AML Project Final Writeup

Harry Cui (hzc4), Min Tae Lee (ml2633), Sara Wang (sw2255)

Abstract

Image classification and segmentation can be applied to food and refrigerator images to potentially combat food waste, help individuals manage their nutritional needs, and generally make food-related decision making easier. Current work either focuses primarily on prepared dishes or fragmented categories of individual ingredients. This project aims to build a suitable dataset of individual ingredients that are commonly found in refrigerators and apply image classification and segmentation algorithms to cataloguing items in refrigerator images. Some algorithms we explored for the classification task include Inception ResNet V2, Efficient Net B0, and VGG-16. For segmentation, we tried edge detection and contour approximation algorithms with limited success, so we further trained a Mask RCNN model and demonstrated the model's capacity for refrigerator image segmentation and classification.

Motivation

Our project aims to classify images of food items within a refrigerator in order to help individuals catalogue their kitchen food stock and inform their food-related decision making. This is an application of image classification and image segmentation to the challenge of helping chefs and home cooks gain a more complete understanding of what ingredients they have in their fridge and what they can do with those ingredients.

Previous work in this space has been primarily focused on identifying prepared dishes (e.g. pasta, soup, pie, etc), rather than individual ingredients. Mainstream adoption of smartphones has made personal food habit recording more commonplace, however the process of taking photos and manually inputting labels and nutritional data for the dish can be tedious and time-consuming. Improving the ease of logging and identification of these items would allow users to log their food consumption more consistently. This would ultimately help users better meet their goals for using such services, for example tracking nutritional information to help manage a health condition like diabetes or simply gaining a better awareness of the food that they eat each day.

We wanted to expand upon the food image classification task to incorporate standalone ingredients such as vegetables, fruits, meats, and eggs. By doing so, we hope to help individuals improve their understanding of the ingredients that they already have. This task could then be applied to suggesting recipes based on existing ingredients a user has in their fridge, suggesting ingredients that a user needs to buy given a recipe they would like to make, and even predicting when food might spoil to help combat the issue of food waste.

Datasets

One challenge we encountered was the availability of good data. First, many of the standard machine learning food image datasets we found, e.g. Food-101, Food-256, and Food-524, are populated with images of cooked dishes [4]. Second, datasets we found that included images of standalone ingredients, such as the Grocery Store and Fruits and Vegetable Image Recognition datasets, were not independently sufficient for the goal of our project [5]. These datasets either did not have sufficient commonly-occurring classes of items or did not have enough images in some of their existing classes.

To address these deficiencies, we identified items we felt were commonly found in refrigerators, including vegetables, fruits, dairy products, condiments, and drinks. Using the Grocery Store dataset as our base dataset, we incorporated images and categories from the Fruits and Vegetable Image Recognition, Fruits 360, Freiburg Groceries, and Meat Images datasets [3, 5, 6]. We then wrote a Python script to scrape Google Images for pictures of other categories we felt were necessary (e.g. beer, red meat, chicken, and yogurt) and to even out the number of images in each category. After manually reviewing and editing each category to make sure that the images pulled were representative and appropriate, our final dataset included 63 classes and 70,189 images. See Appendix Figures 1 and 2 for sample images. See Appendix Figure 3 for the class distribution.

In order to train a neural network to do both image segmentation and classification, we also needed an annotated dataset of refrigerator images. Using the University of Oxford VGG Image Annotator (VIA) tool, we curated and annotated a collection of 76 refrigerator images that were either taken of our own fridges or pulled from the r/FridgeDetective subreddit [1]. See Appendix figures 4 and 5 for sample annotated images.

Method

Since our project is an application of image classification and image segmentation algorithms to cataloguing refrigerators, we experimented with a couple of different state-of-the-art neural networks to solve this problem. To begin with, we utilized transfer learning to reapply different successful image classification neural networks to the task of distinguishing items in our dataset consisting of 63 different refrigerator items.

Some pre-trained neural networks that we opted to experiment with initially were ResNet-50, Inception ResNet V2 [2], VGG16, and Efficient Net B0. To retrain these models, we replaced the last couple of layers with some dense layers that gave 63 different outputs corresponding to our refrigerator items. We also created an 80-20 split of our dataset into a training set and validation set to retrain our models and validate the results in order to compare the different models. We loaded in each neural network pretrained on ImageNet from Tensorflow and used our train-validation split to retrain and fine-tune the model for multiple epochs. To avoid overfitting, we used the validation split to calculate validation loss at each epoch and saved the model weights for the best epoch.

