iLog 3.0 - Estimating Food Volume from 2D Images Using Mask R-CNN and Monocular Depth Estimation

Indira Devi Siripurapu
Dept. of Computer Science and Engineering
University of North Texas, USA
Email: indiradevisiripurapu@my.unt.edu

Alakananda Mitra
Nebraska Water Center
University of Nebraska-Lincoln, USA
Email: amitra6@unl.edu

Saraju P. Mohanty
Dept. of Computer Science and Engineering
University of North Texas, USA
Email: saraju.mohanty@unt.edu

Elias Kougianos
Dept. of Electrical Engineering
University of North Texas, USA
Email: elias.kougianos@unt.edu

***Abstract*—In order to have a healthy and balanced lifestyle, one must consume food in balanced proportions with respect to their requirements. Proportions of the food consumed need to be calculated to find the exact calorie intake or to keep a log. The proposed system, iLog 3.0, automatically determines the volume/quantity of the food item when uploaded using a mobile application, with state-of-the-art object detection and depth estimation techniques when a 2D RGB image is uploaded. The food item will be identified/detected using the Mask R-CNN technique, and to determine the height of the food item from an image MiDaS technique is used to generate the depth map and determine height from it. A high success rate has been achieved, and quantification is accurate compared to the previously used models.**

***Index Terms*—Food Volume Estimation, Smart Healthcare; Healthcare, Volume Detection, Quantification, Mask R-CNN, MiDaS, Dietary Tracking, and Image Processing**

I. Introduction

Food plays an essential role in human health and well-being, and consuming the right measured quantity of food is as important as consuming healthy food. Healthy food in higher quantities is not considered healthy just because the food ingredients are good; in the same way, eating less but eating junk food or processed food is not considered good for human health. Eating without measuring or calculating how much food one needs to consume in a day can cause either over-eating if the individual consumes too much or under-eating if one consumes less than what the body requires. Overeating can lead to obesity, cardiovascular diseases, and metabolic disorders, while under-eating may result in malnutrition, weakened immunity, and chronic health conditions. [1]

Overeating or undereating is not always just eating more food or less food, it can also be eating not according to your body's satiety. Satiety is the full or satisfied feeling related to the stomach after finishing a meal. A meal eaten in a stressful situation or in a hurry might be much less compared to the calorie intake needed, but that meal eaten in a hurry exhibits a satisfied or full feeling. But eating the correct amount with a slower eating rate is much more beneficial than eating less food with a faster eating rate [2].

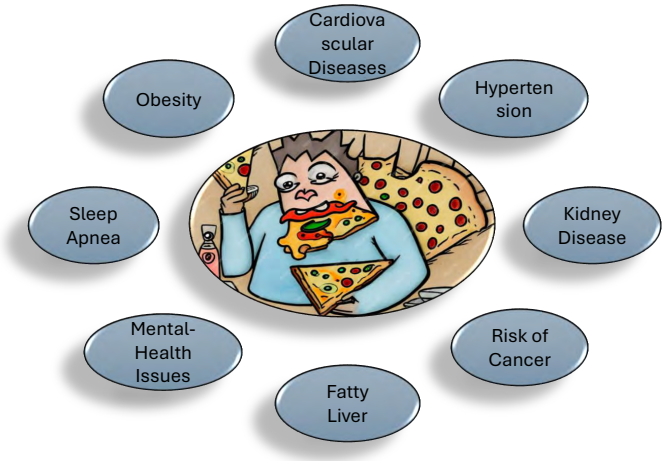

Fig. 1. Effects of Overeating

There are many studies and surveys done on this particular topic of how much an individual should consume in order to stay healthy and fit. In [3], a review paper, the authors focus on the need for combining nutritional sciences with the behavior of the user to reduce and work on obesity and other metabolic disorders.

The study [4] determines the connections between body weight and with effects of dietary habits and physical activity. Multiple combinations of experiments were made with ad-

vanced statistical models, and it was observed that even low to moderate physical activity significantly offsets the impact of high-caloric diets, thus proving that an active lifestyle is important.

