SingaFood: Harnessing Deep Learning for Dietary Analysis of Singapore Cuisine

ML Singapore Group 1

Jefferson Lim (A0217512N); Ong Han Wei (A0234074L); Nicholas Lee Jun Yi (A0233400B); Ryan Chua Zong Xun (A0233636E)

Abstract

Knowledge of nutrition levels in food are influential in people's dietary choices and leads to positive health outcomes. In this project, we design and implement four machine learning models for the task of nutritional information prediction from food images of common Singaporean dishes. We propose enhancements that overcome the various challenges arising from a diversity of dishes in Singapore. Our results show that our `BASELINE-V2-TRANSFER` model has the best performance, with an average of 87.6% MAE.

Introduction

Background

In order to maintain optimal health, it is essential to consume the appropriate types and quantities of food, ensuring that the body receives essential nutrients . In Singapore, the Ministry of Health (MOH) has formed the Health Promotion Board (HPB) in 2001 to tackle the problem of healthy living in Singapore.

One major challenge facing our population is the rise of chronic diseases, namely diabetes (hyperglycemia), high blood pressure (hypertension) and high cholesterol (hyperlipidaemia), or more commonly known as 3Hs. The National Health Survey 2022 shows increasing trends for these chronic diseases, and reports that diet control could be an effective control against health complications [5]. Hence, we need an easy and effective way to figure out the nutrition that we are getting from the food we consume, to remain healthy and free from health conditions.

Currently, some form of laboratory testing is necessary to obtain the accurate estimates of the nutrients present in food [6]. However, this method is impractical for day to day purposes. While it is possible to estimate nutrition by tallying up the information on nutrition labels on food packaging, this only applies to food packaged before distribution, and not served food. The HPB has also introduced "My Healthy Plate" – heuristics for an ideal and well-balanced diet, as well as "Nutrigrade" – a label for the amount of sugar and saturated fats in drinks. Although these solutions are easily accessible and usable, they fail to generalize across many food types and to cater towards specific nutrients that affect hypertension and high cholesterol.

Hence, we turn towards machine learning to devise a potential solution for this problem.

Related Work

One technological solution for nutrition tracking is to use apps like MyFitnessPal. These apps require the user to manually identify the types of food, select macronutrients they are interested in, and visually estimate the portion sizes for each food item. This process is tedious, time consuming and error-prone [6]. A computer-vision based approach can streamline this process while ensuring accuracy.

Some previous work uses the approach of first classifying food images into a type of dish (e.g. chicken rice), and retrieving the nutritional breakdown for that dish from a database [9]. One issue is that there may be different preparation techniques for the same dish (e.g. steamed chicken vs. fried chicken in chicken rice) leading to different nutritional outcomes. Thames et al. constructs a novel dataset with a high-accuracy nutritional annotation to train a computer vision model [6], and Wu et al. uses image segmentation along with semantic content from recipe embeddings to augment nutritional estimation [8].

However, we note that many of these approaches are rarely trained on Asian or Singaporean food. We hope to build a model to make accurate predictions of the nutritional content of Singaporean food, so that Singaporeans will find it much easier to manage their nutrition.

Research Objectives Therefore, our project aims to answer the following research questions:

1. How can we build a model for accurate predictions of nutritional content from common Singaporean food images?
2. How might we use machine learning knowledge to improve training and performance of our models?

Application Requirements Finally, our application must possess these characteristics:

1. Inference should be performed quickly.
2. The end-to-end workflow of our proposed application should be (i) user takes image of food, (ii) nutritional information (fat, sodium etc) is presented to the user, allowing for quicker evaluations for their dietary decisions.

Our Solution

Initial Explorations

`(BASELINE-V1, BASELINE-V2)` We design two black-box convolutional neural network models that predict nutritional information from an image, training on ground truth nutrition data. This model serves as the baseline.

Approaches

We propose 3 alternatives to enhance our solution:

`(CLASSIFY)` Multi-class image classification Instead of treating each image as the raw input to our neural network, we re-frame the problem to image classification of the type of dish. After classifying, we perform a look up on the identified dish on an existing database to retrieve nutritional information.

`(SEGMENT)` Segmentation for portion estimation This approach leverages segmentation techniques to predict individual food types within dishes and estimate its portion. This method enables precise calculation of nutritional information by utilizing the estimated mass of each food type and its corresponding nutritional data.