With our trained models, we wanted to see if we could have the model predict the contents of refrigerators. To accomplish this, we implemented image segmentation methods to segment patches of the refrigerator images and insert them into our trained model. We used a combination of edge detection and contour approximation techniques from the OpenCV library to segment patches of a given image that contained possible points of interests (Appendix Figure 9). Patches that contained points of interests were then fed into our model to see if our model could predict any food in each patch with confidence. From those predictions, we removed any that were predicted with confidence level below 0.8. We also removed any patches that overlapped too much with another patch that was predicted with higher confidence.

Due to the poor performance of the non-neural network segmentation algorithms in conjunction with the refrigerator food classifiers, we decided to pivot and train a neural network to do both image segmentation and classification. Using the Mask RCNN model with a ResNet101 backbone trained on the Microsoft COCO dataset, we replaced the last classification layers to give 63 outputs corresponding to our refrigerator items. We then trained the head at a

learning rate of 0.001 for 10 epochs using our self-annotated dataset consisting of 76 refrigerator images. Our annotated refrigerator dataset was also split 80-20 for training and validation.

Experiments

The first part of our experiment was to see how well our different models worked with our baseline food image dataset, the Food-11 dataset from EPFL. This dataset consisted of 16,643 food images grouped in 11 major food categories such as bread, dairy products, fruits, vegetables, and meat. We used this dataset to choose two better-performing models to use on our larger 63-class dataset.

From the four models, EfficientNet and Inception ResNet V2 performed much better than the other two models. EfficientNet had 93.35% accuracy on the training set and 70.13% accuracy on the validation set. Inception ResNet V2 had 78.56% accuracy on the training set and 84.80% accuracy on the validation set. Since VGG16 had a much smaller number of layers, its training time was much faster than the other models. However, the faster training time was offset by lower accuracy on the same number of epochs--VGG16 had 92.25% training accuracy but 63.79% accuracy on the validation set. Of the four models, the ResNet-50 had the worst performance with the training accuracy remaining ~10-15% over multiple epochs.

Using these results, we retrained the three selected models on our self-created, expanded grocery dataset to see how well our models classified each of the food images into their corresponding food categories. Appendix Figure 12 includes a table and graphs that summarize the results of our models on this dataset.

Using our self annotated dataset consisting of 76 refrigerator images, we fine tuned the Mask RCNN model with a randomly split 60 image training set and 16 image validation set for 10 epochs. At the end of training, we achieved a mean average precision of 0.34535 on our validation set. See Appendix Figure 13 for a graph illustrating our mean average precision over the ten epochs.

Discussion and Context

Unfortunately, our image detection algorithm in conjunction with non-neural network image segmentation algorithms did not perform well in classifying food in our refrigerator images. Even though we were able to segment points of interests in a given image somewhat successfully, our model was not able to classify the contents in the patches accurately.

At first, we had assumed that differences in lighting conditions, overlaps among contents in the refrigerator, and empty spaces caused inconsistencies and inaccuracies in our classification. Even though it is true that the lack of true negative images in our dataset might have caused many false positive classifications in our results, we consistently saw inaccurate classifications even in those areas that our algorithm was able to segment properly (Appendix Figure 10). We further tested the accuracy of our classification algorithm by using images of food in a less complicated environment (Appendix Figure 11), but this was also unsuccessful. Even though our classifier was able to classify one or two items correctly in a given image containing 5 or 6 food items, we consistently saw our classifier making either incorrect classification or no classification on the other items in the image.

This inaccuracy can be attributed to our model overfitting to our dataset and our dataset being inconsistent with the actual food images in real life and in different environments. In fact, this inconsistency can somewhat be seen in our training and validation accuracies. Even though our model was able to achieve fairly high training accuracies with different models, we saw that

the validation accuracies were much lower than their corresponding training accuracies. This phenomenon indicates that our model might be suffering from overfitting and that many of the learnings from our training data might not be applicable to the larger data population.

There are relatively few previous studies applying neural networks to refrigerator food item segmentation and classification, but prior work by Zhu et al. applies a variation of the Faster RCNN model to this task with some success [7]. Given that Faster RCNN only produces bounding boxes, we thought the application of the Mask RCNN to this task could potentially produce even better results with a pixel-level specificity of where refrigerator items are located. For crowded refrigerators with a variety of differently shaped objects, more granular mask segmentations could allow us to pick out individual items without the noisy background.

