Seeing Through the Plate: Estimating Nutrition from Multiview Food Images

1st Andrés Lomelí
AI
Universidad Panamericana
Aguascalientes, Mexico
0242956@up.edu.mx

***Abstract*—Accurate nutritional estimation from food images remains a challenging task due to occlusions, lack of depth cues, and the high variability of food presentation. In this paper, we propose a novel, sensor-free, multi-view framework for predicting detailed macronutrient values—calories, carbohydrates, proteins, fats, and weight—from RGB images alone.**

***Index Terms*—Classification, Carbohydrates, Detector, food**

I. Introduction

Accurate dietary assessment plays a crucial role in addressing global health issues such as obesity, diabetes, and cardiovascular diseases. Manual tracking of food intake is time-consuming, error-prone, and highly subjective, often leading to underreporting or misestimation of caloric and nutritional values. With the ubiquity of smartphones and advancements in computer vision, automated food analysis systems based on images have emerged as promising tools to support healthier lifestyles and dietary self-monitoring.

Estimating the nutritional content of a meal from an image, however, is a highly challenging task. Visual information alone is often insufficient to determine the exact type, portion size, or composition of a dish. Different ingredients can appear similar, and the same dish may vary significantly across cultures, restaurants, or individual preparations. Moreover, the volume of food—crucial for calorie estimation—is difficult to infer from a single 2D image without depth cues or scale references.

To overcome these challenges, researchers have developed a wide range of approaches combining deep learning, computer vision, and nutritional databases. Some models classify food types and infer average values from known datasets, while others aim to estimate portion size through geometric reasoning or depth sensors. A more recent and promising direction involves directly regressing caloric and macronutrient values from food images using end-to-end deep learning models.

This paper presents a review of the state of the art in food image-based nutrition estimation, comparing various methodologies and highlighting their strengths and limitations. We also propose a novel multi-view transformer-based framework that leverages multiple RGB images per dish to improve prediction accuracy without requiring depth sensors or manual calibration

II. State of the art

The automatic estimation of nutritional content from food images has become a significant area of research, driven by the growing need for tools that assist individuals in monitoring their dietary intake. This section explores the evolution of methodologies in this domain, focusing on food recognition, portion size estimation, and nutritional analysis, while addressing the challenges and limitations inherent in these approaches.

A. Food Recognition

Accurate food recognition serves as the foundation for any image-based nutritional estimation system. Early approaches relied on traditional image processing techniques, but the advent of deep learning, particularly Convolutional Neural Networks (CNNs), has significantly enhanced recognition accuracy.

1) Convolutional Neural Networks (CNNs) for Food Classification: CNNs have demonstrated remarkable success in image classification tasks, including food recognition. Models such as AlexNet, VGGNet, and ResNet have been employed to classify food items from images. For instance, the work by Ciocca et al. provides a comprehensive overview of deep learning techniques applied to food recognition, emphasizing the effectiveness of CNNs in handling the variability of food appearances [1]. However, food classification presents unique challenges:

- Intra-Class Variability: The same food item can appear differently due to variations in cooking methods, presentation, and lighting conditions.
- Inter-Class Similarity: Visually similar foods may belong to different categories, making differentiation difficult.

To address these challenges, hybrid models combining CNNs with other techniques have been proposed. Min et al. introduced a hybrid deep learning algorithm that integrates multiple CNN architectures to improve food recognition accuracy, demonstrating the potential of ensemble methods in this context [2].

B. Portion Size and Volume Estimation

Estimating the portion size or volume of food items is crucial for accurate nutritional assessment. Various methodologies have been explored, ranging from geometric models to advanced deep learning techniques.

1) Depth-Based Methods: Incorporating depth information can enhance volume estimation accuracy. Pouladzadeh et al. developed a system utilizing RGB-D images (combining color and depth data) to estimate food volume, leveraging the additional spatial information provided by depth sensors [3]. Similarly, Lee and Kwon proposed a method that combines color and depth images to estimate food intake amounts, demonstrating improved accuracy over color-only approaches [4].

