and subsequently tuned to the FLIC dataset after preliminary assessment. This includes 50 epochs of training guided by binary cross-entropy loss, with gradients updated through the Adam optimizer using a fixed learning rate of 0.005 and no weight decay. No data augmentation was found to be beneficial. Experiments were conducted in a two-fold cross-validation setup to ensure fairness. The same configuration was adopted for DINOv2+FU; however, in this case, only the head was tuned to be applied on top of the pre-trained feature extractor. The fourth approach, SAM+GDINO, is the only one completely free from any tuning on our end, and it is selected to test the limits of zero-shot foundation models.

4.2.2. Food Classification

At the second stage of assisted food recognition, we consider a user-assisted classification of the food items in order to specialize mass estimation for each class. As a first approximation, in fact, mass estimation could be provided globally at the tray level by relying on a general-purpose digital density obtained with the application of Equation (3) to all food instances indiscriminately. However, it is of interest to quantify the extent to which having food-specific information can lead to a more accurate estimation of mass.

Given the constrains of fixed daily menus, we devise a two-level semi-automatic food classification procedure that involves user interaction to provide reliable identification of the food classes. This allows us to decouple and exclude the error of food recognition from that of food mass estimation.

1. A first level of broad classification is related to identifying each portion of food as belonging to either "first course", "second course", "side dish", or "bread'. For this input, we assume the interaction of our system with a dedicated operator or with the users themselves.
2. A second level of fine classification assigns a specific class to the food items (e.g., tomato pasta, white pasta, etc.). In our case, we rely on the specific constraints of our acquisition setup, where the menu is fixed for any given day; i.e., if the user chose to eat the first course, that plate can only contain one specific class. Combining the broad classification with the daily menu, it is therefore possible to automatically obtain a fine association of classes related to first and second courses. No details on the side-dish class are available.

If the accuracy gain in mass estimation as obtained by food classification is found to be significant, future developments might involve the use of additional computer vision techniques for the automated recognition of food classes [49].

5. Experiments for Mass Estimation

The methodology described in Section 4 serves as a transparent baseline to investigate the relationship between visual measurements and physical mass within the FLIC dataset. To assess this baseline, we present several experiments in the following. In Section 5.1, we evaluate different methods for food/no-food binary segmentation. In Section 5.2, we test the hypothesis of linear correlation between the apparent area in pixels and measured mass in grams. In Section 5.3, we apply and test our model for mass estimation based on digital densities. In Section 5.4, we report ablation studies for mass estimation based on non-linear predictors and the impact of food/no-food segmentation.

5.1. Food/No-Food Segmentation

Results for food/no-food binary segmentation are measured in terms of Intersection over Union (IoU):

$$\mathrm{IoU} = \frac{\sum_{i=1}^{N}(\hat{p}_i \wedge p_i)}{\sum_{i=1}^{N}(\hat{p}_i \vee p_i)}, \tag{4}$$

where $\hat{p}_i$ is the predicted set of food pixels and p_i is the ground truth of the i-th image. Two-fold cross validation is used for evaluation, with samples from both full trays and leftover trays used at training and test time.

Results are reported in Table 2: all methods considered for comparison show similar performance on full trays, landing at around 92–93% IoU. However, a significant difference is shown on leftover trays. As a general comment, we can state that leftover trays represent a significantly more challenging scenario than full trays, even for state-of-the-art segmen-

tation models: while more traditional encoder–decoder architectures U-Net and DABNet achieve between 70% and 71%, DINOV2+FeatUp reaches the highest 77.22% IoU, making it the model of choice for our subsequent experiments. SAM+GDINO, which is the only method that does not involve any form of training, also shows competitive performance on full plates but drops significantly in leftover detection.

Table 2. Intersection over union results for food/no-food binary segmentation with different methods.

Method	Training	IoU (Full Trays)↑	IoU (Leftover Trays)↑
U-Net-like [42]	Full	92.52%	70.05%
DABNet [43]	Full	93.64%	71.12%
DINOv2+FU [38,45]	Head only	93.26%	77.22%
SAM+GDINO [35,47]	None	91.98%	31.77%

Visual examples of binary food/no-food segmentation are provided in Figure 7.

In Section 5.4, we also quantify the impact of imperfect food/no-food segmentation on the downstream task of food mass estimation.

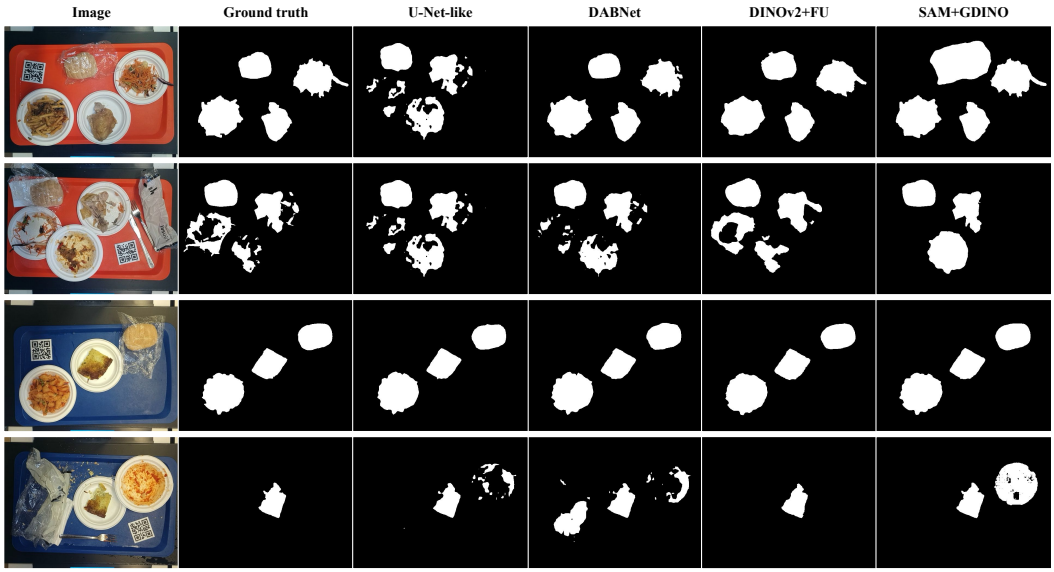

Figure 7. Examples of binary food/no-food segmentation as obtained with different methods on full and leftover trays.