In training the Mask RCNN on our self-annotated dataset, we noticed a couple strengths of the model as well as issues with our dataset and refrigerator item classification. In Appendix Figure 6, we can see that the model is able to capture more pixel level details such as the individual grapes in the top left grape bunch and each pepper, even those in the background, on the second shelf. However, refrigerator liquids such as beers, sodas, and juices appear similarly in cans and bottles, making it difficult to distinguish between these classes. We see this in Appendix Figure 7, where the model labels the same can as both beer and soda. Given the limitations of our small dataset, there are not many distinguishing features between these classes other than the branding.

When manually annotating the refrigerator images, we also noticed that 63 classes was not enough for the refrigerator cataloguing task, and we added catch-all classes like leftovers and condiments to capture items we could not specifically label. Despite the small size of our annotated dataset, we saw that Mask RCNN can achieve some success with labeling (for example, with Appendix Figure 8), which suggests that Mask RCNN may perform even better for this task with a more expansive dataset.

Conclusion

Overall, we were able to test two different ways of image classification using deep learning. Even though our results for image classification using transfer learning and non-neural network image segmentation were not very successful, we were able to demonstrate that Mask RCNN is capable of cataloguing refrigerator items with some success even in crowded refrigerator images. However, the performance of our model was limited by our small annotated dataset of 76 refrigerator images representing only 63 classes.

To further improve the model's cataloguing ability, future work could be done to assemble a larger dataset with a wider variety of items. Collecting a larger dataset or performing text recognition on segmented items may help distinguish similar beverages like soda and beer. Furthermore, using a more detailed variety of items and labels could detect common condiments and brands like Heinz Ketchup and Coca-Cola. Finally, future work should experiment with state-of-the-art semantic segmentation models other than Mask RCNN such as YOLO and RetinaNet for the refrigerator cataloguing task.

Appendix

Figure 1: Example dataset class (Carrots)

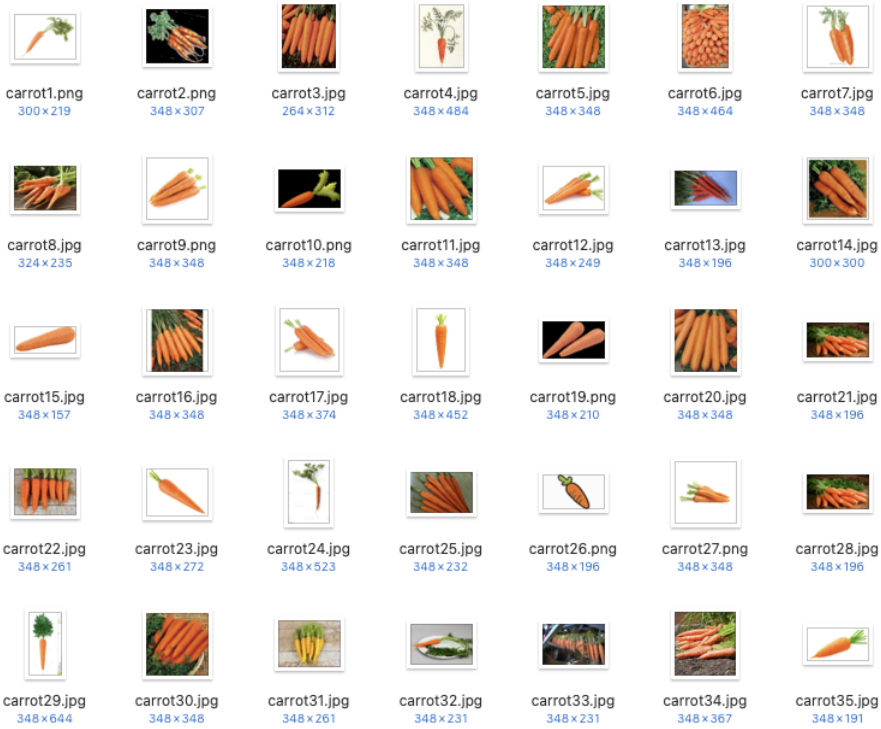


Figure 2: Example dataset class (Soda)