However, the reliance on depth sensors poses limitations:

- Hardware Requirements: Depth sensors are not commonly available in standard consumer devices, restricting the applicability of these methods.
- Environmental Constraints: Depth sensing can be affected by lighting conditions and the reflective properties of food surfaces.

2) RGB Image-Based Methods: To overcome hardware limitations, researchers have explored volume estimation using only RGB images. Shao et al. introduced a framework that reconstructs 3D shapes from monocular images to estimate food portions, employing domain adaptation techniques to enhance accuracy [5]. Similarly, Han et al. proposed DPF-Nutrition, an end-to-end nutrition estimation method that predicts depth maps from monocular images and fuses them with RGB data to improve portion estimation [6].

While these methods eliminate the need for specialized hardware, challenges remain:

- Depth Ambiguity: Inferring depth from a single RGB image is inherently ambiguous and prone to errors.
- Occlusions and Complex Food Arrangements: Overlapping food items and complex presentations can complicate volume estimation.

C. Nutritional Content Estimation

Beyond recognition and volume estimation, determining the nutritional content—calories and macronutrients—of food items is the ultimate goal. This involves correlating visual information with nutritional databases or directly predicting nutritional values.

1) Direct Prediction Models: Thames et al. introduced the Nutrition5k dataset, comprising over 5,000 real-world dishes with detailed nutritional annotations, including calories, fats, proteins, and carbohydrates [7]. Using this dataset, they trained deep learning models to predict nutritional content directly from images, achieving accuracy levels surpassing those of professional nutritionists.

However, direct prediction models face several challenges:

- Dataset Limitations: Large, diverse, and accurately annotated datasets are essential for training robust models, but such datasets are scarce.
- Generalization: Models trained on specific datasets may not generalize well to foods from different cultures or regions.

2) Multispectral Imaging: To enhance accuracy, some studies have incorporated multispectral imaging. Lee et al. utilized images captured at various wavelengths, including ultraviolet and near-infrared, to improve food classification and caloric estimation, demonstrating that multispectral data can provide additional discriminative features [8]. Despite its potential, multispectral imaging is limited by:

- Equipment Accessibility: Specialized cameras are required, which are not commonly available to consumers.
- Complexity: Processing multispectral data increases computational complexity and may not be feasible for real-time applications.

D. In-depth Review: Im2Calories

A seminal work in the field of nutritional estimation from images is Im2Calories by Meyers et al. [9] [14], which presents a comprehensive vision-based framework aimed at building an automated dietary logging tool. The goal of their system is to estimate the caloric content of food using mobile camera images, eliminating manual tracking.

Method Overview The Im2Calories pipeline consists of multiple stages designed to address the complex task of estimating calories from a single RGB image:

1) Food Detection and Segmentation: The system begins by segmenting the food from the background using a deep convolutional neural network. Instance segmentation is used to identify and isolate individual food items on a plate.
2) Food Classification: Each segmented region is classified into a predefined food category using a CNN trained on a food-specific dataset. Fine-grained classification is crucial here, as misclassification can significantly affect calorie estimates.
3) 3D Shape Estimation: A novel aspect of the Im2Calories framework is its depth estimation from RGB. Instead of using dedicated hardware, the authors train a CNN to predict depth maps from single RGB images. These depth maps are then used to infer 3D volume of the food.
4) Calorie Estimation: With known volume and class label, the system queries a food density database to convert volume into mass, and then into calories, using standardized nutritional values.

E. Challenges and Limitations

Despite advancements, several challenges persist in the field of image-based nutritional estimation:

- Reliance on Specialized Hardware: Methods requiring depth sensors or multispectral cameras are impractical for widespread use due to hardware limitations.
- Data Availability and Quality: The scarcity of large-scale, diverse, and accurately annotated datasets hampers model training and evaluation.
- Generalization Across Diverse Cuisines: Models often struggle to generalize to foods from different cultures, limiting their applicability in

Technical Challenges and Research Implications Meyers et al [9]. explicitly identify numerous sub-problems that their system must address—many of which remain open today:

- Fine-grained recognition: Distinguishing between visually similar variants (e.g., baked vs. fried chicken).
- Open-world classification: Recognizing foods outside the training set.
- Visual attribute detection: Understanding food preparation methods (with/without sauce, grilled/fried).
- Real-time on-device inference: Making the system usable in mobile devices with limited resources.