`(BASELINE-RELEM, BASELINE-TRANSFER)` Base model enhancements We enhance our neural network by introducing additional layers of information, such as using text embeddings (Bert, Word2Vec, BiLSTM) learned from recipe information, or transfer learning by swapping out the image encoder for a pre-trained one.

Methods

Dataset & Preprocessing

We use these datasets to train and evaluate our model – Recipes5k/Recipe1M dataset [2], Nutrition 5K [6], Health Promotion Board (HPB) nutritional information [1], FoodSG-233 [9].

Recipes5k/Recipe1M dataset contains images of dishes along with corresponding ingredients list. Nutrition5k dataset contains dishes images along with a breakdown of each dish's ingredients and their corresponding nutritional values as well as the overall nutritional values of the dish.

To streamline the dataset, we cleaned the data and extracted only the overall nutritional information, while discarding the detailed ingredient breakdown. From FoodSG-233 dataset, we obtained various images of different Singaporean cuisine. However, it did not provide the nutritional values. Hence, we retrieved a standardised nutritional values for a given dish from HPB ENCF. These data are added to enrich existing dataset with more Singaporean food.

During preprocessing, various techniques such as random flipping, rotating, scaling and adjusting the contrast and brightness of images were experimented with to augment the dataset and ensure normalization of images.

Base model `(BASELINE)`

The initial architecture, `BASELINE-V1` (Figure 1), comprised a sequence of three convolutional layers followed by max pooling layers, culminating in an adaptive average pooling layer and three fully connected layers. The output layer consisted of five nodes corresponding to mass, calories, fats, protein and carbohydrates.

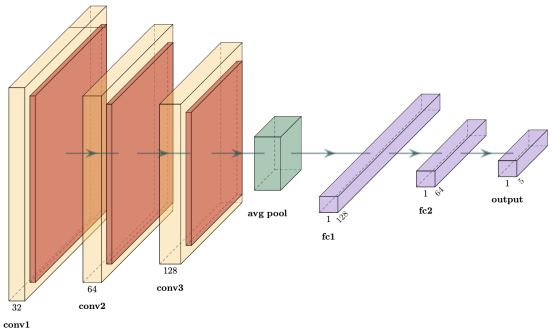


Figure 1: Architecture of `BASELINE-V1`

To introduce increased complexity to better learn nutritional information, we create an enhanced `BASELINE-V2` (Figure 2). Specifically, we increased the convolutional layers to five, each followed by a max pooling layer. Following the adaptive average pooling layer, a multitask architecture was implemented. This architecture branched out after the second fully connected layer, leading to five additional fully connected layers dedicated to predicting each nutritional component.

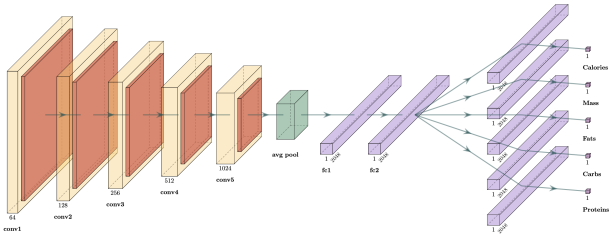


Figure 2: Architecture of `BASELINE-V2`

Throughout the training phase of the model, we optimize based on the loss function used in Google's research paper [6]. The model aims to minimize the overarching multi-loss function l_{multi}, which combines three distinct sub loss functions.

$$
\begin{aligned}
l_{cal}(I, y^{cal}|W) &= |\hat{y}^{cal} - y^{cal}| \\
l_{w}(I, y^{w}|W) &= |\hat{y}^{w} - y^{w}| \\
l_{m}(I, Y^{m}|W) &= \frac{1}{|M|} \sum_{j \in M} |\hat{y}_j^{m} - y_j^{m}| \\
l_{multi}(D|W) &= \frac{1}{N} \sum_{i=1}^{N} [l_m(I, Y^m|W) \\
&\quad + l_w(I, y^w|W) \\
&\quad + l_{cal}(I, y^{cal}|W)]
\end{aligned}
$$

The components consist of: l_{cal}, measuring the absolute error between predicted ($\hat{y}^{cal}$) and actual (y^{cal}) calories; l_w, quantifying the absolute error between predicted ($\hat{y}^w$) and actual (y^w) mass; and l_m, assessing the mean absolute error across predicted and actual values of fats, carbohydrates and protein.