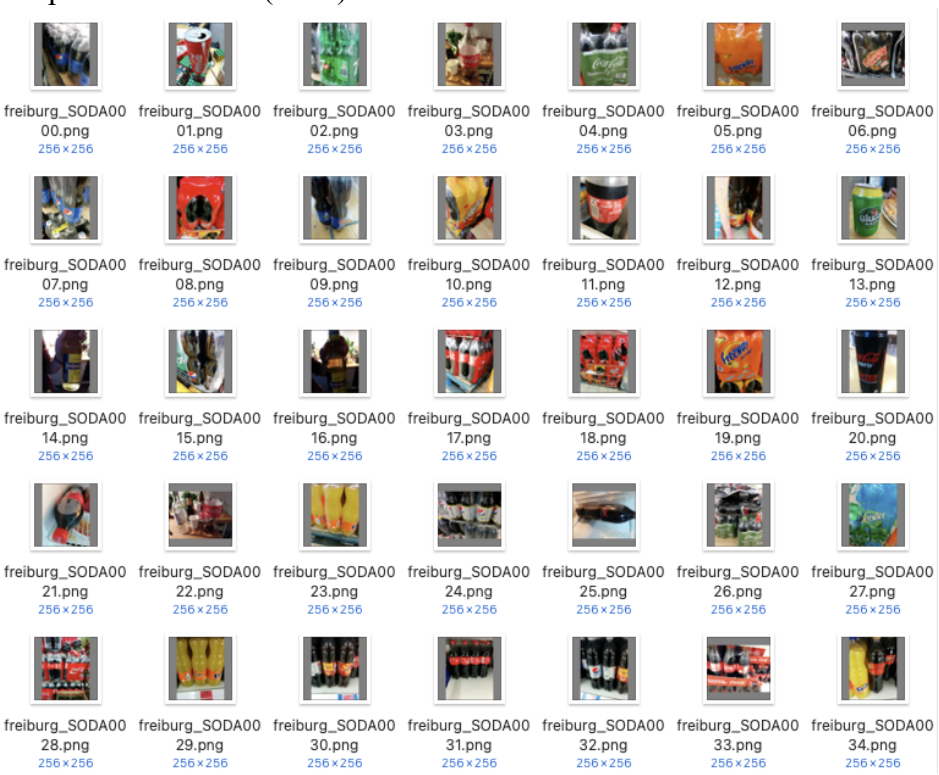


Figure 3: The class distribution of our food item dataset.

apple	condiments	kiwi	peaches	soybeans
avocado	corn	leek	pear	spinach
banana	cucumber	leftovers	peas	strawberries
beer	dates	lemon	pepper	sweetpotato
beetroot	eggplant	lettuce	pineapple	tomato
blueberries	eggs	limes	plum	tomatosauce
cabbage	figs	mango	pomegranate	turnip
cantaloupe	garlic	milk	potato	water
carrot	ginger	mushroom	radish	watermelon
cauliflower	grapefruit	onion	raspberries	yogurt
cheese	grapes	orange	redmeat	zucchini
cherries	guava	papaya	soda	
chicken	juice	passionfruit	sourcream	

Figure 4: Sample annotated refrigerator image using VGG Image Annotator

Figure 5: Sample annotated refrigerator image using VGG Image Annotator


1
2
3
4
5
43
19
12
10
9
11
13
8
7
16
18
17
14
15
24
23
21
29
28
25
26
31
33
27
30
32
38
39
40
41
35
36
34
37
42

Figure 6: Sample refrigerator segmentation and classification by our trained Mask RCNN model showing detection of crowded food items.

Figure 7: Sample refrigerator segmentation and classification by our trained Mask RCNN model showing difficulty distinguishing between beer and soda.

Figure 8: Sample refrigerator segmentation and classification by our trained Mask RCNN model showing some success.

Figure 9: Sample result of finding points of interests in a refrigerator image

Figure 10: Sample refrigerator classification using food classifier in conjunction with non-neural network image segmentation