5.2. Area-to-Mass Linear Correlation Test

Estimating mass values via digital density assumes the existence of a linear relationship between food area in pixels and mass in grams. Here, we test this hypothesis by conducting a correlation analysis of the collected dataset.

The results are shown in Figure 8. Every point represents a plate: full-plate observations are reported in blue, and leftovers are reported in orange. Plates with mixed dishes (second course and side dish on the same plate) are excluded from this analysis. The red line represents a purely linear regression with null offset. For the first and second courses (first row) we also report a subset plot of the best and worst correlated classes in Figure 9. For reference, the semi-transparent red line is the same linear regressor from Figure 8, i.e., at the broad level (first course and second course) instead of the fine level. Note the different ranges of the plots.

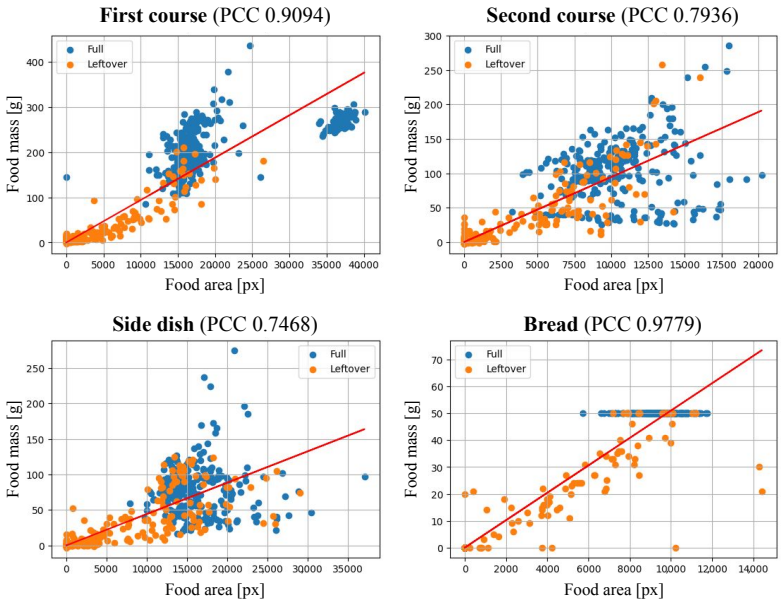

Figure 8. Scatter plots correlating apparent area in pixels and mass in grams,for broad-level classes. Full-tray acquisitions in blue and leftovers in orange. Linear correlations described with a red line and quantified by Pearson Correlation Coefficients (PCCs).

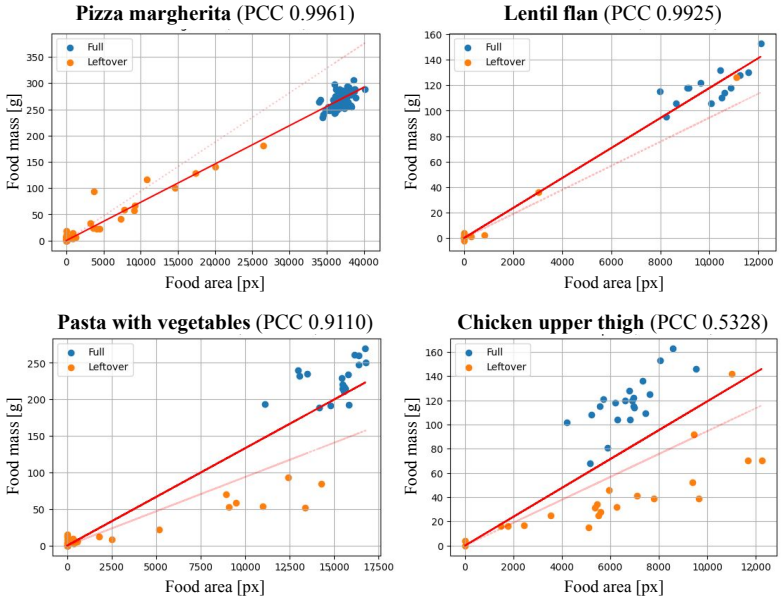

Figure 9. Scatter plots correlating apparent area in pixels and mass in grams for fine-level sample classes. Full-tray acquisitions in blue and leftovers in orange. Linear correlations described with a red line and quantified by Pearson Correlation Coefficients (PCCs).

The first course exhibits a very good Pearson Correlation Coefficient (PCC) of 90.04%. Its plot, however, shows that a non-linear correlation would better fit the great majority of the data and that there is a visible outlier. The outlier is class "Pizza margherita", which incidentally has the highest correlation among first-course sub-classes (99.61%). The worst correlation is for "Pasta with vegetables", although still reaching a solid 91.10%. In this case, the non-linearity can be explained by the heterogeneous nature of the meal: the pasta itself and the vegetables have inherently different digital densities, which are not modeled separately in our scenario. The full meal (blue points) is a mix of the two components, while the leftover (orange points) tends to be mostly composed of vegetables, as revealed by visual inspection. This effectively leads to two different linear relationships between

full and leftover state for the same meal. Visual examples of the aforementioned classes of interest are provided in Figure 10.