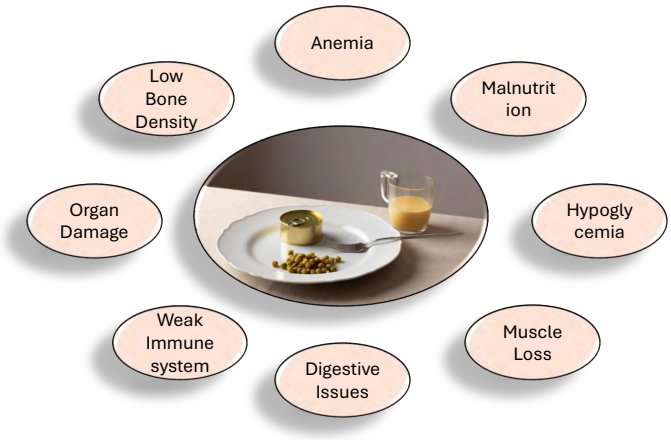

Fig. 2. Effects of Undereating

Paper [5] provides an entire roadmap for addressing food addiction. This paper introduces a computational model for assessing the overeating behavior of an individual by using and combining multiple factors, reinforcement learning, neuroscience, and psychology.

iLog [6] and iLog 2.0 [7] calculate the food quantity automatically and give information about the nutrients present in the food being consumed. iLog 2.0 uses an algorithm that uses a credit card beside the food as a reference object which is not a very convenient way to upload an image in every case scenario and the next algorithm is to find the depth or height of the food in an image using a preset value which was calculated only with limited set of food items.

This paper, iLog 3.0, is an extension to iLog 2.0, covering the gaps in those algorithms. This paper's proposed method does not calculate the nutritional values of the food images uploaded by the user but instead calculates the exact volume of the food in the images using Mask R-CNN trained on polygonal annotated images and uses monocular depth estimation techniques which is also the state of the art technique to find out the depth/height of each food images by converting them into depth maps. Then multiply the values of pixel area from the images by the depth calculated to get the exact volume of the image.

II. Novelty of the Proposed Solution

The novelties of the proposed work of iLog 3.0 are:

1) The quantification system is entirely automated. All it needs is a 2-dimensional RGB image as the user input.
2) Manual input regarding the details of the food item is not required.
3) The image of the food can be uploaded from any angle or direction, the system will figure out the exact height of the food item from the plate.
4) It can be accessed through a mobile application
5) No reference image or reference object is required to find the volume of the food item.
6) It calculates precise volume, making it efficient to log intake.
7) The dataset focuses on fewer items in breakfast items and two different world cuisines for accurate training purposes.
8) The proposed system uses a polygonally annotated training dataset to reduce space issues from bounding box annotations.
9) This proposed system can also be employed on edge devices for more precision.

III. Related Work

With advancements in artificial intelligence, machine learning, and ICT, technology has reached new application levels. Smart city [8] and smart agriculture [9] models are now implemented in daily life.

There are various existing methods for food detection, recognition, calorie counting, and monocular depth estimation in food logging applications. Some rely on user inputs, while others estimate volumes more accurately. Recent methods use deep learning but often involve complex volume calculations or preset values. The proposed system combines multiple models for improved volume estimation. The core of the proposed system is Monocular Depth Estimation. Research [10] validates monocular depth estimation in computer vision and 3D reconstruction using multi-scale Laplacian pyramid fusion frameworks, employing models like MiDaS and DPT to refine depth maps while preserving details.

To select the depth estimation model, [11] compares MiDaS convolutional neural networks and dense transformers. Although dense transformers outperform MiDaS in quantitative metrics like RMSE and mAP by about 12%, MiDaS is more suitable for this proposed system based on other factors. Study [12] compares six deep learning models across three datasets for food detection. Faster R-CNN (MobileNet-V3) achieved 93.1% mAP on the School Lunch Dataset, while YOLOv5 showed strong real-time performance with 77.4% mAP on UEC FOOD 100 and 70.1% on UEC FOOD 256, making Faster R-CNN ideal for precision and YOLOv5 ideal for real-time processing.