These problems are not only important for practical deployment but are also of significant research interest across computer vision and human-computer interaction fields.

III. Proposed Method

We propose a two-branch pipeline for nutrition estimation that fuses predictions from two complementary sources: a multiview CNN-based regression model and a semantic segmentation model combined with ingredient metadata. This design avoids the need for depth sensors or 3D modeling and enables interpretable, compositional estimates of food macronutrients.

A. Overview of the Fusion Framework

Our method consists of two main components:

1) **Multiview EfficientNetB4 Regressor:** A deep CNN trained on multiple views of a meal to directly regress calories, protein, fat, and carbohydrates.
2) **Mask2Former-Based Segmentation + Metadata Fusion:** A semantic segmentation branch that identifies visible food ingredients in one view using a pretrained Mask2Former model. Each segment is mapped to nutrient values (per gram) using Nutrition5k ingredient metadata. These values are combined with estimated area proportions to infer a second prediction.

Finally, both predictions are fused using a weighted average (learned or fixed), producing a final nutritional estimate. This approach balances learned global priors (EffNet) with interpretable, visible evidence (segmentation + metadata).

B. Multiview EfficientNetB4 Regressor

Following the idea of direct regression, we trained an EfficientNetB4 model adapted to accept N image views of a meal. Each image is processed individually, and the resulting features are aggregated using average pooling across views. The pooled representation is passed through four fully connected layers to predict total calories, protein, fat, and carbohydrate content.

This model is trained using the Nutrition5k dataset, using 2–4 randomly selected views per dish and MSE loss for each macronutrient.

C. Mask2Former-Based Segmentation

For the segmentation branch, we fine-tune a Mask2Former model on the FoodSeg103 dataset to obtain pixel-level segmentations of ingredient categories. We apply this model to the first image view of each meal.

Each segment is assigned a class label (e.g., tomato, lettuce, rice), and we use the Nutrition5k ingredient metadata to assign average nutrient values (per gram) to each class. Using the relative area of each segment as a proxy for mass proportion, we compute an ingredient-level estimate of the full meal's macronutrient profile.

D. Fusion Strategy

The outputs from both branches are fused using a weighted average. Let y_{eff} be the output of the EfficientNet regressor and y_{seg} be the output of the segmentation-based method. The final prediction is:

$$y_{\text{final}} = \alpha \cdot y_{\text{eff}} + (1 - \alpha) \cdot y_{\text{seg}}$$

Where α is a tunable parameter (e.g., 0.6) controlling the trust in each branch. Future work could explore learning this fusion dynamically.

E. Pipeline Diagram

F. Comparison with Transformer-Based Design (Prior Work)

In our early design iterations, we explored a transformer-based multiview model that utilized a cross-view attention encoder to combine visual features extracted from multiple RGB views of a dish. Each image was first pre-processed using the Segment Anything Model (SAM) to isolate visible food regions. The segmented outputs were then passed through a shared CNN backbone (e.g., EfficientNet) and fused via a transformer encoder that performed inter-view attention.

This approach was conceptually powerful—it allowed for implicit amodal completion (inferring hidden parts of food) and spatial reasoning across views without requiring depth estimation. However, in practice, several limitations emerged:

- **SAM Dependency**: The reliance on SAM added computational overhead and occasional segmentation artifacts, especially on complex or cluttered dishes.
- **Training Instability**: The transformer was sensitive to view order and required careful tuning to generalize across food types.
- **Interpretability Issues**: The end-to-end model made it difficult to trace predictions back to specific ingredients or visual cues.

Based on these challenges, we transitioned to a more interpretable and modular fusion strategy. Our current system consists of two components:

1) A multiview EfficientNetB4 model that directly regresses the nutritional values from 2–4 RGB images.
2) A segmentation pipeline using a fine-tuned Mask2Former that segments visible ingredients and estimates their nutritional contribution via an ingredient metadata database.