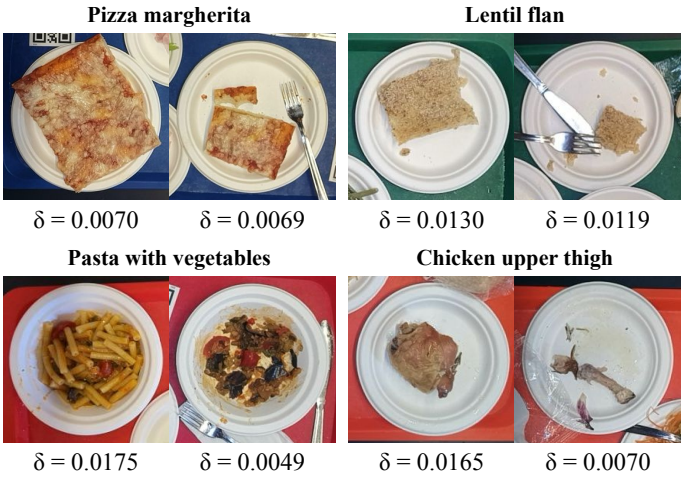

Figure 10. Examples of classes from our FLIC dataset with corresponding digital density values (δ) according to the specific sample shown. The classes of "Pizza margherita" and "Lentil flan" are the most consistent between full and leftover state, whereas "Pasta with vegetables" and "Chicken upper thigh" show significant differences before and after consumption.

The second course exhibits worse linearity overall, with correlation at around 80% and a visibly scattered plot. Within this group, some individual dishes, such as "Lentil flan", do have excellent linear correlation (99.25%). The worst second course is "Chicken upper thigh", at 53.28% correlation. Similarly to "Pasta with vegetables", this is also a case where two components of the dish have inherently different digital densities: the meat and the bone. In this case, however, the bone is mostly invisible on the full plate and mostly visible on the leftover plate. This creates two independent point clouds that cannot be modeled, even with a non-linear correlation. It also suggests that for given food classes, some leftover quantity always be present and should not be counted as waste.

The side-dish class exhibits decent behavior overall, not dissimilar to that of the second-course class, with 74.68% correlation. In this case, however, it is not possible to define food-specific relationships, as the canteen does not provide detailed sub-class labels for side dishes.

Bread shows excellent linear correlation (97.79%). The plot also clearly shows how the full-tray bread information was set to a default of 50 g without actual physical weighing.

This analysis highlights the variable level of reliability that can be attributed to linear modeling of the area-to-mass relationship. In the next section, we will quantify how this linear modeling actually impacts the task of mass estimation.

5.3. Mass Estimation

5.3.1. Experimental Setup

We learn digital densities by applying Equation (3) to data extracted from full-tray images. We then verify the effectiveness of digital densities for the estimation of mass from leftover tray images. This setup allows us to better observe the distribution of estimation errors, since the leftover trays naturally present a more widespread range of mass values.

Our food recognition pipeline based on binary food/no-food segmentation does not allow us to distinguish mixed dishes (i.e., multiple types of food in close contact within the same plate). Since 111 trays out of 401 contain a second course and a side dish in the same plate, we exclude those specific mixed-dish plates from the mass estimation experiment.

Let $\mathcal{F} = \{1, 2, \ldots, 10\}$ denote the set of possible training-set sizes (i.e., the number of samples), and let $R = 100$ be the number of independent runs for each size. We adhere to the following procedure:

- For each $F \in \mathcal{F}$, collect F random training samples of the class of interest from full-tray images, along with the corresponding measured mass annotation. This emulates a hypothetical "enrollment" phase in the canteen, where few instances of food are individually measured for reference.
- Estimate the digital density, relating area in pixels to mass in grams by application of Equation (3).
- Select all L food regions that correspond to the class of interest identified on leftover food images.
- Apply the learned digital density to the L leftover food regions, remapping the food area to estimated leftover mass.
- Compute the Mean Absolute Error (MAE) between the measured leftover mass (m) and estimated leftover mass ($\hat{m}$):

$$\text{MAE} = \frac{1}{L} \sum_{i=1}^{L} |m_i - \hat{m}_i|. \tag{5}$$

- Repeat the procedure R times with different random selections of F training samples and compute the mean and standard deviation of the resulting MAE values.

In order to assess the impact of classification errors on the final quality estimation results, we consider three levels of food classification in our experiments: no classification (instances from all food classes are considered), broad classification (instances of first course, second course, side dish, and bread are considered separately), and fine classification (instances of all specific first-course and second-course classes are considered separately).

5.3.2. Experimental Results

In Table 3, we report results for global statistics, i.e., errors computed on all leftover plates (only excluding mixed-dish plates, as in all subsequent experiments). MAE-1 is the error obtained by learning the digital density of just $T = 1$ random training example, as averaged for $R = 100$ runs. MAE-1/full is a relative error provided to gauge the impact of MAE-1 by dividing it for the average mass of a full plate. An alternative option would be to normalize MAE-1 for the average leftover mass. This, however, would lead to numerically unstable and uninformative results, since leftover mass values are often extremely small (especially for the detailed tables presented in the following). The effect of discriminating food classes with specialized digital densities, as indicated by the "Classification" column, is measurable but limited, lowering the relative error by two percentage points.

Table 3. Global statistics for leftover mass estimation. The "Classification" column identifies different specializations of the digital density used in mass estimation. For each, we report both absolute and relative errors.

Course	Classification	No. of Samples	Avg Full [g]	Avg Leftover [g]	MAE-1 [g]	MAE-1/Full
All	No classification	1166	121.56	18.54	14.30	11.76%
	Broad classification				13.04	10.73%
	Fine classification				11.47	9.44%

In Table 4, we report a detailed view of the impact of broad classification and fine classification on course-level statistics. Fine classification (i.e., identifying the specific food class) is only available for the first and second course, since the canteen did not provide detailed information about various side dishes. This experiment shows that fine

classification has a significant impact in terms of reducing the error for the second course and virtually no impact (if not a slightly negative impact) on the first course. This can be explained by the relatively uniform nature of first courses in this dataset (different types of pasta) against the heterogeneity of second courses. It should also be noted that several food classes exhibit limited representation in the evaluation set; therefore, performance metrics for these classes should be interpreted with caution due to statistical instability and limited generalization capability.

Table 4. Course-level statistics for leftover mass estimation. The "Classification" column identifies different specializations of the digital density used in mass estimation. For each, we report both absolute and relative errors.