The survey paper [13] categorizes depth estimation into active and passive methods, focusing on machine learning-driven monocular approaches. Similarly, the review [14] explores deep learning-based food detection, emphasizing CNNs, vision transformers, and the need for diverse datasets. Paper [15] benchmarks YOLOv5 and EfficientDet for multi-class

food detection, showing YOLOv5 achieving 4% higher mAP and faster inference on datasets like UNIMIB2016 and UEC-Food256.

The study MIDAS [16] supports using monocular depth estimation, demonstrating that multi-scale residual Laplacian refinement improves depth accuracy by 15% on datasets like NYU Depth V2. Research [17] uses a CNN-based approach similar to iLog [6] and iLog 2.0 [7] for calorie estimation, focusing on Thai cuisine and using feature extraction for food item recognition.

The study [18] applies Mask R-CNN for food segmentation and volume estimation using RGB images from a monocular camera, achieving 85.43% accuracy at 0.5 IoU, although it currently works on a single food item. Study [19] introduces the largest food recognition dataset with 1,036,564 images across 2000 categories, using deep progressive region enhancement networks (ResNet-50, ResNet-101, and EfficientNet) trained for fine-grained, ingredient-level food recognition.

IV. METHODOLOGY OF iLOG 3.0

A. Overview of the Proposed System

The system-level overview of the proposed system is shown in Fig.3, The mobile camera works as the end platform of the system, where the user takes a picture of the food platter and then uploads the 2D RGB image into the edge platform. This data, which is in the edge platform, goes through the iLog 3.0 server and passes through the multi-model method, which is proposed in this paper. The information from that model, after calculating the volume of the food item, sends back to the user and stored the data in the server.

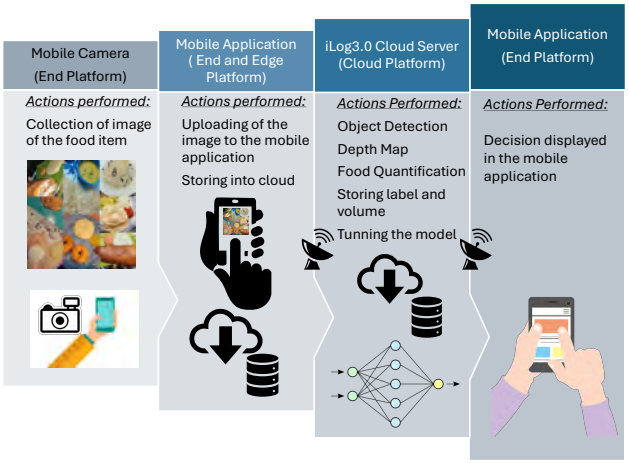

Fig. 3. System Overview

B. Dataset Collection

For this study, we have considered a main class, which is divided into 2 sub-classes. The main class for this dataset is Breakfast foods, the sub-classes are American breakfast cuisine and Indian breakfast cuisine. Each cuisine set has Vegetarian and Non-vegetarian options to cover the two main classes of the world's eating division, but does not go into the depth like eggetarian, vegan, only red meat, or any such classes. It only consists of images that come under 2 categories: Vegetarian and Non-Vegetarian. All the images that were used to train, test, and validate the model were collected by the authors, either making or ordering the food items from multiple food chains, no images were taken from the internet or any copyrighted sources. The total images collected were ∼900, where a major portion was used for training the model, and the other were used for testing and validation of the model. The images were annotated using 50 labels of food items, including sides, fruits, and dips, and were annotated using the polygonal annotation method for accurate training purposes.