Base model + ResNet50 encoder (`BASELINE-TRANSFER`)

In the Base model, we used our own network architecture. We propose a modification of including a pre-trained Convolutional Network (ResNet50), removing its top fully connected layer and using the resulting network as a feature extractor instead (Figure 3) Via transfer learning, we leverage the abilities of the pre-trained network trained on a large dataset to extract image features. We hypothesize that the learned weights from ResNet50 might already encode important knowledge about food, thus improving the model.

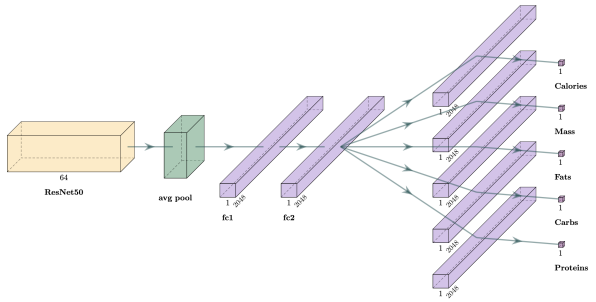


Figure 3: Architecture of `BASELINE-V2-TRANSFER`

In training this model, the weights for the pre-trained ResNet50 are frozen, and so only the weights for the rest of the model after ResNet50 are updated.

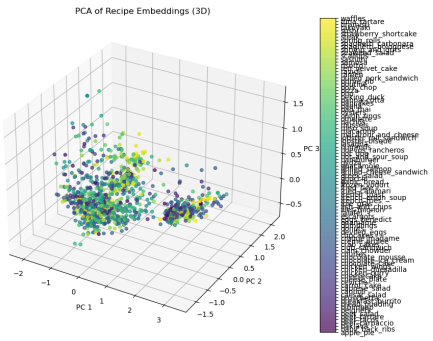


Figure 4: 3D plot of 1K Word2Vec recipe embeddings

Cross-modal embedding alignment (`BASELINE-RELEM`)

We attempt to further enhance the base model by training a separate image encoder which can extract cooking-method independent features from food images using recipes as a semantic anchor. We refer to Wu et al. for implementation details [8]. We train an image encoder to align its image-derived embeddings with the text embeddings derived from their recipes, with the goal of achieving a semantic alignment for the visual encoder's embeddings to ingredients. We hypothesize that combining visual semantics and text-based semantics will achieve better results.

For this task, we use the Recipes5K dataset [2], which contains 4826 unique recipes consisting of an image and a corresponding list of ingredients.

BERT We attempt to encode the recipes in a way that preserves semantic integrity. Early trials using a pre-trained BERT encoder did not yield successful results due to its sensitivity to sequential ordering of words in a sentence. This made it unsuitable for recipe ingredients where ordering is arbitrary. Subsequent training utilizing a ResNet encoder architecture resulted in model collapse, where the model predicted a constant value across different inputs in an attempt to reduce validation set loss.

BiLSTM As an alternative to the BERT, in order to introduce independence of ingredient order, we use a bi-directional LSTM to train our text embeddings. Each ingredient in a recipe is embedded using Word2Vec pre-trained on the `word2vec-google-news-300` corpus, and padded to reach the max recipe length. These embeddings are passed into a bi-directional LSTM and trained for 500 epochs to predict the class of recipe (many-to-one). The final layer of the LSTM was taken to represent the trained embeddings. However, during testing, our LSTM-generated embeddings also exhibited model collapse and produced constant, incorrect outputs.

Word2Vec Finally, we train a Word2Vec model on our recipe corpus to generate 500-dimensional embeddings for each ingredient. For example, in a recipe such as *"spaghetti, bacon, garlic, egg, cheese, black pepper, salt"*, we first extract Word2Vec embeddings for each ingredient. We average these embeddings to form a single vector representing the entire recipe's semantic content. Our embedding results are visualized in Figure 4. We identify 3 primary clusters: Western foods, desserts and Asian foods. Exotic foods were generally positioned far from these major clusters in the 3D space, proving that our embedding technique could semantically distinguish food types. However, this simplistic approach introduced variability in embeddings for dishes with the same name but different ingredient lists.

We modified the ResNet50 architecture by replacing the original fully connected layer with a layer matching the dimension of our recipe embeddings. Training involved a combination of positive samples (correct recipe-image pairs) and negative samples (recipe-image pairs from different food categories), equally balanced in our final training set. Our loss function reduces the cosine similarity between both embeddings, utilizing the Pytorch implementation of `nn.CosineEmbeddingLoss` with a margin of 0.1, learning rate of 0.001 and a batch size of 64.