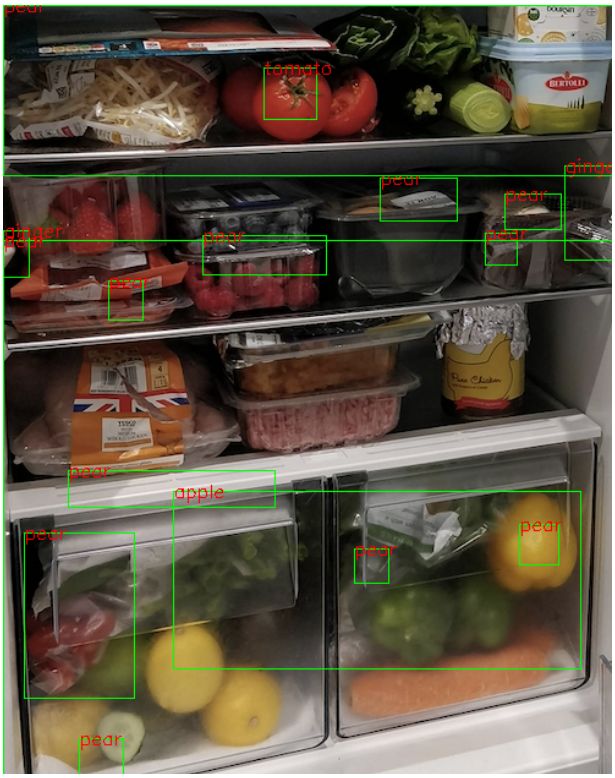


Figure 11: Sample classification using food classifier in conjunction with non-neural network image segmentation (simpler environment)

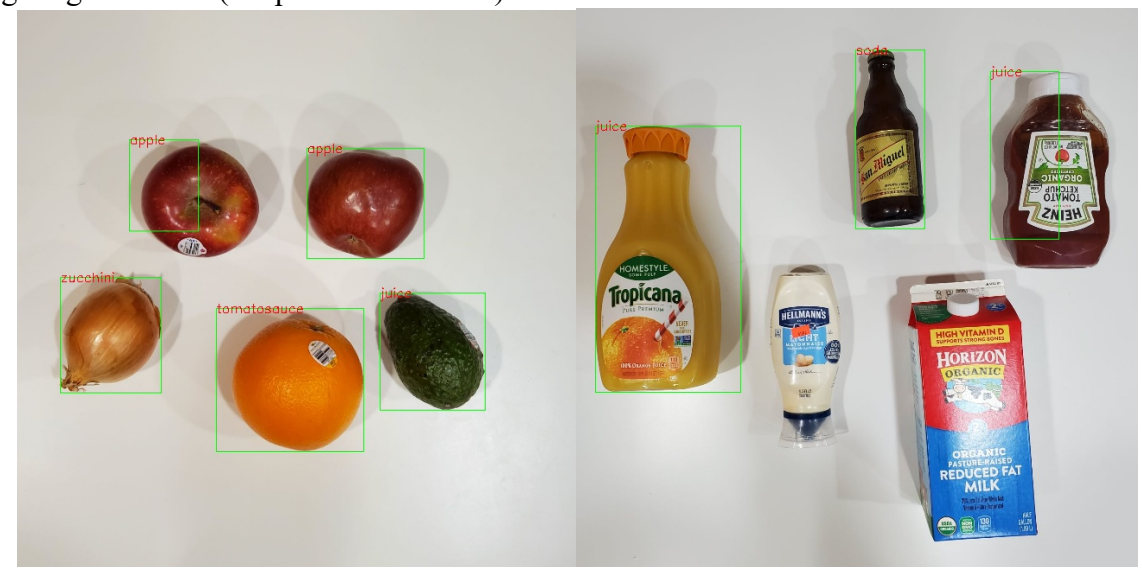
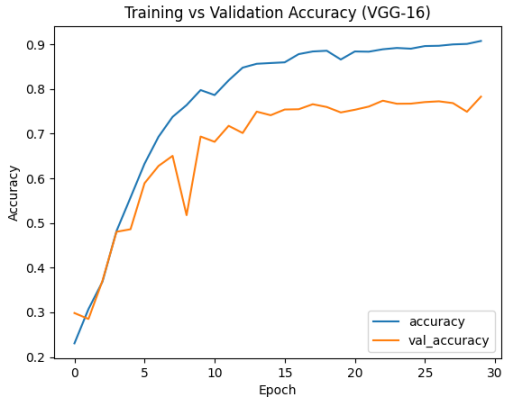
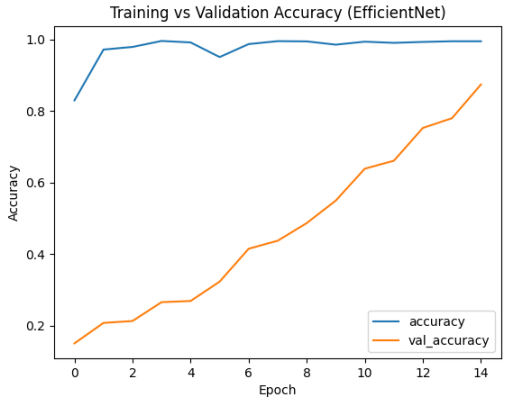