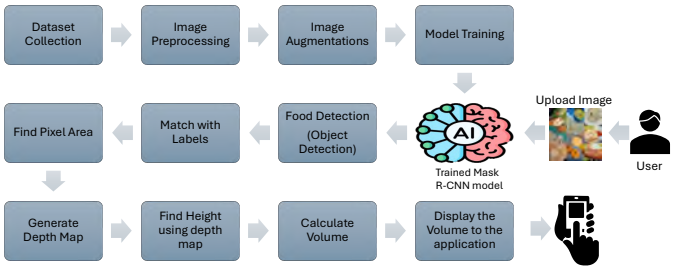

Fig. 4. Development Workflow

C. Data Preprocessing

Each image in the dataset was preprocessed to ensure consistent dimensions and quality for the input, which were given to the Mask R-CNN and MiDaS models for further steps to be performed. All images were resized to a standard dimension and pixels, which allows for all the inputs to be of uniform size. Normalization was done to scale the pixel values between 0 and 1, which improves the consistency of the model and robustness during the inference.

D. Segmentation with Mask R-CNN

Mask R-CNN, the state-of-the-art technology [18], is employed to detect and segment individual food items within each image that is uploaded to the application by the user. The model was trained on the COCO.json file of the dataset, which was generated after all the images were annotated with the given set of labels.

Fig.6 Shows the Mask R-CNN model outputs segmentation masks for each item detected in the image uploaded, which

Fig. 5. Input by the user

then allows for the calculation of the pixel-based area for every food item on the plate of the image. These masks segment and provide a clear boundary around each object, and that is one of the most critical steps in the process.

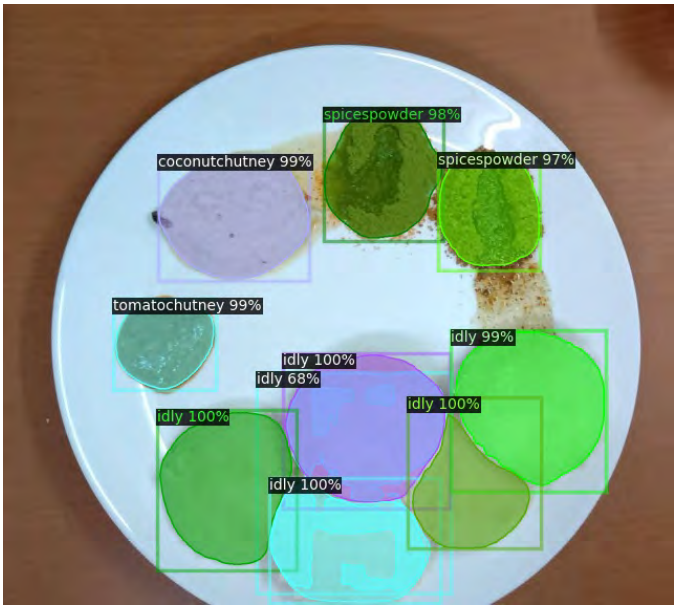

Fig. 6. Mask R-CNN model after identifying food items

E. Monocular Depth Estimation Using MiDaS

After segmenting and calculating the pixel area of each food item in the image, the model proceeds to calculate the height of each food item. To estimate the height of each food item, MiDaS, a state-of-the-art model for monocular depth estimation, is used. MiDaS [10] generates a depth map for each image, where each pixel's intensity value represents the relative depth of the item, i.e., the distance from the camera and the height from the plate. For each of the segmented food items, the segmentation mask is applied to the depth map to extract the depth values that are specific to that particular item. The average depth value of each food item is computed to give the approximate height of the food item.

This method of extracting depth information allows us to get the approximate height of the food item without requiring any additional hardware references, preset values, or multiple angles of the images. Just a simple RGB image of the plate of food items is sufficient.