Despite the clear semantic differentiation achieved with

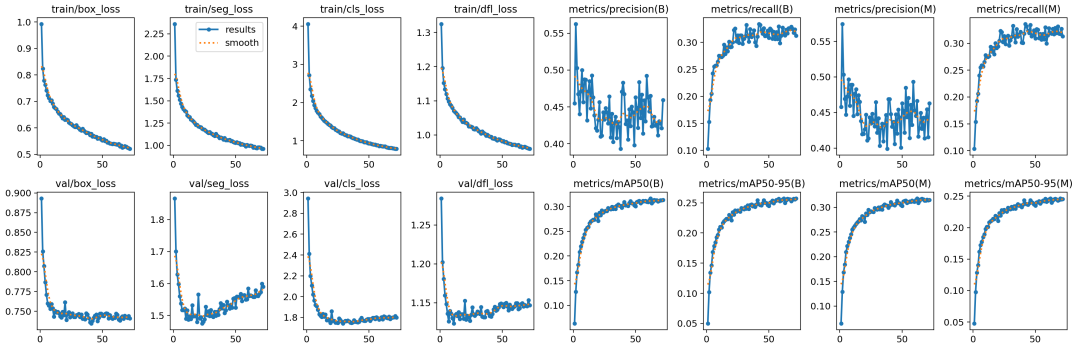


Figure 5: Training losses and precision-recall curves for YoloV8s-seg (SEGMENT) on FoodSeg103 dataset

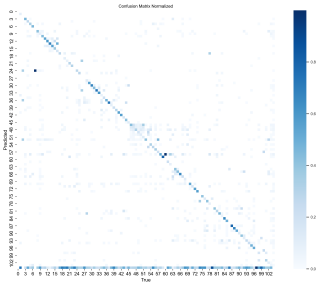


Figure 6: Confusion matrix for SEGMENT on FoodSeg103

the recipe embeddings, the ResNet50 + Word2Vec combination struggled to converge significantly. We hypothesize that the averaging of the embeddings might have diluted some critical semantic detail and contributed to the observed performance issues. The lowest observed cosine loss was 0.5. Unfortunately, the model did not perform well in accurately retrieving recipes based on images of food.

Image classification model (CLASSIFY)

Inspired by previous works [4, 9], we train an image classification model where images are classified into various menu items. The item labels are then cross-referenced against the HBP ENCF database [1] to generate an estimate of nutritional. This approach is limited by the classes present in the training dataset. We use FoodSG-233, containing 233 common Singaporean dishes [9]. Our model was able to classify images to an average of 77% accuracy on our validation set.

Segmentation model (SEGMENT)

We hypothesize that better portion estimation leads to more accurate predictions of nutritional values from images of food. This process not only requires identifying the volume of food but also the specific food present in each serving. We train a model based on YOLOv8 [7] to perform this segmentation task.

YOLOv8 is a model that performs instance segmentation with a deep convolutional neural network (CNN) architecture. This segmentation allows pixels to be classified into specific food types, and enables us to accurately calculate the pixel-wise area that each food type occupies. In an ideal scenario where all the food items on the plate are fully visible, it becomes easy to determine their nutritional content by referencing a standard database. We use the USDA's National Nutrient Database [3] for this purpose. This approach assumes no occlusions or overlapping food items, and relies on established data to quantify each food item's nutritional values.

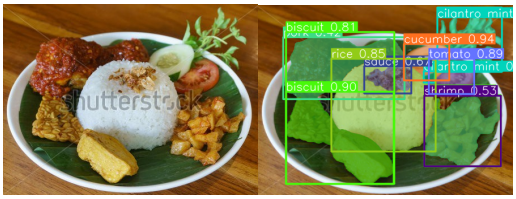
We trained a YOLOv8s-seg (SEGMENT) instance segmentation model on the FoodSeg103 dataset [8]. FoodSeg103 comprises 103 food categories across 7118 dish images, each accompanied by pixel-level annotations specifying the foods present within the dish, with non-food pixels classified as the background. After training for 72 epochs with batch size of 16 and scheduled learning rate from 0.05, we obtained our best model with a food detection precision of **0.459** and recall of **0.312**. Figure 5 presents an analysis of our model's performance. The normalized confusion matrix (Figure 6) depicts the model's predicted food classes over the actual food classes. Notably, there were many food types which were misclassified as background pixels. Figure 7 highlights the predicted food types and segmentation masks against the original labels.