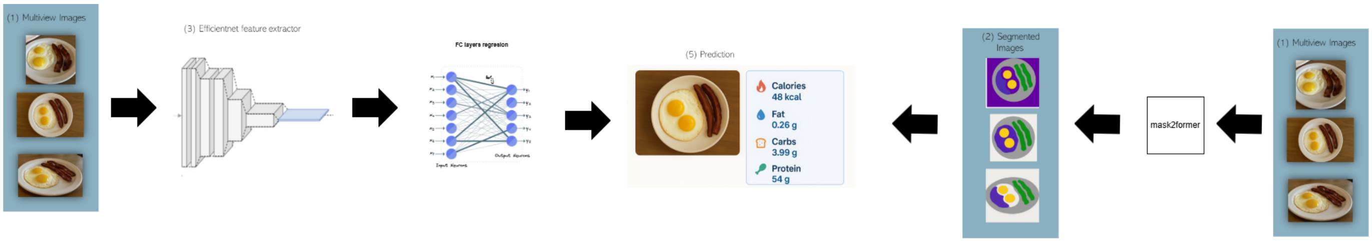

Fig. 1. Overview of the proposed fusion-based nutrition estimation pipeline.

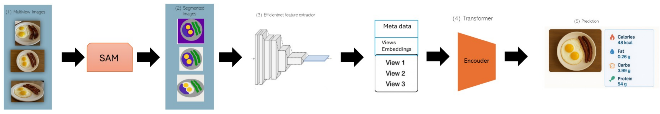

Fig. 2. First Sam-pipeline.

The predictions from both components are combined through a rule-based or learnable fusion scheme, leveraging both visual priors and compositional reasoning.

This revised approach provides several benefits:

- **Scalability**: No need for pre-segmentation during inference.
- **Interpretability**: Ingredient-level segmentation enables nutritional attribution per region.
- **Robustness**: Fusing two independent estimates improves generalization on unseen food compositions.

Overall, the fusion-based approach outperforms the original transformer pipeline in both accuracy and usability, while reducing architectural complexity and inference time.

The overall pipeline is illustrated in Figure 1. The segmentation module uses a fine-tuned Mask2Former to produce instance masks of food items, which are then matched to an ingredient-nutrient database. Simultaneously, the EffNet model aggregates multiview context. Their predictions are fused via a weighted sum.

IV. Experiments and Results

We evaluated our models using the Nutrition5k dataset, with separate splits for training and validation. Performance is measured using both Mean Squared Error (MSE) and Mean Absolute Error (MAE) across five nutritional targets: calories, carbohydrates, protein, fat, and total weight (grams).

A. Baseline Comparison: EfficientNet Variants

We first trained two variants of EfficientNet to directly regress macronutrients from multi-view RGB images. The models received either 3 or 4 views per dish.

- **EfficientNetB0 (3 views):**
  - Calories: MAE = 154.20
  - Carbs: MAE = 13.41
  - Protein: MAE = 11.32
  - Fat: MAE = 10.63
  - Weight: MAE = 100.64
- **EfficientNetB4 (4 views):**
  - Calories: MAE = 87.08
  - Carbs: MAE = 9.65
  - Protein: MAE = 7.74
  - Fat: MAE = 7.09
  - Weight: MAE = 68.25

B. Transformer-Based Variant (with SAM Preprocessing)

We next tested a more complex pipeline based on a cross-view transformer encoder and SAM (Segment Anything Model) for food region isolation. Each image was first segmented with SAM before feature extraction.

Note: All images had to be pre-segmented using SAM, which was computationally expensive.

- Calories: MAE = 78.13
- Carbs: MAE = 10.68
- Protein: MAE = 6.90
- Fat: MAE = 6.89
- Weight: MAE = 63.86

C. EfficientNetB4 (New Standalone Pipeline)

In our updated and streamlined model—without segmentation or transformer components—we achieve:

- Calories: MAE = 46.32
- Carbs: MAE = 10.68
- Protein: MAE = 6.85
- Fat: MAE = 5.03

D. Fusion Pipeline (EffNet + Mask2Former)

To be computed. Placeholder for full dataset fusion MAEs.