Course	Classification	No. of Samples	Avg Full [g]	Avg Leftover [g]	MAE-1 [g]	MAE-1/Full
First course	Broad classification	393	207.98	17.59	12.00	5.77%
First course	Fine classification				12.38	5.95%
Second course	Broad classification	274	102.31	23.76	20.81	20.34%
Second course	Fine classification				13.60	13.29%
Side dish	Broad classification	241	79.12	25.14	16.19	20.46%
Bread	Broad classification	258	50.00	8.29	3.43	6.86%

Figure 11 shows the MAE curves and confidence intervals for the broad classification of "first course", "second course", "side dish", and "bread". Here, a blue line plots the MAE with the number of training samples varying from $T = 1$ up to $T = 10$. The blue area encodes the standard deviation resulting from $R = 100$ independent runs. The dashed red line is the average mass for leftovers. The average mass for full plates is reported numerically in the figure but not visualized for the sake of compactness. The plot shows how, with more training examples, it is possible to reduce the average MAE. This is visible in the descending blue line. Furthermore, relying on multiple training examples reduces the difference in behavior across runs, since the aleatority of random sample selection has a lesser impact. This is visible in the reducing blue area. In some cases (such as "second course"), the improvement is more apparent than in others ("first course"), a behavior that can also be explained by the more heterogeneous nature of second courses.

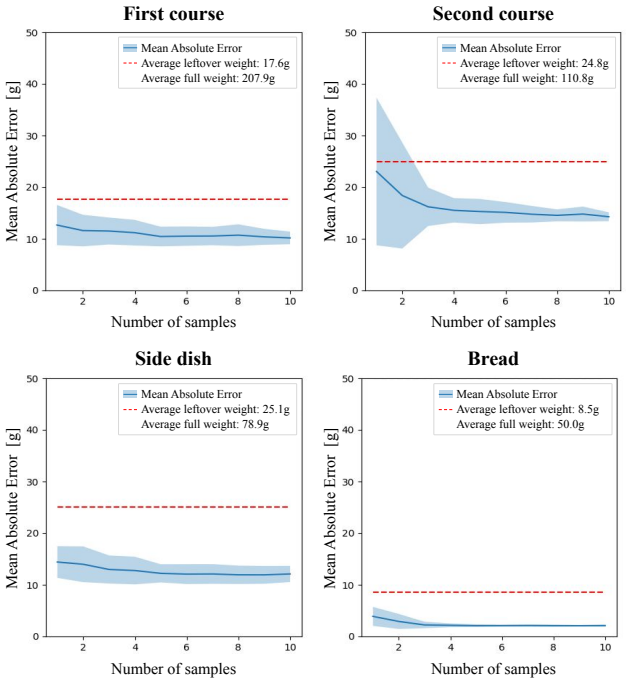

Figure 11. MAE with varying numbers of training samples used to define the digital density. The blue area indicates the standard deviation across multiple runs.

Tables 5 and 6 provide additional details about the error distribution in first courses and second courses, respectively. The impact on mass estimation can be seen to be generally in line with the correlation analysis from Section 5.2. For example, "Pasta with vegetables" shows the worst absolute and relative MAE-1 values among first courses, and it also shows the least correlation between area and mass. Similarly, "Chicken upper thigh" has the worst MAE-1 for second courses. Other notable examples include the excellent mass estimation for the first course of "Ricotta and spinach tortelli" and for the second course of "Gratin-style plaice fillets".

Table 5. First-course details for leftover mass estimation. For each dish, we report both absolute and relative errors.

First-Course Dish	No. of Samples	Avg Full [g]	Avg Leftover [g]	MAE-1 [g]	MAE-1/Full
Pasta Bolognese style	45	195.27	4.51	5.49	2.81%
Whole pasta with tomato and basil	38	227.37	23.71	16.52	7.27%
Whole pasta with zucchini	12	177.00	85.33	18.71	10.57%
Pasta with eggplant and tomatoes	15	147.60	15.80	9.06	6.14%
Rice with tomato and basil	14	226.79	27.36	24.48	10.79%
Pizza margherita	71	269.75	20.06	7.76	2.88%
Pasta with olive oil and Parmesan	16	119.69	2.94	3.24	2.71%
Whole pasta with eggplant	12	185.92	17.42	7.82	4.21%
Pasta with tomato and tuna	14	176.57	30.71	18.73	10.61%
Ricotta and spinach tortelli	10	204.10	3.80	3.52	1.72%
Pasta with zucchini	11	184.18	6.09	10.21	5.54%
Pasta with tomato and basil	23	193.57	8.30	11.75	6.07%
Rice with oil and Parmesan	22	172.27	22.82	14.97	8.69%
Pasta with vegetables	19	226.68	30.05	33.46	14.76%
Pasta with tomato and peas	23	237.96	11.48	23.34	9.81%
Pasta with pesto sauce	25	184.96	3.96	4.23	2.29%
Three-color cold pasta salad	23	187.61	14.04	13.14	7.00%

Table 6. Second-course details for leftover mass estimation. For each dish, we report both absolute and relative errors.

Second Course Dish	N Samples	Avg Full [g]	Avg Leftover [g]	MAE-1 [g]	MAE-1/Full
Zucchini omelet	23	135.70	46.57	13.91	10.25%
Chickpea and potato pattie	18	144.06	6.44	2.60	1.80%
Roast pork	12	94.00	28.33	23.12	24.60%
Gratin-style cuttlefish	15	79.73	31.27	10.62	13.32%
Savory chard and spinach pie	12	206.08	116.67	32.03	15.54%
Cooked ham	46	45.33	5.87	4.38	9.66%
Lentil flan	14	118.93	12.14	3.12	2.62%
Turkey bites with lemon	12	116.50	28.92	14.39	12.35%
Spinach omelet	11	104.73	8.09	3.46	3.30%
Parmesan cheese	10	40.00	8.30	33.09	82.73%
Cod fillets with potatoes	11	85.55	13.18	7.25	8.47%
Legume pattie	22	117.00	23.82	6.95	5.94%
Gratin-style plaice fillets	23	90.43	0.17	0.26	0.29%
Chicken upper thigh	21	117.71	39.71	66.20	56.24%
Swiss chard and ricotta omelet	22	115.59	25.73	4.81	4.16%
Cod with tomato and olives	2	106.00	41.00	8.26	7.79%

5.4. Ablation Studies

In this section, we report ablation studies aimed at quantifying the impact of super-linear models and the impact of food/no-food segmentation errors.