Figure 12: Summary of the results of our models on the new larger dataset

Model	Training Accuracy	Validation Accuracy
Inception ResNet V2	92.94%	85.86%
Efficient Net B0	99.51%	87.45%
VGG - 16	90.75%	78.29%





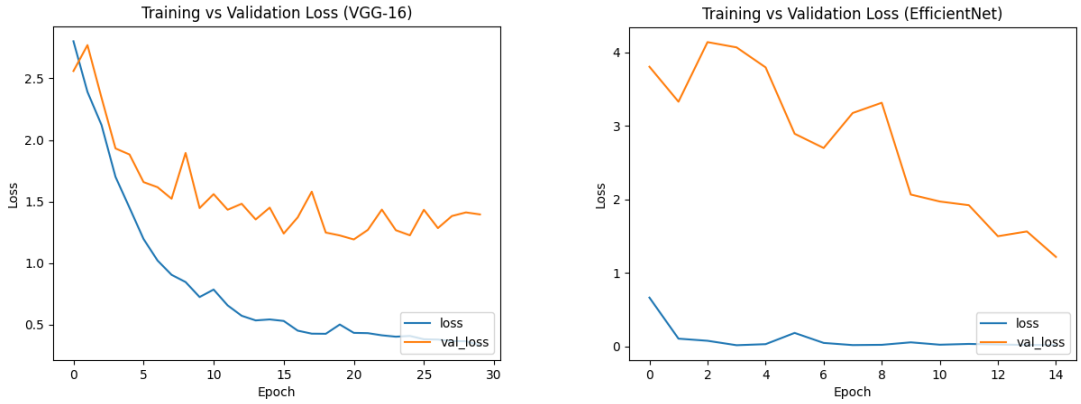
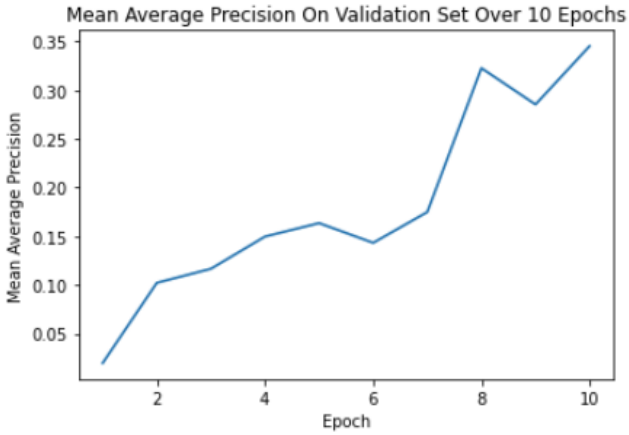


Figure 13: Mean Average Precision Over 10 Epochs on Validation Set for Mask RCNN Training



References

[1] A. Dutta and A. Zisserman, "The VIA Annotation Software for Images, Audio and Video," *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.

[2] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *arXiv.org*, 23-Aug-2016. [Online]. Available: https://arxiv.org/abs/1602.07261. [Accessed: 14-Dec-2020].

[3] H. Mureşan and M. Oltean, "Fruit recognition from images using deep learning," *Acta Universitatis Sapientiae, Informatica*, vol. 10, no. 1, pp. 26–42, 2018.

[4] L. Bossard, M. Guillaumin, and L. V. Gool, “Food-101 – Mining Discriminative Components with Random Forests,” *Computer Vision – ECCV 2014 Lecture Notes in Computer Science*, pp. 446–461, 2014.

[5] M. Klasson, C. Zhang, and H. Kjellstrom, “A Hierarchical Grocery Store Image Dataset With Visual and Semantic Labels,” *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[6] P. Jund, N. Abdo, A. Eitel, and W. Burgard, “The Freiburg Groceries Dataset,” *arXiv.org*, 17-Nov-2016. [Online]. Available: https://arxiv.org/abs/1611.05799. [Accessed: 14-Dec-2020].

[7] Y. Zhu, X. Zhao, C. Zhao, J. Wang, and H. Lu, “Food det: Detecting foods in refrigerator with supervised transformer network,” *Neurocomputing*, vol. 379, pp. 162–171, 2020.