- **Calories: MAE = ____________**
- **Carbs: MAE = ____________**
- **Protein: MAE = ____________**
- **Fat: MAE = ____________**

E. Case Studies

We include two example dishes to illustrate how fusion improves prediction when ingredient segmentation aligns with reality, and how errors can arise when important ingredients are missing from the segmentation mask.

1) Dish ID: dish_1550708327: **Fusion Prediction:**

- Calories: 46.18
- Protein: 2.20
- Fat: 2.54
- Carbs: 4.32

EffNetB4 Prediction:

- Calories: 76.96
- Protein: 3.66
- Fat: 4.23
- Carbs: 7.21

Ground Truth:

- Calories: 67.07
- Protein: 5.51
- Fat: 3.15
- Carbs: 6.17

Note: Only two out of three ingredients were correctly segmented in this example, which may explain the fusion error (further discussed in Section VI).

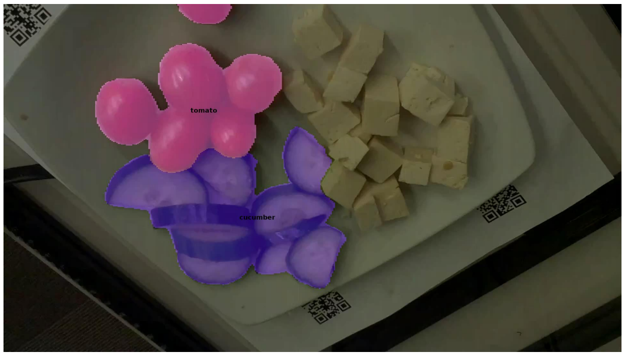

Fig. 3. Segmentation result for dish_1550708327. Two ingredients correctly detected.

2) Dish ID: dish_1550772454: **Fusion Prediction:**

- Calories: 0.88
- Protein: 0.79
- Fat: -0.39
- Carbs: 0.88

EffNetB4 Prediction:

- Calories: 1.09
- Protein: 0.99
- Fat: -0.49
- Carbs: 1.10

Ground Truth:

- Calories: 22.75
- Protein: 1.56
- Fat: 0.26
- Carbs: 4.55

Note: All visible ingredients were captured in the segmentation mask, yet both models underpredict the true caloric value—likely due to low-contrast food or data imbalance.

V. Discussion

Our experiments explored two primary approaches for nutritional estimation: an earlier pipeline using image segmentation with SAM and a cross-view transformer, and a new streamlined design that fuses EfficientNet predictions with ingredient-aware segmentation from Mask2Former. Both approaches offer different insights into the challenges of image-based nutrition prediction.

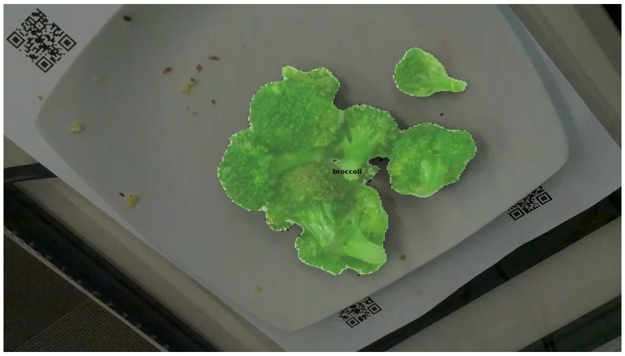

Fig. 4. Segmentation result for dish_1550772454. All ingredients covered.

A. Limitations of SAM-Based Segmentation

In the initial pipeline, we employed the Segment Anything Model (SAM) to isolate food regions prior to prediction. While SAM is effective at producing clean segmentation masks without supervision, we observed limited benefit in our pipeline. The core issue was a mismatch between SAM's generic masks and the ingredient-specific annotations in the Nutrition5k dataset.

Although masks looked visually accurate, the model could not learn meaningful associations between regions and their nutritional impact without ingredient-level supervision. Consequently, even with a transformer-based fusion across views, improvements plateaued—our best MAE came from EfficientNetB4 with SAM masks and a cross-view transformer, but this was only marginally better than simpler baselines.