Tables 7 and 8 focus on replacing the linear model of digital density with a second-degree polynomial and a multilayer perceptron. We report MAE-1 and MAE-10 errors computed with $T = 1$ and $T = 10$ training examples, respectively, based on results averaged from $R = 100$ runs. We also report the MAE differences such that a positive difference indicates an improvement with super-linear models.

As can be observed, both for the polynomial and the perceptron approaches, the overall effect on MAE-1 is markedly negative: this is explained by the numerical instability intrinsic in learning more complex models with little data. The overall differences are

evened out once more training samples are available (MAE-10), which suggests that more complex models are not, in an average case, more useful than the digital density approach. Specific dishes of interest, as emerged in the previous sections, are also reported to assess the specific positive impact of super-linear modeling (e.g., "Pasta with vegetables" and "Chicken upper thigh" in Table 8).

Table 7. Ablation study that compares the linear model of digital density against a more complex second-degree polynomial (2nd deg). Reported are MAE-1 and MAE-10 errors, together with Differences (Diff.). A positive difference indicates an improvement achieved by the super-linear model.

Course	Classification	MAE-1			MAE-10		
		Linear	2nd deg	Diff.	Linear	2nd deg	Diff.
All	No classification	14.30	414.85	−400.55	10.99	12.33	−1.34
All	Broad classification	13.04	81.62	−68.58	9.03	11.91	−2.88
All	Fine classification	11.47	14.88	−3.41	9.54	10.36	−0.82
First course	Broad classification	12.00	10.88	+1.12	10.01	14.61	−4.60
First course	Fine classification	12.38	10.37	+2.01	12.02	12.13	−0.11
↳ Pizza margherita		7.76	14.00	−6.24	7.21	9.83	−2.62
↳ Pasta with vegetables		33.46	16.88	+16.58	33.69	37.70	−4.01
Second course	Broad classification	20.81	311.48	−290.67	11.44	14.17	−2.73
Second course	Fine classification	13.60	28.18	−14.58	10.70	11.11	−0.41
↳ Lentil flan		3.12	4.98	−1.86	1.33	1.38	−0.05
↳ Chicken upper thigh		66.20	102.94	−36.74	64.57	58.47	+6.10
Side dish	Broad classification	16.19	18.77	−2.58	12.21	14.81	−2.60
Bread	Broad classification	3.43	3.99	−0.56	2.02	2.69	−0.67

Table 8. Ablation study that compares the linear model of digital density against a more complex multilayer perceptron (MLP). Reported are MAE-1 and MAE-10 errors, together with Differences (Diff.). A positive difference indicates an improvement achieved by the super-linear model.

Course	Classification	MAE-1			MAE-10		
		Linear	MLP	Diff.	Linear	MLP	Diff.
All	No classification	14.30	190.96	−176.66	10.99	12.12	−1.13
All	Broad classification	13.04	81.39	−68.35	9.03	10.81	−1.78
All	Fine classification	11.47	90.23	−78.76	9.54	11.40	−1.87
First course	Broad classification	12.00	15.86	−3.86	10.01	11.17	−1.16
First course	Fine classification	12.38	16.05	−3.67	12.02	13.06	−1.04
↳ Pizza margherita		7.76	12.38	−4.62	7.21	9.97	−2.76
↳ Pasta with vegetables		33.46	39.95	−6.49	33.69	32.54	+1.15
Second course	Broad classification	20.81	28.08	−7.27	11.44	13.88	−2.44
Second course	Fine classification	13.60	65.45	−51.85	10.70	13.68	−2.98
↳ Lentil flan		3.12	8.82	−5.70	1.33	3.23	−1.90
↳ Chicken upper thigh		66.20	186.60	−120.40	64.57	60.60	+3.97
Side dish	Broad classification	16.19	303.74	−287.55	12.21	15.00	−2.79
Bread	Broad classification	3.43	30.11	−26.68	2.02	3.10	−1.08

Table 9 focuses on the impact of replacing automated food/no-food segmentation, as obtained via the DINOV2+FeatUp approach presented in Section 4.2.1, with the human annotations used to train said model. In this context, the human annotations are used as a proxy for perfect segmentation. The information is presented following the same scheme as Tables 7 and 8, and here, a positive difference indicates an improvement gained from perfect human segmentation. The results show that, overall, the effect of imperfect segmentation (measured at 93.26% IoU on full trays and 77.22% on leftover trays in Section 5.1) is actually negligible on the downstream task of mass estimation. This is explained by the fact that,

for each configuration, the digital densities are learned from the corresponding semantic annotation (human or annotated) and, as such, sufficiently compensate for over- and under-segmentations.

Table 9. Ablation study that compares the automated food/no-food segmentation obtained through DINOV2+FeatUp with human-annotated masks. Reported are MAE-1 and MAE-10 errors, together with Differences (Diff.). A positive difference indicates an improvement gained from human annotation.