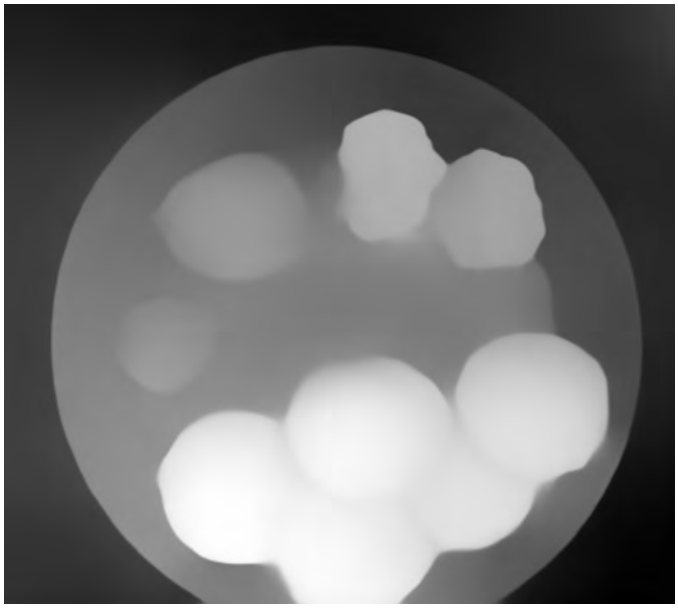

Fig. 7. Depth Map generated for the input given

F. Volume Calculation

The volume of each food item will be estimated by combining the pixel-segmented area of the food item obtained from the Mask R-CNN model's output and the height information from the depth map generated from the MiDaS model.

The formula for volume is:

$$\text{Volume} = \text{Area (cm}^2\text{)} \times \text{Height (cm)} \quad (1)$$

To convert the pixel area into the real-world area, a scaling factor based on the image dimensions and typical real-world sizes is applied. This conversion gives the estimated volume in cubic centimeters (cm^3). Additionally, the calculated volume for every segment can be converted into grams using the densities that are specific to the food type, which can be implemented in dietary use and portion control.

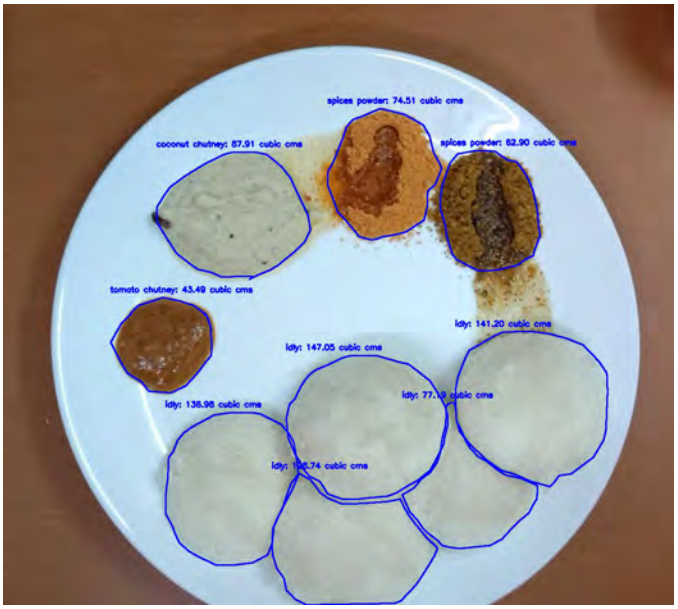

Fig. 8. Volume calculated and displayed

V. Results

The proposed segmentation model is based on Mask R-CNN and is much more efficient than the current methods for a food instance, as it was trained using Detectron2, a robust object detection and segmentation framework built on PyTorch 2.0, on a dataset of 900 where 600 were annotated images covering 50 distinct food categories and the rest were used for testing and validation. The training was conducted on Google Colab, which utilized an NVIDIA A100 GPU with CUDA 11.8 acceleration to optimize both training and inference. Timm and OpenCV were used to enhance efficiency in image preprocessing and model optimization.

The training involved 3000 iterations, utilizing a ResNet-50 FPN backbone, 128 ROI proposals per image, and Stochastic Gradient Descent (SGD) with Momentum as the optimizer. A cyclic learning rate schedule, beginning from 0.00025, was employed to ensure consistent convergence and prevent overfitting.

As shown in Table 1, the model achieves an mAP of 72.1% for bounding box detection and 70.3% for segmentation, which is significantly higher than previous studies that typically show mAP values between 60-68%. Additionally, the model achieves an AP50 accuracy of 85.3% for bounding boxes and 83.9% for segmentation, which is significantly better than conventional methods that typically show AP50 values below 80%.