This suggests that unless segmentation masks are annotated and semantically aligned with ground truth ingredients, they may inject noise into the learning process rather than improve it. The SAM-based setup remains promising if a supervised mask annotation effort is carried out in the future.

B. Mask2Former and Metadata Fusion

Our second pipeline eliminates manual segmentation and uses a trained Mask2Former model for food ingredient detection, combined with a metadata lookup table of per-gram nutritional values. When segmentation is successful, this allows us to estimate nutritional values per ingredient and combine them for a total prediction.

We found that this strategy improves estimates, especially for simpler dishes where all ingredients are part of the FoodSeg103 taxonomy. However, the model is inherently limited by the coverage of its training set—only 103 food categories. Many real dishes include ingredients that are either not detected or misclassified, leading to incomplete predictions. In those cases, we default to using the EfficientNet-only regression.

The fusion system helps improve MAE slightly in some examples, but suffers when segmentations are partial or missing. Case studies such as `dish_1550708327` highlight how missing one of three ingredients can significantly skew results.

C. Why the Transformer Pipeline Didn't Succeed

From an AI standpoint, several reasons may explain why our transformer-based cross-view encoder failed to outperform EfficientNet alone:

- **Weak supervision:** Without strong labels tied to regions or views, the transformer cannot learn reliable correspondences across angles.
- **Insufficient view diversity:** Many dishes have only two or three slightly different perspectives, which limits the transformer's ability to infer geometry.
- **Overhead vs. gain:** The added architectural complexity didn't translate into meaningful accuracy gains—possibly due to overfitting, especially with limited training data.

D. Summary and Insights

Our most effective model remains the simple EfficientNetB4 trained end-to-end with 4 views, achieving a substantial reduction in MAE over prior methods. The fusion model using Mask2Former and ingredient metadata showed promising results, but remains constrained by dataset coverage and the quality of segmentation.

Future directions should include:

- Expanding segmentation models to cover a broader ingredient taxonomy.
- Annotating masks in Nutrition5k to enable true pixel-wise supervision.
- Exploring late fusion strategies where EffNet and segment-based predictions are adaptively combined based on mask confidence.

In conclusion, while multiview RGB imagery enables reasonable nutrient estimation, further gains require better supervision, richer metadata, and perhaps multimodal inputs (e.g., user context or textual recipe).

VI. Conclusion and Future Work

In this work, we investigated two distinct approaches for nutritional estimation from multiview RGB food images: a transformer-based pipeline incorporating semantic segmentation masks and a more practical model leveraging EfficientNet and ingredient-aware fusion using Mask2Former. Through extensive experimentation on the Nutrition5k dataset, we observed that while multiview information improves macronutrient prediction accuracy, the real bottleneck lies in the availability and quality of ingredient-level annotations.

Our findings reveal that even visually accurate segmentation (e.g., from SAM) is insufficient if not semantically aligned with the nutritional metadata. Additionally, while Mask2Former can detect some ingredients effectively, its limited taxonomy (103 classes) restricts its real-world applicability. Nevertheless, combining its outputs with ingredient metadata proved helpful in specific cases, showing potential for more modular and interpretable nutrition prediction.

A. Lessons Learned

- Segmentations must be meaningfully linked to ingredients to improve prediction.
- Pretrained segmentation models help, but coverage is critical—unseen ingredients cause incomplete predictions.
- EfficientNet alone performs surprisingly well, but combining it with ingredient-level logic can further refine outputs.

B. Future Directions

To fully realize the promise of food segmentation-based nutrition estimation, future work should focus on:

- **Building a large-scale dataset** of ingredient-segmented food images, covering hundreds or thousands of common ingredients. This dataset would be the backbone for training segmentation models that understand nutritional composition.
- **Combining tools like SAM** with manual refinement or weak supervision from recipe metadata to create annotated masks at scale.
- **Expanding and fine-tuning segmentation models** beyond FoodSeg103 to capture more realistic and mixed-ingredient meals.
- **Leveraging multimodal input** (e.g., natural language descriptions or menus) to complement visual ambiguity.