		MAE-1			MAE-10		
Course	**Classification**	**Automated**	**Human**	**Diff.**	**Automated**	**Human**	**Diff.**
All	No classification	14.30	14.62	−0.32	10.99	11.18	−0.19
All	Broad classification	13.04	11.61	+1.43	9.03	8.94	+0.10
All	Fine classification	11.47	11.46	+0.01	9.54	9.72	−0.18
First course	Broad classification	12.00	12.50	−0.50	10.01	10.31	−0.30
First course	Fine classification	12.38	12.77	−0.40	12.02	12.47	−0.45
↳ Pizza margherita		7.76	7.21	+0.55	7.21	6.71	+0.50
↳ Pasta with vegetables		33.46	40.87	−7.41	33.69	41.07	−7.38
Second course	Broad classification	20.81	13.98	+6.83	11.44	10.66	+0.78
Second course	Fine classification	13.60	12.97	+0.62	10.70	10.87	−0.17
↳ Lentil flan		3.12	3.41	−0.29	1.33	1.54	−0.21
↳ Chicken upper thigh		66.20	68.66	−2.46	64.57	66.15	−1.58
Side dish	Broad classification	16.19	16.49	−0.30	12.21	12.49	−0.28
Bread	Broad classification	3.43	3.16	+0.27	2.02	1.70	+0.32

6. Conclusions

This work is primarily intended as a dataset contribution, providing a realistic experimental foundation for studying food leftover estimation rather than proposing a novel estimation model. Specifically, we introduced FLIC: a novel annotated dataset for image-based estimation of food mass in collective catering environments. The dataset includes 401 full and 401 leftover tray images with per-pixel semantic labels and precise physical mass measurements collected in a real-world canteen over 22 days. We defined the concept of digital density to relate image area to food mass and provided baseline results through a lightweight estimation pipeline combining food/no-food segmentation and assisted food recognition. These experiments serve to validate the dataset's applicability and establish an initial benchmark for future research on visual food-waste estimation.

While the FLIC dataset provides high-fidelity paired visual and mass annotations, its current scale remains relatively modest. This may affect the statistical stability of some findings and limit its future usage for deep learning methods. Furthermore, its single-site origin and fixed menu may introduce inherent geographic and cultural biases. These constraints reflect the inherent difficulty of acquiring large-scale, culturally diverse data, particularly the significant logistical challenge of manually weighing every individual food item in a high-throughput operational canteen. To mitigate the risk of overfitting and enhance generalization, future efforts will focus on multi-site collection to encompass diverse food compositions, lighting conditions, and tableware types. Furthermore, we intend to explore advanced data augmentation strategies such as the generation of synthetic leftover tray configurations or the use of generative models to simulate the morphological variability and food mixing typically seen in post-consumption scenarios.

The proposed digital density model relies on the assumption of a linear relationship between image area in pixels and food mass in grams. While this assumption is supported in several cases by our correlation analysis, it does not hold uniformly across all food types, especially for heterogeneous or multi-component dishes. For this reason, future work will investigate the selective application of non-linear models to mass estimation

of specific foods that are found to benefit from increased complexity. Closely related to this limitation is the fact that our current formulation uses only the apparent area as a predictor. A more physically grounded approach would relate food mass to estimated volume rather than projected area, as demonstrated in research utilizing overhead RGB-D sensing [22], voxel-based reconstruction from depth maps [50], or 3D object scaling [24]. Such methods leverage depth estimation or multi-view reconstruction to achieve precise portion quantification, though they typically require specialized hardware or strong 3D priors not available in our standard 2D setup.

Our experiments also show that broad and fine food classifications lead to uneven gains in estimation accuracy across courses. Although class-specific modeling does not consistently improve performance, detailed food categorization remains desirable for reporting and monitoring purposes. For this reason, future developments will explore more advanced computer vision methods for automatic food classification, reducing the need for user assistance while preserving class-level reporting.

Finally, an open challenge between leftover food mass estimation and true waste estimation is the treatment of non-edible fractions and of edible residues that are practically unrecoverable with standard utensils. Explicit identification of non-waste leftovers will be necessary to more accurately translate visual mass estimates into measures of effective food waste.

Author Contributions: Conceptualization, S.B., G.C., R.S. and F.S.; methodology, F.P., D.C., D.M., M.B., C.F. and G.C.; software, F.P., D.M. and M.B.; validation, F.P., D.C., M.B., C.F. and L.S.; formal analysis, D.C., M.B., C.F. and L.S.; investigation, D.C., M.B., C.F. and L.S.; resources, D.M., R.S. and F.S.; data curation, F.P., D.C., M.B., C.F. and L.S.; writing—original draft preparation, D.C., D.M., M.B. and C.F.; writing—review and editing, F.P., D.C., D.M., M.B., C.F., L.S., S.B., G.C., R.S. and F.S.; visualization, M.B.; supervision, S.B., G.C., R.S. and F.S.; project administration, R.S. and F.S.; funding acquisition, S.B., G.C., R.S. and F.S. All authors have read and agreed to the published version of the manuscript.

Funding: Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for tender No. 341 of 15 March 2022 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU. Project code PE00000003, Concession Decree No. 1550 of 11 October 2022 adopted by the Italian Ministry of University and Research, CUP D93C22000890001, Project title “ON Foods - Research and innovation network on food and nutrition Sustainability, Safety and Security – Working ON Foods”. This work was partially supported by the MUR under the grant “Dipartimenti di Eccellenza 2023-2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Because the study involved only anonymized photographs and plate-level mass measurements without collecting personal or identifiable data, formal ethical approval was not required.

Data Availability Statement: The dataset generated and analyzed during the current study is made available at https://drive.google.com/drive/folders/1IGCjr7y-_zsZfHTsL4CyxdkspTp8Uhnt (accessed on 26 May 2026).

Acknowledgments: The authors wish to thank all participants involved in the study. Moreover, we gratefully thank the Giocampus project and Cimas Ristorazione for their support in data acquisition.

Conflicts of Interest: No authors of this manuscript have any competing financial or non-financial interests, currently or previously, including serving in an editorial capacity for the journal we are submitting to.