The approach has a success rate at IoU = 0.75 (AP75), with an accuracy of 78.5 % and 76.8 %, respectively, indicating improved localization accuracy and robustness in detecting overlapping food items. The model's optimized hyperparameters, larger proposal batch size, and advanced learning rate scheduling strategy have led to its superior performance in generalizing across a range of food groups. Figure 9, displays the classification loss, localization loss, total loss, and learning rate progression, demonstrating rapid convergence and stable training dynamics.

Metric	Proposed Model	iLog 2.0
Bounding Box (BB) Accuracy (%)		
mAP	**72.1**	60–68
AP_{50}	**85.3**	80
AP_{75}	**78.5**	–
Segmentation Accuracy (%)		
mAP	**70.3**	60–68
AP_{50}	**83.9**	80
AP_{75}	**76.8**	–

TABLE I
Comparison between the proposed model and iLog 2.0

The first 500 iterations show sharp decreases in both classification loss (Figure 9a) and total loss (Figure 9c), indicating that feature extraction and model learning are relatively stable. This is in contrast to previous methods that show a sharp decline in localization loss (Figure 9b), which is significantly lower than previously observed, leading to a more stable reduction.

The cyclic learning rate strategy illustrated in Figure 9d prevents overfitting and leads to faster convergence, increasing model performance. These results demonstrate that our method achieves the highest level of accuracy reported for food segmentation tasks, a new standard for automated food volume estimation applications.

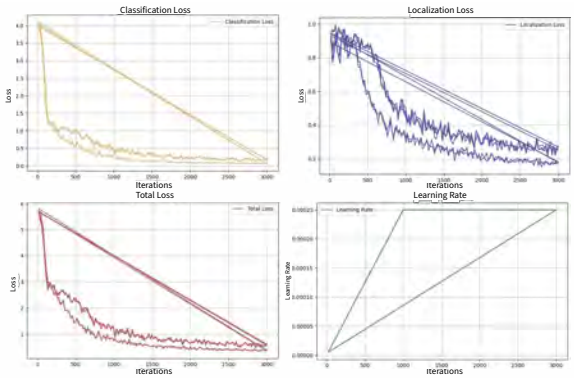

Fig. 9. Validation Graph

The superior performance of our model is attributed to its optimized hyperparameters, increased proposal batch size, and advanced learning rate scheduling strategy, ensuring better generalization across diverse food types. Figure 9 illustrates the classification loss, localization loss, total loss, and learning rate progression, confirming rapid convergence and stable training dynamics. The classification loss (Figure 9a) and total loss (Figure 9c) decrease sharply within the first 500 iterations, signifying effective feature extraction and model

learning stability. Unlike previous methods, where the localization loss (Figure 9b) fluctuates significantly, our approach demonstrates a smoother and more stable reduction, resulting in better bounding box placement and mask accuracy. The cyclic learning rate strategy (Figure 9d) plays a crucial role in preventing overfitting while accelerating convergence, further enhancing model performance. These results confirm that our method achieves the highest accuracy reported for food segmentation tasks, setting a new benchmark for automated food volume estimation applications.

The model has a few limitations as it assumes a relatively uniform height within each segmented region, which may lead to minor inaccuracies for foods with varied shapes or uneven surfaces. Additionally, depth estimation can be affected by shadows and lighting variations, which, while minimized in this setup, could be further improved with refined preprocessing techniques.

VI. Conclusion

This paper presents a practical approach for food volume estimation from a single 2D image, leveraging Mask R-CNN for segmentation and MiDaS for depth estimation. Our solution offers a significant improvement in accuracy over traditional 2D approaches and simplifies hardware requirements by eliminating the need for multi-angle or 3D imaging setups. This approach has potential applications in dietary tracking, portion control, and health monitoring. Future work will focus on improving depth estimation for irregularly shaped foods and testing the model across larger and more diverse datasets.