Ultimately, while multiview deep learning is a step forward, we conclude that achieving accurate and explainable nutritional estimation will require bridging the gap between visual perception and ingredient-level semantic understanding—powered by better datasets and more holistic models.

References

[1] G. Ciocca, P. Napoletano, and R. Schettini, "Food recognition: A new dataset, experiments, and results," IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 3, pp. 588–598, May 2017.

[2] W. Min, S. Jiang, J. Sang, H. Wang, and L. Herranz, "Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration," IEEE Transactions on Multimedia, vol. 19, no. 5, pp. 1100–1113, May 2017.

[3] P. Pouladzadeh, S. Shirmohammadi, and R. Al-Maghrabi, "Measuring calorie and nutrition from food image," IEEE Transactions on Instrumentation and Measurement, vol. 63, no. 8, pp. 1947–1956, Aug. 2014.

[4] S. Lee and H. Kwon, "A two-stage approach for estimating food intake in mobile diet monitoring," in Proc. IEEE Int. Conf. on Image Processing (ICIP), 2015, pp. 2761–2765.

[5] Y. Shao, T. S. Huang, and Z. Yang, "Food volume estimation using 3D reconstruction from a single-view image," IEEE Transactions on Multimedia, vol. 21, no. 5, pp. 1190–1199, May 2019.

[6] Y. Han, M. Zhu, and S. Wang, "DPF-Nutrition: Depth prediction and fusion for food nutrition estimation," arXiv preprint arXiv:2310.11702, 2023.

[7] M. Thames, A. Spence, T. R. Hayes, and A. Farhadi, "Nutrition5k: Towards automatic nutritional understanding of generic food," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1911–1921.

[8] C. Lee, S. Lin, and H. Wang, "Multispectral imaging for real-time food recognition and caloric estimation," in Proc. ACM Int. Conf. on Multimedia (ACM MM), 2018, pp. 1204–1212.

[9] S. Fang, Y. Zhu, C. Boushey, and E. J. Delp, "Im2Calories: Towards an automated mobile vision food diary," in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2015, pp. 1233–1241.

[10] S. Myers et al., "Im2Calories: Towards an automated mobile vision food diary," in Proc. ICCV, 2015, pp. 1233–1241.

[11] M. Anthimopoulos, J. Dehais, P. Diem, and S. Mougiakakou, “Computer vision-based carbohydrate estimation for type 1 patients with diabetes using deep learning,” Multimedia Tools and Applications, vol. 77, no. 9, pp. 10433–10451, May 2018.

[12] M. Mezgec and B. Koroušić Seljak, “NutriNet: A deep learning food and drink image recognition system for dietary assessment,” Nutrients, vol. 9, no. 7, pp. 657, Jul. 2017.

[13] K. Bolliger, R. He, R. Kaehr, and J. Amft, “Deep learning based food image segmentation for dietary assessment,” in Proc. Int. Conf. on Smart Health, 2019, pp. 1–14.

[14] A. Meyers, S. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, and S. Guadarrama, “Im2Calories: Towards an Automated Mobile Vision Food Diary,” in Proc. ICCV, 2015.

[15] H. Kawano and K. Yanai, “Real-time mobile food recognition system,” in Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 2013, pp. 1–7.

[16] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. on Machine Learning (ICML), 2019, pp. 6105–6114.

[17] Z. Wu, Y. Zhang, W. Min, and S. Jiang, “FoodSeg103: A dataset of 103 annotated food categories for semantic segmentation,” in Proc. ACM Int. Conf. on Multimedia (ACM MM), 2021, pp. 3635–3644.

[18] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, “Masked-attention Mask Transformer for Universal Image Segmentation,” in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1290–1299.

[19] N. Vahdat, “NimaVahdat/foodseg/mask: Ingredient segmentation from food images,” GitHub repository, 2024. [Online]. Available: https://github.com/NimaVahdat/foodseg_mask

[20] M. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, A. Rokas, and R. Girshick, “Segment Anything,” arXiv preprint arXiv:2304.02643, 2023.