Abbreviations

The following abbreviations are used in this manuscript:

2D	Two-Dimensional
3D	Three-Dimensional
CVAT	Computer Vision Annotation Tool
DABNET	Depth-wise Asymmetric Bottleneck Network
DETR	DEtection with TRansformers
DINO	self-DIstillation with NO labels
DSLR	Digital Single-Lens Reflex
ECUSTFD	East China University of Science and Technology Food Dataset
EU	European Union
FLIC	Food Leftovers In Canteens
FU	FeatUp
GB	Giga Bytes
GDINO	Grounding DETR with Improved deNoising anchor boxes
IoU	Intersection over Union
MAE	Mean Absolute Error
MedGRFood	Mediterranean Greek Food
MLP	Multi-Layer Perceptron
MLLM	Multimodal Large Language Model
QR code	Quick Response code
LED	Light Emitting Diode
PCC	Pearson Correlation Coefficient
RAM	Random Access Memory
RGB	Red Green Blue
RGB-D	Red Green Blue and Depth
SAM	Segment Anything Model
UNIMIB	University of Milano-Bicocca
YOLO	You Only Look Once

References

1. Liu, D.; Zuo, E.; Wang, D.; He, L.; Dong, L.; Lu, X. Deep learning in food image recognition: A comprehensive review. *Appl. Sci.* **2025**, *15*, 7626. [CrossRef]
2. Liu, C.; Cao, Y.; Luo, Y.; Chen, G.; Vokkarane, V.; Ma, Y. Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment. In *Proceedings of the International Conference on Smart Homes and Health Telematics*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 37–48.
3. He, J.; Shao, Z.; Wright, J.; Kerr, D.; Boushey, C.; Zhu, F. Multi-task image-based dietary assessment for food recognition and portion size estimation. In Proceedings of the 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Shenzhen, China, 6–8 August 2020; IEEE: New York, NY, USA, 2020; pp. 49–54.
4. Gonzalez, B.; Garcia, G.; Velastin, S.A.; GholamHosseini, H.; Tejeda, L.; Farias, G. Automated Food Weight and Content Estimation Using Computer Vision and AI Algorithms. *Sensors* **2024**, *24*, 7660. [CrossRef] [PubMed]
5. Ciocca, G.; Napoletano, P.; Schettini, R. Food recognition and leftover estimation for daily diet monitoring. In *Proceedings of the International Conference on Image Analysis and Processing*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 334–341.
6. Ciocca, G.; Napoletano, P.; Schettini, R. Food recognition: A new dataset, experiments, and results. *IEEE J. Biomed. Health Inform.* **2016**, *21*, 588–598. [CrossRef] [PubMed]
7. Dalakleidi, K.V.; Papadelli, M.; Kapolos, I.; Papadimitriou, K. Applying image-based food-recognition systems on dietary assessment: A systematic review. *Adv. Nutr.* **2022**, *13*, 2590–2619. [CrossRef] [PubMed]
8. Amugongo, L.M.; Kriebitz, A.; Boch, A.; Lütge, C. Mobile computer vision-based applications for food recognition and volume and calorific estimation: A systematic review. *Healthcare* **2022**, *11*, 59. [CrossRef] [PubMed]
9. Andrews, L.; Kerr, G.; Pearson, D.; Mirosa, M. The attributes of leftovers and higher-order personal values. *Br. Food J.* **2018**, *120*, 1965–1979. [CrossRef]
10. Aloysius, N.; Ananda, J.; Mitsis, A.; Pearson, D. Why people are bad at leftover food management? A systematic literature review and a framework to analyze household leftover food waste generation behavior. *Appetite* **2023**, *186*, 106577. [CrossRef] [PubMed]

11. UN Environment. Food Waste Index Report 2024 | UNEP-UN Environment Programme. 2024. Available online: https://www.unep.org/resources/publication/food-waste-index-report-2024 (accessed on 26 May 2026).
12. Eurostat. Food Waste and Food Waste Prevention-Estimates-Statistics Explained-Eurostat. 2025. Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Food_waste_and_food_waste_prevention_-_estimates (accessed on 26 May 2026).
13. Sbraga, L.; Erba, G.R.; Ferretti, D.; Boroni, B.; Silvestri, S.; Pelò, F.; Zatelli, G. Rapporto Ristorazione FIPE 2025. 2025. Available online: https://www.fipe.it/wp-content/uploads/2025/08/RAPPORTO-RISTORAZIONE-2025-1-1.pdf (accessed on 26 May 2026).
14. Mordor Intelligence. Europe Foodservice Market Size & Share Analysis-Industry Research Report-Growth Trends. 2023. Available online: https://www.mordorintelligence.com/industry-reports/europe-foodservice-market (accessed on 26 May 2026).
15. Principato, L.; Marchetti, S.; Barbanera, M.; Ruini, L.; Capoccia, L.; Comis, C.; Secondi, L. Introducing digital tools for sustainable food supply management: Tackling food loss and waste in industrial canteens. *J. Ind. Ecol.* **2023**, *27*, 1060–1075. [CrossRef]
16. Roka, K. Environmental and social impacts of food waste. In *Responsible Consumption and Production*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 216–227.
17. Amicarelli, V.; Bux, C. Food waste measurement toward a fair, healthy and environmental-friendly food system: A critical review. *Br. Food J.* **2021**, *123*, 2907–2935. [CrossRef]
18. Ferreira, J.; Cerqueira, P.; Ribeiro, J. Food Waste Detection in Canteen Plates Using YOLOv11. *Appl. Sci.* **2025**, *15*, 7137. [CrossRef]
19. Boelter, F.; Venner, F. Food Waste Dataset. 2024. Available online: https://huggingface.co/datasets/AI-ServicesBB/food-waste-dataset (accessed on 26 May 2026).
20. Sari, Y.A.; Nakazawa, A.; Wani, Y.A. LeFood-set: Baseline performance of predicting level of leftovers food dataset in a hospital using MT learning. *PLoS ONE* **2025**, *20*, e0320426. [CrossRef] [PubMed]
21. Liang, Y.; Li, J. Computer vision-based food calorie estimation: Dataset, method, and experiment. *arXiv* **2017**, arXiv:1705.07632. [CrossRef]
22. Thames, Q.; Karpur, A.; Norris, W.; Xia, F.; Panait, L.; Weyand, T.; Sim, J. Nutrition5k: Towards automatic nutritional understanding of generic food. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8903–8911.
23. Tai, C.e.A.; Nair, S.; Markham, O.; Keller, M.; Wu, Y.; Chen, Y.; Wong, A. Nutritionverse-real: An open access manually collected 2d food scene dataset for dietary intake estimation. *arXiv* **2023**, arXiv:2401.08598.
24. Vinod, G.; He, J.; Shao, Z.; Zhu, F. Food portion estimation via 3D object scaling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 3741–3749.
25. Boelter, F.; Venner, F.; Rueda-Toicen, A. Food Waste Dataset with FiftyOne Enhancements. 2024. Available online: https://huggingface.co/datasets/andandandand/food-waste-dataset (accessed on 26 May 2026).
26. Wang, A.; Liu, L.; Chen, H.; Lin, Z.; Han, J.; Ding, G. YOLOE: Real-Time Seeing Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–20 October 2025; pp. 24591–24602.
27. Sari, Y.A.; Adinugroho, S.; Maligan, J.M.; Candra, E.N.; Utaminingrum, F.; Nur'Aini, N. Leftovers food recognition using deep neural network and regression approach for objective visual analysis estimation. In *Proceedings of the 2021 4th International Conference of Computer and Informatics Engineering (IC2IE), Depok, Indonesia, 14–15 September 2021;* IEEE: New York, NY, USA, 2021; pp. 24–29.
28. Hsieh, Y.; Dai, X.; Tanimoto, H. Recognition of Leftover Food Based on Deep Learning. In *Proceedings of the International Symposium on Automation, Mechanical and Design Engineering*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 207–215.
29. Rokhva, S.; Teimourpour, B. Artificial Intelligence in the Food Industry: Food Waste Estimation based on Computer Vision, a Brief Case Study in a University Dining Hall. *arXiv* **2025**, arXiv:2507.14662. [CrossRef]
30. Li, Y.; Zhang, C.; Xu, H.; Yang, Y.; Lu, H.; Deng, L. Strategies for Automated Identification of Food Waste in University Cafeterias: A Machine Vision Recognition Approach. *Appl. Sci.* **2025**, *15*, 5036. [CrossRef]
31. Health Canada. Canadian Nutrient File (CNF)-Search by Food. 2024. Available online: https://open.canada.ca/data/en/dataset/1b6139bd-ed7e-4043-bc28-ff00e10f3109 (accessed on 28 April 2026).
32. Konstantakopoulos, F.S.; Georga, E.I.; Fotiadis, D.I. An automated image-based dietary assessment system for mediterranean foods. *IEEE Open J. Eng. Med. Biol.* **2023**, *4*, 45–54. [CrossRef]
33. Giocampus. Giocampus-Project-Giocampus. 2021. Available online: https://www.giocampus.it/en/giocampus-project/ (accessed on 26 May 2026).
34. CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT). 2023. Available online: https://github.com/cvat-ai/cvat (accessed on 26 May 2026).
35. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 4015–4026.