Evaluation We found that SEGMENT performed the segmentation tasks very well, especially if the food types were correctly identified. However, the inherent challenge lies in the diverse preparation styles that similar foods can undergo. The wide range of appearances resulting from different cooking methods often led to misclassifications, as SEGMENT failed to recognise and categorize certain food types correctly. Moreover, SEGMENT is only limited to estimating nutritional information based on the food classes available in the training dataset. As a result, the model struggled with most Singaporean foods. A test on our local dishes (Figure 8) shows that our model does not generalize well to Asian food. For instance, it classifies noodles in laksa as "bean sprouts" and tofu puffs as "sausages". This could potentially be due to FoodSeg103 being composed largely of Western dishes.

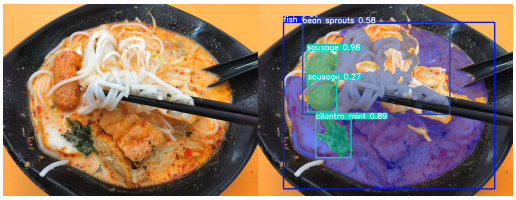
Thus, we chose not to pursue this proposed method of food type identification and portion estimation, because having high accuracy in the predicted food type was critical for

(a) Left: Ground truth labels and masks (b) Right: Predicted labels and masks

Figure 7: SEGMENT predictions on FoodSeg103 test set



(a) Nasi Lemak (b) Predicted masks



(c) Laksa (d) Predicted masks

Figure 8: SEGMENT predictions on Singaporean foods

producing reliable nutritional estimates.

Experiments & Results

Developing our Validation Set

Nutrition 5k validation set We use the Nutrition5k [6] validation set, which comprises of overhead images taken in a cafeteria. In addition, we also augment this validation set with the FoodSG-233 [9] validation set, cross-referenced against HPB ENCF [1], providing a representation of Singaporean dishes to benchmark our models against.

Custom validation set Finally, we developed our own 2 validation sets, 'Easy' and 'Hard', made up of 21 food items commonly found in Singapore. We collected our own images to ensure that the images were not present in the training data. The 'Easy' category was made up of easily identifiable food items, such as a banana or egg tart. The 'Hard' category was made up of more complex foods that we observed our models having trouble with, such as soups, stews and noodles.

For each image item, we did manual portion estimation and referenced the HPB ENCF dataset to derive nutritional information.

Results

We used various combinations of datasets, learning rates, and batch sizes to fine-tune our models. We present only those models that yielded the highest performance in the tables. Notably, the BASELINE-V2-TRANSFER model with the pre-trained ResNet50 encoder performed the best on both validation sets, achieving an average error rate of 61.3% on the 'Easy' set and 55.6% on the 'Hard' set. When examining models trained and validated exclusively on the Nutrition5K dataset, we consistently noted lower multi-loss values.

Specifically, the BASELINE-V1 model registered a multi-loss of 166.38, and the BASELINE-V2 model recorded a multi-loss of 263.49, and the BASELINE-V2-TRANSFER model showed a multi-loss of 244.94. This could suggest that the BASELINE-V2 variants, despite having higher validation losses during training, actually benefited from the integration of the FoodSG-233 dataset, evidenced by their superior performance on our custom validation set.

Finally, the CLASSIFY model could classify to an accuracy of 75% for the 'Easy' dataset and 80% for the 'Hard' dataset. We did not calculate MAE as our validation sets consisted of data from the HPB ENCF data, thus MAE was likely to be unnaturally low.

Discussion

In this study, we experiment with various ML algorithms to solve the nutrition prediction problem. By exploring a

Mean Absolute Error (MAE) on 'Easy' Validation Set

Model	Calories	Mass	Fats	Carbs	Proteins	Dataset(s)
BASELINE-V1	220.7 / 71.1%	182.7 / 113.6%	10.6 / 69.7%	25.5 / 53.6%	10.0 / 78.1%	Nutrition5K
BASELINE-V2	200.7 / 64.6%	70.4 / 43.8%	11.5 / 75.8%	34.2 / 71.9%	5.9 / 46.5%	Nutrition5K + FoodSG-233
BASELINE-V2-TRANSFER	**206.0 / 66.3%**	**83.8 / 52.1%**	**12.4 / 81.7%**	**24.2 / 50.8%**	**7.0 / 55.6%**	Nutrition5K + FoodSG-233