References

[1] M. Pakhare and A. Anjankar, "Critical correlation between obesity and cardiovascular diseases and recent advancements in obesity," *Cureus*, vol. 16, no. 1, p. e51681, 2024. PMID: 38314003; PMCID: PMC10838385.

[2] G. Argyrakopoulou, S. Simati, G. Dimitriadis, and A. Kokkinos, "How important is eating rate in the physiological response to food intake, control of body weight, and glycemia?," *Nutrients*, vol. 12, p. 1734, Jun. 2020.

[3] R. R. Subramanian, M. Kancharla, S. H. Duddekula, A. Harshith, G. S. Kamisetty, and R. R. Sudharsan, "Assessing and monitoring dietary intake," in *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, pp. 1–4, 2021.

[4] T. Prioleau, Y. Heng, A. Veeraraghavan, and A. Sabharwal, "Exploring the effect of food intake and physical activity on body weight," in *2017 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 165–168, 2017.

[5] K. S. Burger, "Food reinforcement architecture: A framework for impulsive and compulsive overeating and food abuse," *Obesity (Silver Spring)*, vol. 31, pp. 1734–1744, Jul. 2023.

[6] L. Rachakonda, S. P. Mohanty, and E. Kougianos, "ilog: An intelligent device for automatic food intake monitoring and stress detection in the iomt," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 2, pp. 115–124, 2020.

[7] A. Mitra, S. Goel, S. P. Mohanty, E. Kougianos, and L. Rachakonda, "ilog 2.0: A novel method for food nutritional value automatic quantification in smart healthcare," in *2022 IEEE International Symposium on Smart Electronic Systems (iSES)*, pp. 683–688, 2022.

[8] S. P. Mohanty, U. Choppali, and E. Kougianos, "Everything you wanted to know about smart cities: The internet of things is the backbone," *IEEE Consumer Electronics Magazine*, vol. 5, no. 3, pp. 60–70, 2016.

[9] A. Mitra, S. L. Vangipuram, A. K. Bapatla, V. K. Bathalapalli, S. P. Mohanty, E. Kougianos, and C. Ray, "Smart agriculture: A comprehensive overview," *SN Computer Science*, vol. 5, no. 8, p. 969, 2024.

[10] J.-H. Lee and C.-S. Kim, "Monocular depth estimation using relative depth maps," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9721–9730, 2019.

[11] S. Howells and O. Abuomar, "Depth maps comparisons from monocular images by midas convolutional neural networks and dense prediction transformers," in *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pp. 1–6, 2022.

[12] S. W. Tan, C. P. Lee, K. M. Lim, and J. Y. Lim, "Food detection and recognition with deep learning: A comparative study," in *2023 11th International Conference on Information and Communication Technology (ICoICT)*, pp. 283–288, 2023.

[13] N. Jamwal, N. Jindal, and K. Singh, "A survey on depth map estimation strategies," in *2016 International Conference on Signal Processing (ICSP)*, pp. 1–5, 2016.

[14] A. Banerjee, P. Bansal, and K. Thomas, "Food detection and recognition using deep learning – a review," in *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pp. 1221–1225, 2022.

[15] R. Morales, J. Quispe, and E. Aguilar, "Exploring multi-food detection using deep learning-based algorithms," in *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*, pp. 1–7, 2023.

[16] A. Zhang, Y. Ma, J. Liu, and J. Sun, "Promoting monocular depth estimation by multi-scale residual laplacian pyramid fusion," *IEEE Signal Processing Letters*, vol. 30, pp. 205–209, 2023.

[17] R. Sombutkaew and O. Chitsobhuk, "Image-based thai food recognition and calorie estimation using machine learning techniques," in *2023 20th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 1–4, 2023.

[18] Y. Dai, S. Park, and K. Lee, "Utilizing mask r-cnn for solid-volume food instance segmentation and calorie estimation," *Applied Sciences*, vol. 12, no. 21, p. 10938, 2022.

[19] W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, and S. Jiang, "Large scale visual food recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9932–9949, 2023.