36. Caffagni, D.; Cocchi, F.; Barsellotti, L.; Moratelli, N.; Sarto, S.; Baraldi, L.; Cornia, M.; Cucchiara, R. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 13590–13618.
37. Romero-Tapiador, S.; Tolosana, R.; Lacruz-Pleguezuelos, B.; Marcos-Zambrano, L.J.; Bazán, G.X.; Espinosa-Salinas, I.; Fierrez, J.; Ortega-Garcia, J.; de Santa Pau, E.C.; Morales, A. Are vision-language models ready for dietary assessment? Exploring the next frontier in AI-powered food image recognition. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 430–439.
38. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. Dinov2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193.
39. Bianco, S.; Buzzelli, M.; Ciocca, G.; Piccoli, F.; Schettini, R. A study on the generalization of DINOv2 features for food recognition tasks: A unified evaluation framework. *Intell. Syst. Appl.* **2026**, *29*, 200632. [CrossRef]
40. Tanabe, H.; Yanai, K. Calorievol: Integrating volumetric context into multimodal large language models for image-based calorie estimation. In *Proceedings of the International Conference on Multimedia Modeling*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 353–365.
41. Penrose, R. A generalized inverse for matrices. In *Proceedings of the Mathematical Proceedings of the Cambridge Philosophical Society*; Cambridge University Press: Cambridge, UK, 1955; Volume 51, pp. 406–413.
42. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
43. Li, G.; Yun, I.; Kim, J.; Kim, J. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. *arXiv* **2019**, arXiv:1907.11357.
44. Piccoli, F.; Buzzelli, M.; Marelli, D.; Bianco, S.; Ciocca, G.; Schettini, R. Efficient Deep Learning Methods for Food Localization in Canteen Trays. In Proceedings of the 2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI), Milano, Italy, 18–20 September 2024; IEEE: New York, NY, USA, 2024; pp. 208–213.
45. Fu, S.; Hamilton, M.; Brandt, L.E.; Feldmann, A.; Zhang, Z.; Freeman, W.T. FeatUp: A Model-Agnostic Framework for Features at Any Resolution. In Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
46. Buzzelli, M.; Moronta-Montero, F.; Fernández-Gualda, R.; López-Baldomero, A.B.; Nieves, J.L.; Valero, E.M. Handwritten ink segmentation algorithms for hyperspectral images of historical documents. *Multimed. Tools Appl.* **2025**, *84*, 39551–39575. [CrossRef]
47. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *Proceedings of the Computer Vision–ECCV 2024*; Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G., Eds.; Springer: Cham, Switzerland, 2025; pp. 38–55.
48. Medeiros, L. Language Segment-Anything. 2023. Available online: https://github.com/luca-medeiros/lang-segment-anything (accessed on 26 May 2026).
49. Bianco, S.; Buzzelli, M.; Chiriaco, G.; Napoletano, P.; Piccoli, F. Food recognition with visual transformers. In Proceedings of the 2023 IEEE 13th International Conference on Consumer Electronics-Berlin (ICCE-Berlin), Berlin, Germany, 3–5 September 2023; IEEE: New York, NY, USA, 2023; pp. 82–87.
50. Meyers, A.; Johnston, N.; Rathod, V.; Korattikara, A.; Gorban, A.; Silberman, N.; Guadarrama, S.; Papandreou, G.; Huang, J.; Murphy, K.P. Im2Calories: Towards an automated mobile vision food diary. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1233–1241.