Mean Absolute Error (MAE) on 'Hard' Validation Set

Model	Calories	Mass	Fats	Carbs	Proteins	Dataset(s)
BASELINE-V1	201.7 / 43.9%	215.8 / 46.0%	11.1 / 64.0%	27.1 / 52.0%	54.2 / 80.8%	Nutrition5K
BASELINE-V2	193.9 / 42.2%	261.2 / 55.7%	8.3 / 47.9%	33.7 / 64.6%	56.3 / 83.8%	Nutrition5K + FoodSG-233
BASELINE-V2-TRANSFER	**191.7 / 41.8%**	**265.9 / 56.4%**	**8.2 / 46.7%**	**25.3 / 48.4%**	**59.1 / 84.5%**	Nutrition5K + FoodSG-233

breadth of techniques, including transfer learning, neural network design, and computer vision segmentation models, we discover that the challenge of creating an accurate model was harder than expected.

For our final models, we achieved our research objectives of (1) quick inference and (2) direct image-to-nutrition information. The most significant challenge was not being able to predict the nutrition values to a close degree of accuracy.

For future work, we suggest some improvements that can be made. Firstly, the FoodSG-233 dataset only consisted of various food images grouped by the types of food, and our data preparation involved applying the same set of nutritional values for each type of food to all food images of that same type with the HPB ENCF dataset. Thus, the dataset did not account for different serving sizes, different makeup of ingredients, etc.

Although we attempted to circumvent this by including textual semantics in our model, we hypothesize that an accurately labelled FoodSG-233 dataset with proper nutritional information would be better.

In addition, if we had the computational resources, we could do ensemble learning by combining prediction results of our BASELINE-V2-TRANSFER and BASELINE-V2 models, combining the prediction results.

Conclusion

In summary, we contribute four models to solve the task of food nutrition estimation from images. We construct a novel validation dataset consisting of Singaporean food to conduct our analysis. Our results show that the BASELINE-V2-TRANSFER has the best results. However, our results are limited by the lack of compute required to train more complex models, and the lack of well-annotated Singaporean food datasets.

Team members & Roles

Jefferson Lim Contributed to BASELINE-RELEM, CLASSIFY models, Dataset preparation (validation, FoodSG-233)

Ong Han Wei Contributed to BASELINE, BASELINE-TRANSFER models

Nicholas Lee Contributed to BASELINE, BASELINE-TRANSFER, BASELINE-RELEM, SEGMENT models, Dataset preparation (Recipes5k, USDA Nutrient Database)

Ryan Chua Contributed to BASELINE-TRANSFER models, Dataset preparation (HPB ENCF)

References

[1] Health Promotion Board. Energy & nutrient composition search. https://focos.hpb.gov.sg/eservices/ENCF/.

[2] Marc Bolaños, Aina Ferrà, and Petia Radeva. Food ingredients recognition through multi-label learning. 07 2017.

[3] David B. Haytowitz, Jaspreet K.C. Ahuja, Xianli Wu, Meena Somanchi, Melissa Nickle, Quyen A. Nguyen, Janet M. Roseland, Juhi R. Williams, Kristine Y. Patterson, Ying Li, and Pamela R. Pehrsson. USDA National Nutrient Database for Standard Reference, Legacy Release. 5 2019.

[4] Simon Mezgec and Barbara Koroušić Seljak. Nutrinet: A deep learning food and drink image recognition system for dietary assessment. *Nutrients*, 9(7), 2017.

[5] Ministry of Health and Singapore Health Promotion Board. National health survey 2022. https://www.moh.gov.sg/docs/librariesprovider5/resources-statistics/reports/nphs-2022-survey-report-(final).pdf, 2022.

[6] Quin Thames, Arjun Karpur, Wade Norris, Fangting Xia, Liviu Panait, Tobias Weyand, and Jack Sim. Nutrition5k: Towards automatic nutritional understanding of generic food, 2021.

[7] Ultralytics. ultralytics/ultralytics: Yolov8 in pytorch. https://github.com/ultralytics/ultralytics, 2024.

[8] Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven CH Hoi, and Qianru Sun. A large-scale benchmark for food image segmentation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 506–515, 11 2021.

[9] Kaiping Zheng, Thao Nguyen, Jesslyn Hwei Sing Chong, Charlene Enhui Goh, Melanie Herschel, Hee Hoon Lee, Changshuo Liu, Beng Chin Ooi, Wei Wang, and James Yip. From plate to prevention: A dietary nutrient-aided platform for health promotion in singapore. *arXiv preprint arXiv:2301.03829*, 2023.